

LEARNING MADE EASY



2nd Edition

Microsoft®  
**Excel® Sales  
Forecasting**

for  
**dummies®**  
A Wiley Brand

Choose, manage,  
and present data

Select the right forecasting  
method for your business

Use moving averages and  
predict seasonal sales

**Conrad Carlberg, PhD**

[www.allitebooks.com](http://www.allitebooks.com)





---

# Excel<sup>®</sup> Sales Forecasting

for  
**dummies**<sup>®</sup>  
A Wiley Brand





# Excel<sup>®</sup> Sales Forecasting

for  
**dummies**<sup>®</sup>  
A Wiley Brand

2nd edition

by **Conrad Carlberg, Ph.D**

for  
**dummies**<sup>®</sup>  
A Wiley Brand

## Excel® Sales Forecasting For Dummies®, 2nd Edition

Published by: **John Wiley & Sons, Inc.**, 111 River Street, Hoboken, NJ 07030-5774, [www.wiley.com](http://www.wiley.com)

Copyright © 2016 by John Wiley & Sons, Inc., Hoboken, New Jersey

Published simultaneously in Canada

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without the prior written permission of the Publisher. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permissions>.

**Trademarks:** Wiley, For Dummies, the Dummies Man logo, Dummies.com, Making Everything Easier, and related trade dress are trademarks or registered trademarks of John Wiley & Sons, Inc. and may not be used without written permission. Excel is a registered trademark of Microsoft Corporation. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

**LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY:** THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002. For technical support, please visit <https://hub.wiley.com/community/support/dummies>.

Wiley publishes in a variety of print and electronic formats and by print-on-demand. Some material included with standard print versions of this book may not be included in e-books or in print-on-demand. If this book refers to media such as a CD or DVD that is not included in the version you purchased, you may download this material at <http://booksupport.wiley.com>. For more information about Wiley products, visit [www.wiley.com](http://www.wiley.com).

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Control Number: 2016942855

ISBN: 978-1-119-29142-8

ISBN 978-1-119-29143-5 (ePub); ISBN 978-1-119-29144-2 (ePDF)

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

# Contents at a Glance

<b>Introduction</b> .....	1
<b>Part 1: Understanding Sales Forecasting and How Excel Can Help</b> .....	5
CHAPTER 1: A Forecasting Overview .....	7
CHAPTER 2: Forecasting: The Basic Issues .....	23
CHAPTER 3: Understanding Baselines .....	41
CHAPTER 4: Predicting the Future: Why Forecasting Works .....	53
<b>Part 2: Organizing the Data</b> .....	69
CHAPTER 5: Choosing Your Data: How to Get a Good Baseline .....	71
CHAPTER 6: Setting Up Tables in Excel .....	87
CHAPTER 7: Working with Tables in Excel .....	103
<b>Part 3: Making a Basic Forecast</b> .....	119
CHAPTER 8: Summarizing Sales Data with Pivot Tables .....	121
CHAPTER 9: Charting Your Baseline: It's a Good Idea .....	141
CHAPTER 10: Forecasting with Excel's Data Analysis Add-in .....	159
CHAPTER 11: Basing Forecasts on Regression .....	173
<b>Part 4: Making Advanced Forecasts</b> .....	189
CHAPTER 12: Entering the Formulas Yourself .....	191
CHAPTER 13: Using Moving Averages .....	219
CHAPTER 14: Changing Horses: From Moving Averages to Smoothing .....	237
CHAPTER 15: Smoothing: How You Profit from Your Mistakes .....	259
CHAPTER 16: Fine-Tuning a Regression Forecast .....	285
CHAPTER 17: Managing Trends .....	311
CHAPTER 18: Same Time Last Year: Forecasting Seasonal Sales .....	329
<b>Part 5: The Part of Tens</b> .....	349
CHAPTER 19: Ten Fun Facts to Know and Tell about Array Formulas .....	351
CHAPTER 20: The Ten Best Excel Tools .....	363
<b>Index</b> .....	373





# Table of Contents

<b>INTRODUCTION</b> .....	1
About This Book .....	1
Foolish Assumptions .....	2
Icons Used in This Book .....	2
Beyond the Book .....	3
Where to Go from Here .....	3
<b>PART 1: UNDERSTANDING SALES FORECASTING AND HOW EXCEL CAN HELP</b> .....	5
<b>CHAPTER 1: A Forecasting Overview</b> .....	7
Understanding Excel Forecasts .....	8
Method #1: Moving averages .....	9
Method #2: Exponential smoothing .....	9
Method #3: Regression .....	10
Getting the Data Ready .....	10
Using tables .....	10
Ordering your data .....	12
Making Basic Forecasts .....	13
Putting moving averages to work for you .....	14
Making sense of exponential smoothing .....	16
Using regression to get what you want .....	16
Charting Your Data .....	19
Forecasting with Advanced Tools .....	20
<b>CHAPTER 2: Forecasting: The Basic Issues</b> .....	23
Why Forecast? .....	24
To plan sales strategies .....	24
To size inventories .....	25
Talking the Talk: Basic Forecasting Lingo .....	26
Autoregressive integrated moving averages (ARIMA) .....	26
Baseline .....	27
Correlation .....	27
Cycle .....	28
Damping factor .....	28
Exponential smoothing .....	28
Forecast period .....	28
Moving average .....	29
Predictor variable .....	29
Regression .....	29
Seasonality .....	30
Trend .....	30

	Understanding the Baseline . . . . .	30
	Charting the baseline . . . . .	31
	Looking for trends . . . . .	32
	Setting Up Your Forecast . . . . .	33
	Smoothing data . . . . .	34
	Regression: It's all about relationships . . . . .	34
	Using Your Revenue and Cost Data . . . . .	35
<b>CHAPTER 3:</b>	<b>Understanding Baselines . . . . .</b>	<b>41</b>
	Using Qualitative Data . . . . .	42
	Asking the right questions . . . . .	42
	Keeping your eye on the ball: The purpose of your forecast . . . . .	43
	Recovering from Mistakes in Sales Forecasting . . . . .	45
	Getting over it . . . . .	45
	Using revenue targets as forecasts . . . . .	46
	Recognizing Trends and Seasons . . . . .	47
	Identifying trends . . . . .	48
	Understanding seasonality . . . . .	50
<b>CHAPTER 4:</b>	<b>Predicting the Future: Why Forecasting Works . . . . .</b>	<b>53</b>
	Understanding Trends . . . . .	54
	Watching revenues go up — and down . . . . .	55
	Testing for trends . . . . .	59
	Matchmaker, Matchmaker: Finding Relationships in the Data . . . . .	63
	Choosing the predictors . . . . .	64
	Analyzing the correlations . . . . .	67
	<b>PART 2: ORGANIZING THE DATA . . . . .</b>	<b>69</b>
<b>CHAPTER 5:</b>	<b>Choosing Your Data: How to Get a Good Baseline . . . . .</b>	<b>71</b>
	Early to Bed: Getting Your Figures in Order . . . . .	72
	Why order matters: Moving averages . . . . .	72
	Why order matters: Exponential smoothing . . . . .	75
	Why order doesn't matter: Regression . . . . .	76
	Staying Inside the Lines: Why Time Periods Matter . . . . .	77
	Deciding how far to forecast . . . . .	78
	Choosing your time periods . . . . .	81
	Spacing Time Periods Equally . . . . .	82
	Using periodic relationships . . . . .	83
	When missing data causes unequal time periods . . . . .	85
<b>CHAPTER 6:</b>	<b>Setting Up Tables in Excel . . . . .</b>	<b>87</b>
	Understanding Table Structures . . . . .	88
	Creating a Table . . . . .	91
	Using the Total row . . . . .	93
	Using other table features . . . . .	96

	Filtering Lists . . . . .	96
	Using Excel's table filters . . . . .	96
	Using the Advanced Filter . . . . .	98
	Importing Data from a Database to an Excel Table . . . . .	100
<b>CHAPTER 7:</b>	<b>Working with Tables in Excel . . . . .</b>	<b>103</b>
	Turning Tables into Charts . . . . .	103
	Understanding chart types . . . . .	104
	Creating the chart from your table . . . . .	108
	Refining charts . . . . .	109
	Using the Data Analysis Add-in with Tables . . . . .	112
	Avoiding the Data Analysis Add-in's Traps . . . . .	115
	<b>PART 3: MAKING A BASIC FORECAST . . . . .</b>	<b>119</b>
<b>CHAPTER 8:</b>	<b>Summarizing Sales Data with Pivot Tables . . . . .</b>	<b>121</b>
	Understanding Pivot Tables . . . . .	122
	Making baselines out of sales data . . . . .	123
	Totaling up the data . . . . .	128
	Building the Pivot Table . . . . .	130
	Grouping Records . . . . .	134
	Knowing when to group records . . . . .	135
	Creating the groups . . . . .	135
	Avoiding Grief in Excel Pivot Tables . . . . .	137
	Don't use blank dates . . . . .	137
	Making multiple groups . . . . .	138
<b>CHAPTER 9:</b>	<b>Charting Your Baseline: It's a Good Idea . . . . .</b>	<b>141</b>
	Digging into a Baseline . . . . .	142
	Using date and time data in Excel . . . . .	142
	Charting dates and times in Excel . . . . .	143
	Using Line charts . . . . .	146
	Using XY (Scatter) charts . . . . .	151
	Making Your Data Dance with Pivot Charts . . . . .	152
	Using Two Value Axes . . . . .	157
<b>CHAPTER 10:</b>	<b>Forecasting with Excel's Data Analysis Add-in . . . . .</b>	<b>159</b>
	Installing Add-ins: Is the Add-in Even There? . . . . .	160
	Using Moving Averages . . . . .	162
	Moving day: Getting from here to there . . . . .	163
	Moving averages and stationary baselines . . . . .	164
	Using Exponential Smoothing . . . . .	165
	Using the Regression Tool . . . . .	168

<b>CHAPTER 11: Basing Forecasts on Regression</b> .....	173
Deciding to Use the Regression Tool .....	174
Adopting the Regression approach .....	175
Using more than one predictor variable .....	177
Understanding the Data Analysis Add-in's Regression Tool .....	178
Checking the forecast errors .....	182
Plotting your actual revenues .....	183
Understanding confidence levels .....	184
Avoiding a zero constant .....	185
Using Multiple Regression .....	186
New predictor with forecast variable .....	187
New predictor with existing variable .....	188
<b>PART 4: MAKING ADVANCED FORECASTS</b> .....	189
<b>CHAPTER 12: Entering the Formulas Yourself</b> .....	191
About Excel Formulas .....	192
Doing it yourself: Why bother? .....	192
Getting the syntax right .....	198
Using Insert Function .....	199
Understanding Array Formulas .....	205
Choosing the range for the array formula .....	206
Excel's three-finger salute: Ctrl+Shift+Enter .....	207
Recognizing array formulas .....	208
A special problem with array formulas .....	209
Using the Regression Functions to Forecast .....	210
Using the LINEST function .....	210
Selecting the right range of cells .....	213
Getting the statistics right .....	213
Using the TREND function .....	215
<b>CHAPTER 13: Using Moving Averages</b> .....	219
Choosing the Length of the Moving Average .....	220
Signaling: Left turn coming up? .....	220
A little less noise, please .....	221
Stepping it up .....	224
Reacting Quickly versus Modeling Noise .....	226
Getting a smoother picture .....	227
Calculating and charting moving averages .....	228
Using the Data Analysis Add-in to Get Moving Averages .....	229
Using the Data Analysis add-in's Moving Average tool .....	230
Charting residuals .....	234

<b>CHAPTER 14: Changing Horses: From Moving Averages to Smoothing</b> .....	237
Losing Early Averages .....	238
Understanding Correlation .....	240
When did they start going together? .....	240
Charting correlated data .....	243
Understanding Autocorrelation .....	244
Calculating autocorrelation .....	253
Diagnosing autocorrelation .....	254
<b>CHAPTER 15: Smoothing: How You Profit from Your Mistakes</b> .....	259
Correcting Errors: The Idea Behind Smoothing .....	260
Adjusting the forecast .....	260
Why they call it “exponential smoothing” .....	263
Fooling around with the smoothing constant .....	266
Using the Smoothing Tool’s Formula .....	269
Getting a forecast from the Exponential Smoothing tool .....	269
Modifying the smoothing constant .....	273
Finding the Smoothing Constant .....	275
Developing the yardstick .....	275
Minimizing the square root of the mean square error .....	278
Problems with Exponential Smoothing .....	282
Losing an observation at the start .....	282
The Regression tool’s standard errors: They’re wrong .....	283
<b>CHAPTER 16: Fine-Tuning a Regression Forecast</b> .....	285
Doing Multiple Regression .....	286
Using more than one predictor .....	286
The thinking person’s approach to multiple regression .....	292
Interpreting the coefficients and their standard errors .....	296
Getting a Regression Trendline into a Chart .....	301
Evaluating Regression Forecasts .....	306
Using autoregression .....	306
Regressing one trend onto another .....	309
<b>CHAPTER 17: Managing Trends</b> .....	311
Knowing Why You May Want to Remove the Trend from a Baseline .....	312
Understanding why trend is a problem .....	312
Diagnosing a trend .....	313
Getting a Baseline to Stand Still .....	315
Subtracting one value from the next value .....	316
Dividing one value by another .....	318

Getting rates . . . . .	320
The downside of differencing . . . . .	321
And All the King's Men: Putting a Baseline Together Again . . . . .	325
<b>CHAPTER 18: Same Time Last Year: Forecasting</b>	
<b>Seasonal Sales</b> . . . . .	329
Doing Simple Seasonal Exponential Smoothing . . . . .	330
Relating a season to its ancestors . . . . .	330
Using the smoothing constants . . . . .	334
Getting Farther into the Baseline . . . . .	338
Calculating the first forecast . . . . .	338
Smoothing through the baseline level . . . . .	341
'Tis the seasonal component . . . . .	342
Finishing the Forecast . . . . .	344
Modifying the formulas . . . . .	344
Using the worksheet . . . . .	345
Using the workbook . . . . .	346
Excel 2016's new Forecast Sheet . . . . .	348
<b>PART 5: THE PART OF TENS</b> . . . . .	349
<b>CHAPTER 19: Ten Fun Facts to Know and</b>	
<b>Tell about Array Formulas</b> . . . . .	351
Entering Array Formulas . . . . .	352
Using the Shift Key . . . . .	352
Noticing the Curly Brackets . . . . .	354
Using INDEX to Extract a Value from an Array Formula's Result . . . . .	354
A Quick Route to Unique Values . . . . .	356
Selecting the Range: LINDEX . . . . .	358
Selecting the Range: TRANSPOSE . . . . .	358
Selecting a Range: TREND . . . . .	359
Editing an Array Formula . . . . .	361
Deleting Array Formulas . . . . .	362
<b>CHAPTER 20: The Ten Best Excel Tools</b> . . . . .	363
Cell Comments . . . . .	363
AutoComplete . . . . .	364
Macro Security . . . . .	365
The Customizable Toolbar . . . . .	366
Evaluate Formula . . . . .	367
Worksheet Protection . . . . .	368
Unique Records Only . . . . .	369
Using the Fill Handle . . . . .	369
Quick Data Summaries . . . . .	370
Help with Functions . . . . .	370
<b>INDEX</b> . . . . .	373



# Introduction

---

**Y**ou wouldn't have pulled this book off the shelf if you didn't need to forecast sales. And I'm sure that you're not Nostradamus. Your office isn't filled with the smell of incense and it's not your job to predict the date that the world will come to an end.

But someone — perhaps you — wants you to forecast sales, and you find out how to do that here, using the best general-purpose analysis program around, Microsoft Excel.

## About This Book

---

This book concentrates on using numbers to forecast sales. If you're a salesperson, or a sales manager, or someone yet higher up the org chart, you've run into forecasts that are based not on numbers but on guesses, sales quotas, wishful thinking, and Scotch.

I get away from that kind of thing here. I use numbers instead. Fortunately, you don't need to be a math major to use Excel for your forecasting. Excel has a passel of tools that will do it on your behalf. Some of them are even easy to use, as you'll see.

That said, it's not all about numbers. You still need to understand your products, your company, and your market before you can make a sensible sales forecast, and I have to trust you on that. I hope I can. I think I can. Otherwise, start with Part 1, which talks about the context for a forecast.

You can hop around the chapters in this book, as you can in all books that feature the guy with a pool ball rack for a head. There are three basic approaches to forecasting with numbers — moving averages, smoothing, and regression — and you really don't have to know much about one to understand another. It helps to know all three, but you don't really need to.

# Foolish Assumptions

The phrase *foolish assumptions* is, of course, redundant. But here are the assumptions I'm making:

- » **I'm assuming that you know the basics of how to use Excel.** Entering numbers into a worksheet, like numbers that show how much you sold in August 2015; entering formulas in worksheet cells; saving workbooks; using menus; that sort of thing.  
  
If you haven't ever used Excel before, don't start here. Do buy this book, but also buy *Excel 2013 For Dummies* by Greg Harvey (published by Wiley), and dip into that one first.
- » **I'm assuming that you have access to information on your company's sales history, and the more the better.** The only way to forecast what's about to happen is to know what's happened earlier. Doesn't really matter where that information is — it can be in a database, or in an Excel workbook, or even in a simple text file. As long as you can get your hands on it, you can make a forecast. And I talk about how you can get Excel's "hands" on it.
- » **I'm assuming you don't have a phobia about numbers.** You don't have to be some kind of egghead to make good forecasts. But you can't be afraid of numbers, and I really doubt that you are. Except maybe your quarterly sales quota.
- » **Of course, I'm also assuming you have Excel on your computer.** I'm *not* assuming you have the latest version. But the Excel user interface changed so drastically in 2007 that I have to assume your version uses the Ribbon rather than the original menu structure. Even so, very little of the information in this book has to do with the user interface. Mostly, it's about setting up your sales history, letting yourself be guided by Excel's Data Analysis add-in, and finally working with worksheet formulas, charts, and other tools to get where you're headed on your own.

## Icons Used in This Book

In the margins of this book, you find icons — little pictures that are designed to draw your attention to particular kinds of information. Here's what the icons mean:



TIP

Anything marked with this icon will make things easier for you, save you time, get you home in time for dinner. You get some of what I've distilled from all my years browsing those blasted newsgroups.



WARNING

Not a lot of warnings in this book, but there are a few. These tell you what to expect if you do something that Microsoft hasn't sufficiently protected you against. And there are some of those.



REMEMBER

A string around your finger. There are some things to keep in mind when you're doing your forecasts, and it's usually easier to remember them than to have to look them up over and over. I do want you to read this book over and over, as I do with murder mysteries, but you'll get your work done faster if you remember this stuff.



TECHNICAL  
STUFF

Speaking of stuff, anything marked with this icon is stuff you can probably ignore — but if you're having trouble getting to sleep you may want to read these. I don't get into heavy-duty mathematical issues here, but you see some special things about how Excel prepares your forecasts. Sleep tight.

## Beyond the Book

In addition to what you're reading right now, this product also comes with a free access-anywhere Cheat Sheet that tells you about Excel data analysis add-in tools, how to use forecasting functions, what you get out of the Excel LINST function, and what to do when setting up your baseline in Excel. To get this Cheat Sheet, simply go to [www.dummies.com](http://www.dummies.com) and search for “Excel Sales Forecasting For Dummies Cheat Sheet” in the Search box.

I've also provided files for each chapter so that you can try out what I'm talking about in the leisure of your own home. You can find these files at [www.dummies.com/go/excelsalesforecasting](http://www.dummies.com/go/excelsalesforecasting).

## Where to Go from Here

Are you looking for information about the basics of forecasting? Why it works? Why it's not just a self-licking ice cream cone? Start at Chapter 1.

Do you want to know how to put your data together in a workbook? Head to Chapter 5 to find out more about baselines, and then check out the chapters on using tables in Excel.

If you're already up on forecasting basics and tables, head for Chapter 8, where you'll see how to use pivot tables to set up the baseline for your forecast.

And if you know all that stuff already, just go to Chapter 10 and start looking at how to manage your forecasts yourself, without relying on the various tools that take care of things for you. You'll be glad you did.

# 1 Understanding Sales Forecasting and How Excel Can Help

## **IN THIS PART . . .**

In Part 1, I talk about why forecasting sales can help your business in ways that seem to have little to do with sales. Part 1 also tells you why forecasting isn't simply a matter of using formulas to crunch numbers. But, face it, some numbers have to be crunched, and here you find an introduction to baselines — which are the basis for the number-crunching. I try to convince you that forecasting really does work, and I back up that claim by showing you how.



## IN THIS CHAPTER

Knowing the different methods of forecasting

Arranging your data in an order Excel can use

Getting acquainted with the Analysis ToolPak

Going it alone

# Chapter 1

# A Forecasting Overview

A sales forecast is like a weather forecast: It's an educated guess at what the future will bring. You can forecast all sorts of things — poppy-seed sales, stock market futures, the weather — in all sorts of ways: You can make your own best guess; you can compile and composite other people's guesses; or you can forecast on the basis of wishful thinking.

Unfortunately, none of these options is truly acceptable. If you want to make better forecasts, you need to take advantage of some better options. And there *are* different ways to forecast, ways that have proven their accuracy over and over. They take a little more time to prepare than guessing does, but in the long run I've spent more time explaining bad guesses than doing the forecasts right in the first place.

Microsoft Excel was originally developed as a spreadsheet application, suited to figuring payment amounts, interest rates, account balances, and so on. But as Microsoft added more and more functions — for example, AVERAGE and TREND and inventory-management stuff — Excel became more of a multipurpose analyst than a single-purpose calculator.

Excel has the tools you need to make forecasts, whether you want to prepare something quick and dirty (and who doesn't from time to time?) or something sophisticated enough for a boardroom presentation.

The tools are there. You just need to know which tool to choose for which situation and, of course, how to use it. You need to know how to arrange data for the tool. And you need to know how to interpret what the tool tells you — whether that tool's a basic one or something more advanced.

## Understanding Excel Forecasts

If you want to forecast the future — next quarter's sales, for example — you need to get a handle on what's happened in the past. So you always start with what's called a *baseline* (that is, past history — how many poppy seeds a company sold during each of the last ten years, where the market futures wound up each of the last 12 months, what the daily high temperature was year-to-date).

Unless you're going to just roll the dice and make a guess, you need a baseline for a forecast. Today follows yesterday. What happens tomorrow generally follows the pattern of what happened today, last week, last month, last quarter, last year. If you look at what's already happened, you're taking a solid step toward forecasting what's going to happen next. (Part 1 of this book talks about forecast baselines and why they work.)

An Excel forecast isn't any different from forecasts you make with a specialized forecasting program. But Excel is particularly useful for making sales forecasts, for a variety of reasons:

- » **You often have sales history recorded in an Excel worksheet.** When you already keep your sales history in Excel, basing your forecast on the existing sales history is easy — you've already got your hands on it.
- » **Excel's charting features make it much easier to visualize what's going on in your sales history and how that history defines your forecasts.**
- » **Excel has tools (found in what's called the Data Analysis add-in) that make generating forecasts easier.** You still have to know what you're doing and what the tools are doing — you don't want to just jam the numbers through some analysis tool and take the result at face value, without understanding what the tool's up to. But that's what this book is here for.

» You can take more control over how the forecast is created by skipping the Data Analysis add-in's forecasting tools and entering the formulas yourself. As you get more experience with forecasting, you'll probably find yourself doing that more and more.

You can choose from several different forecasting methods, and it's here that judgment begins. The three most frequently used methods, in no special order, are moving averages, exponential smoothing, and regression.

## Method #1: Moving averages

Moving averages may be your best choice if you have no source of information other than sales history — but you *do* need to know your baseline sales history. Later in this chapter, I show you more of the logic behind using moving averages. The underlying idea is that market forces push your sales up or down. By averaging your sales results from month to month, quarter to quarter, or year to year, you can get a better idea of the longer-term trend that's influencing your sales results.

For example, you find the average sales results of the last three months of last year — October, November, and December. Then you find the average of the next three-month period — November, December, and January (and then December, January, and February; and so on). Now you're getting an idea of the general direction that your sales are taking. The averaging process evens out the bumps you get from discouraging economic news or temporary boomlets.

## Method #2: Exponential smoothing

Exponential smoothing is closely related to moving averages. Just as with moving averages, exponential smoothing uses past history to forecast the future. You use what happened last week, last month, and last year to forecast what will happen next week, next month, or next year.

The difference is that when you use smoothing, you take into account how bad your previous forecast was — that is, you admit that the forecast was a little screwed up. (Get used to that — it happens.) The nice thing about exponential smoothing is that you take the error in your last forecast and use that error, so you hope, to improve your next forecast.

If your last forecast was too low, exponential smoothing kicks your next forecast up. If your last forecast was too high, exponential smoothing kicks the next one down.

The basic idea is that exponential smoothing corrects your next forecast in a way that would have made your *prior* forecast a better one. That's a good idea, and it usually works well.

## Method #3: Regression

When you use regression to make a forecast, you're relying on one variable to predict another. For example, when the Federal Reserve raises short-term interest rates, you might rely on that variable to forecast what's going to happen to bond prices or the cost of mortgages. In contrast to moving averages or exponential smoothing, regression relies on a *different* variable to tell you what's likely to happen next — something other than your own sales history.

# Getting the Data Ready

Which method of forecasting you use does make a difference, but regardless of your choice, in Excel you have to set up your baseline data in a particular way. Excel prefers it if your data is in the form of a *table*. In Part 2, I fill you in on how to arrange your data so that it best feeds your forecasts, but following is a quick overview.

## Using tables



TIP

There's nothing mysterious about an Excel table. A table is something very much like a database. Your Excel worksheet has columns and rows, and if you put a table there, you just need to manage three requirements:

- » **Keep different variables in different columns.** For example, you can put sales dates in one column, sales amounts in another column, sales reps' names in another, product lines in yet another.
- » **Keep different records in different rows.** When it comes to recording sales information, keep different sales records in different rows. Put information about a sale that was made on January 15 in one row, and information about a sale made on January 16 in a different row.
- » **Put the names of the variables in the table's first row.** For example, you might put "Sales Date" in column A, "Revenue" in column B, "Sales Rep" in column C, and "Product" in column D.

Figure 1-1 shows a typical Excel table.

	A	B	C	D
1	Sales Date	Revenue	Sales Rep	Product
2	10/6/2016	\$ 7,678.26	Jones	Services
3	10/7/2016	\$ 8,253.70	Tafoya	Services
4	10/11/2016	\$ 3,052.08	James	Warranty
5	10/12/2016	\$ 4,153.27	Turgidson	Hardware
6	10/15/2016	\$ 5,701.64	Anderson	Hardware
7	10/13/2016	\$ 6,382.83	Jones	Services
8	10/15/2016	\$ 2,864.20	Anderson	Maintenance
9	10/17/2016	\$ 6,379.38	Coulter	Warranty
10	10/20/2016	\$ 1,680.76	James	Services
11	10/10/2016	\$ 6,639.68	Jones	Services
12	10/14/2016	\$ 6,190.50	James	Software
13	10/10/2016	\$ 6,721.69	Wilson	Hardware
14	10/19/2016	\$ 4,996.28	Anderson	Maintenance
15	10/8/2016	\$ 5,383.12	Coulter	Maintenance
16	10/20/2016	\$ 4,742.58	Jones	Hardware
17	10/17/2016	\$ 5,369.55	James	Software
18	10/18/2016	\$ 1,837.77	Wilson	Warranty
19	10/12/2016	\$ 2,727.85	Wilson	Software
20	10/22/2016	\$ 1,141.00	Wilson	Hardware
21	10/15/2016	\$ 7,633.23	Turgidson	Software
22	10/19/2016	\$ 8,797.87	Coulter	Maintenance
23	10/18/2016	\$ 4,859.78	James	Services
24	10/16/2016	\$ 6,296.67	Coulter	Services

**FIGURE 1-1:** You don't have to keep the records in date order — you can handle that later.

Why bother with tables? Because many Excel tools, including the ones you use to make forecasts, rely on tables. Charts — which help you visualize what's going on with your sales — rely on tables. Pivot tables — which are the most powerful way you have for summarizing your sales results in Excel — rely heavily on tables. The Data Analysis add-in — a very useful way of making forecasts — relies on tables, too.

For years, Excel depended on an informal arrangement of data called a *list*. A list looked a lot like a table does now, with field names in its first row, followed by records. But a list did not have built-in properties such as record counts or filters or total rows or even a name. You had to take special steps to identify the number of rows and columns the list occupied.

In Excel 2007, Microsoft added tables as a new feature, and tables have all those things that lists lack. One aspect of tables is especially useful for sales forecasting. As time passes and you get more information about sales figures, you want to add the new data to your baseline. Using lists, you had to define what's called a *dynamic range name* to accommodate the new data. With tables, all you need to do is provide a new record, usually in a new row at the end of the table. When you do so, the table is automatically extended to capture the new data. Anything in the

workbook — charts, formulas, whatever — is also automatically updated to reflect the new information. Tables are a major improvement over lists and this book makes extensive use of them.

You find a lot more about creating and using tables in Chapter 6. In the meantime, just keep in mind that a table has different variables in different columns, and different records in different rows.

## Ordering your data

“Ordering your data” may sound a little like “coloring inside the lines.” The deal is that you have to tell Excel how much you sold in 1999, and then how much in 2000, and in 2001, and so on. If you’re going to do that, you have to put the data in chronological order.

The very best way to put your data in chronological order in Excel is by way of pivot tables. A pivot table takes individual records that are in an Excel table (or in an external database) and combines the records in ways that you control. You may have a table showing a year’s worth of sales, including the name of the sales rep, the product sold, the date of sale, and the sales revenue. If so, you can very quickly create a pivot table that totals sales revenue by sales rep and by product across quarters. Using pivot tables, you can summarize tens of thousands of records, quite literally within seconds. If you haven’t used pivot tables before, this book not only introduces the subject but also makes you dream about them in the middle of the night.

Three particularly wonderful things about pivot tables:

- » **They can accumulate for you all your sales data — or, for that matter, your data on the solar wind, but this book is about sales forecasting.** If you gather information on a sale-by-sale basis, and you then want to know how much your reps sold on a given day, in a given week, and so on, a pivot table is the best way to do so.
- » **You can use a pivot table as the basis for your next forecast, which saves you a bunch of time.**
- » **They have a unique way of helping you group your historical data — by day, by week, by month, by quarter, by year, you name it.** Chapter 8 gives you the details, as well as much more information on pivot tables, including troubleshooting some common problems.



# Making Basic Forecasts

Part 3 gets into the business of making actual forecasts, ones that are based on historical data (that is, what's gone on before). You see how to use the Data Analysis add-in to make forecasts that you can back up with actuals — given that you've looked at Part 2 and set up your actuals correctly. (Your *actuals* are the actual sales results that show up in the company's accounting records — say, when the company recognizes the revenue.)

The Data Analysis add-in is a gizmo that has shipped with Excel ever since 1995. It includes a convenient way to make forecasts, as well as to do general data analysis. The three principal tools that the Data Analysis add-in gives you to make forecasts are:

- » Moving Averages
- » Exponential Smoothing
- » Regression

Those are the three principal forecasting methods, and they form the basis for the more-advanced techniques and models. So it's no coincidence that these tools have the same names as the forecasting methods mentioned earlier in this chapter.



TECHNICAL  
STUFF

The Data Analysis add-in is an *add-in*. An add-in does tasks, like forecasting, on your behalf. An add-in is much like the other tools that are a part of Excel — the difference is that you can choose whether to install an add-in. For example, you can't choose whether the Goal Seek tool (under What-If Analysis on the Ribbon's Data tab) is available to you. If you decide to install Excel on your computer, Goal Seek is just part of the package. Add-ins are different. You can decide whether to install them. When you're installing Excel — and in most cases this means when you're installing Microsoft Office — you get to decide which add-ins you want to use.

The following sections offer a brief introduction to the three Data Analysis tools.



REMEMBER

Given a good baseline, the Data Analysis can turn a forecast back to you. And then you're responsible for evaluating the forecast, for deciding whether it's a credible one, for thinking the forecast over in terms of what you know about your business model. After all, Excel just calculates — you're expected to do the thinking.

# Putting moving averages to work for you

You may already be familiar with moving averages. They have two main characteristics, as the name makes clear:

- » **They move.** More specifically, they move over time. The first moving average may involve Monday, Tuesday, and Wednesday; in that case, the second moving average would involve Tuesday, Wednesday, and Thursday; the third Wednesday, Thursday, and Friday, and so on.
- » **They're averages.** The first moving average may be the average of Monday's, Tuesday's, and Wednesday's sales. Then the second moving average would be the average of Tuesday's, Wednesday's, and Thursday's sales, and so on.

The basic idea, as with all forecasting methods, is that something regular and predictable is going on — often called the *signal*. Sales of ski boots regularly rise during the fall and winter, and predictably fall during the spring and summer. Beer sales regularly rise on NFL Sundays and predictably fall on other days of the week.

But something else is going on, something irregular and unpredictable — often called *noise*. If a local sporting goods store has a sale on, discounting ski boots from May through July, you and your friends may buy new boots during the spring and summer, even though the regular sales pattern (the signal) says that people buy boots during the fall and winter. As a forecaster, you typically can't predict this special sale. It's random and tends to depend on things like overstock. It's noise.

Let's say you run a liquor store, and a Thursday night college football game that looked like it would be the Boring Game of the Week when you were scheduling your purchases in September has suddenly in November turned into one with championship implications. You may be caught short if you scheduled your purchases to arrive at your store the following Saturday, when the signal in the baseline leads you to expect your sales to peak. That's *noise* — the difference between what you predict and what actually happens. By definition, noise is unpredictable, and for a forecaster it's a pain.

If the noise is random, it averages out. Some months, sporting goods stores will be discounting ski boots for less than the cost of an arthroscopy. Some months, a new and really cool model will come out, and the stores will take every possible advantage. The peaks and valleys even out. Some weeks there will be an extra football game or two and you'll sell (and therefore need) more bottles of beer. Some weeks there'll be a dry spell from Monday through Friday, you won't need so much beer, and you won't want to bear the carrying costs of beer you're not going to sell for a while.



TIP

The idea is that the noise averages out, and that what moving averages show you is the signal. To misquote Johnny Mercer, if you accentuate the signal and eliminate the noise, you latch on to a pretty good forecast.

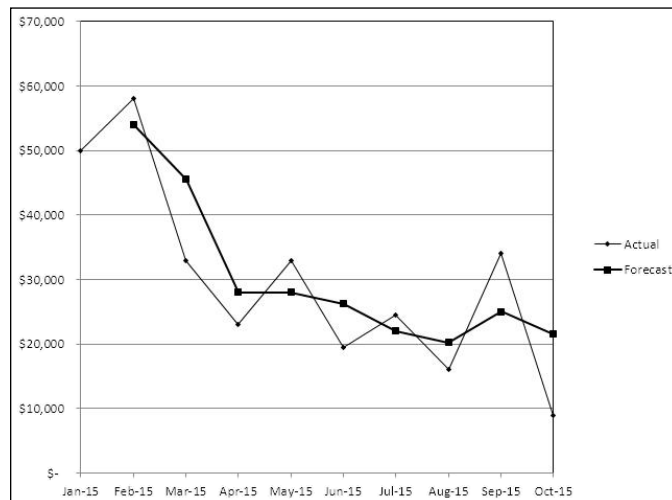
So with moving averages, you take account of the signal — the fact that you sell more ski boots during certain months and fewer during other months, or that you sell more beer on weekends than on weekdays. At the same time you want to let the random noises — also termed *errors* — cancel one another out. You do that by averaging what’s already happened in two, three, four, or more previous consecutive time periods. The signal in those time periods is emphasized by the averaging, and that averaging also tends to minimize the noise.

Suppose you decide to base your moving averages on two-month records. That is, you’ll average January and February, and then February and March, and then March and April, and so on. In that case you’re getting a handle on the signal by averaging two consecutive months and reducing the noise at the same time. Then, if you want to forecast what will happen in May, you hope to be able to use the signal — that is, the average of what’s happened in March and April.

Figure 1-2 shows an example of the monthly sales results and of the two-month moving average.

Chapter 14 goes into more detail about using moving averages for forecasting.

**FIGURE 1-2:** The moving average shows the general direction of the sales (the signal), and deemphasizes the random variations (the noise).



## Making sense of exponential smoothing

I know, the term *exponential smoothing* sounds intimidating and pretentious. I guess it's both — although I promise I'm not responsible for it. (If you really want, you can find out why it's called that in Chapter 15.) In any event, don't worry about what it's called — it's just a kind of self-correcting moving average.

Suppose that in June, you forecast \$100,000 in sales for July. When the July sales results are in, you find that your July forecast of \$100,000 was \$25,000 too low — you actually made \$125,000 in sales. Now you need to forecast your sales for August. The idea behind this approach to forecasting is to adjust your August forecast in a way that would have made the *July* forecast more accurate. That is, because your July forecast was too low, you increase your August forecast above what it would have been otherwise.

More generally:

- » If your most recent forecast turned out to be an underestimate, you adjust your next forecast upward.
- » If your most recent forecast turned out to be an overestimate, you adjust your next forecast downward.

You don't make these adjustments just by guessing. There are formulas that help out, and the Data Analysis add-in's Exponential Smoothing tool can enter the formulas for you. Or you can roll your own formulas if you want. Turn to Chapter 15 to see how to do that.

Figure 1-3 shows what you would forecast if your prior forecast (for July) was too low — then you boost your forecast for August.

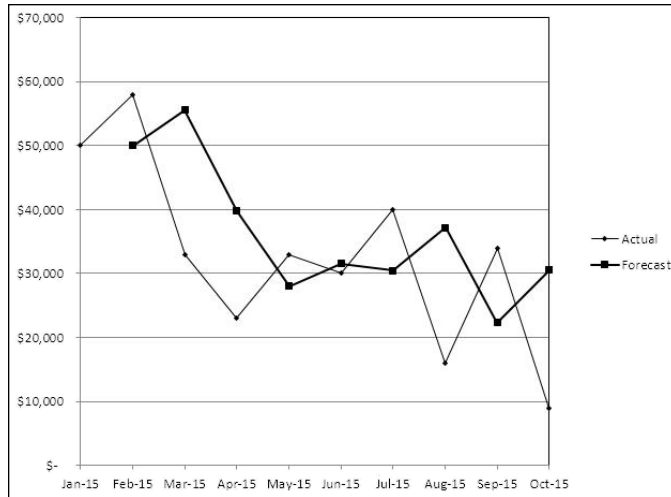
And if your prior, July forecast was too high, you cool your jets a little bit in your August forecast, as shown in Figure 1-4.

## Using regression to get what you want

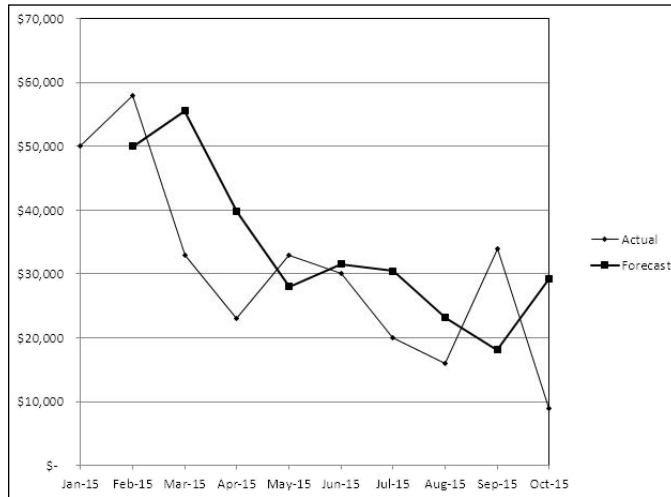
The term *regression* doesn't sound as bad as *exponential smoothing*, but it is — I admit — more complicated, at least in terms of the math.

And that's why the Regression tool in the Data Analysis add-in is convenient. The add-in takes responsibility for the math, just as it does with moving averages and exponential smoothing. **Remember:** You still have to give a good baseline to the tools in the Data Analysis add-in to get accurate results.

**FIGURE 1-3:**  
Here's what happens if your forecast for July was an underestimate. Notice that the August forecast is kicked up.



**FIGURE 1-4:**  
Your forecast for March 2015 was too high, so exponential smoothing makes you back off your forecast for April 2015.



Here's a quick look at forecasting with regression. (You can find a more detailed look in Chapter 11.)

The idea behind regression is that one variable has a relationship with another variable. When you're a kid, for example, your height tends to have a relationship to your age. So if you want to forecast how tall you'll be next year — at least, until you quit growing — you can check how old you'll be next year.

Of course, people differ. When they're 15 years old, some people are 5 feet tall, some are 6 feet tall. On average, though, you can forecast with some confidence

how tall someone will be at age 15. (And you can almost certainly forecast that a newborn kidlet is going to be under 2 feet tall.)

The same holds true with sales forecasting. Suppose your company sells consumer products. It's a good bet that the more advertising you do, the more you'll sell. At least it's worth checking out whether there's a relationship between the size of your advertising budget and the size of your sales revenue. If you find that there's a dependable relationship — and if you know how much your company is willing to spend on advertising — you're in a good position to forecast your sales.

Or suppose your company markets a specialty product, such as fire doors. (A *fire door* is one that's supposed to be resistant to fire for some period of time, and there are a lot of them in office buildings.) Unlike consumer products, something such as a fire door doesn't have to be a particular off-the-shelf color or have a fresher-than-fresh aroma. If you're buying fire doors, you want to get the ones that meet the specs and are the cheapest.

So if you're selling fire doors, as long as your product meets the specs, you'd want to have a look at the relationship between the price of fire doors and how many are sold. Then you check with your marketing department to find out how much they want you to charge per door, and you can make your forecast accordingly.



TIP

The point is that more often than not you can find a dependable relationship between one variable (advertising dollars or unit price) and another (usually, sales revenue or units sold).

You use Excel's tools to quantify that relationship. In the case of regression forecasts, you give Excel a couple of baselines. To continue the examples used so far in this section:

- »» Historical advertising expenses and historical sales revenues
- »» How much you charged per fire door and how many doors you sold

If you give Excel good baselines, it will come back to you with a formula.

- »» Excel will give you a number to multiply times how much you expect to spend on advertising, and the result will be your expected sales revenue.
- »» Or, Excel will give you a number to multiply times the unit cost per door, and the result will be the number of doors you can expect to sell.



TECHNICAL  
STUFF

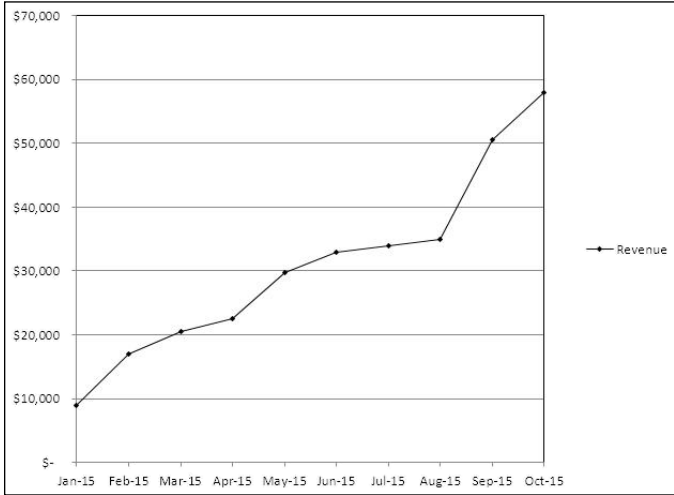
It's just a touch more complicated than that. Excel also gives you a number, called a *constant*, that you need to add to the result of the multiplication. But as Chapter 11 shows, you can get Excel to do that for you.

# Charting Your Data

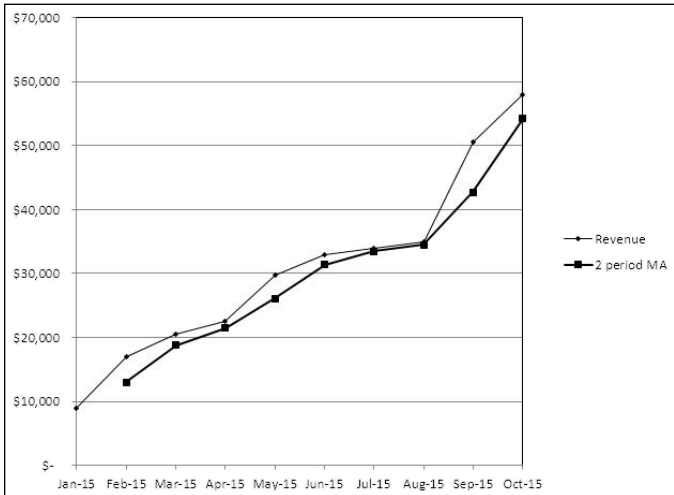
I've been doing this stuff for a long time, and I can't tell you how critical it is to chart your baseline and your forecast. Being able to visualize what's going on is important for several reasons.

Using Excel's charts, you can see how your actuals are doing (see Figure 1-5). And by charting your actuals, you can see how well your sales forecasts do against the actual sales results. Figure 1-6 shows a forecast that's based on moving averages, against the monthly actuals.

**FIGURE 1-5:**  
An Excel chart makes it much easier to see how your sales are doing.



**FIGURE 1-6:**  
Notice how the moving average lags behind the actual results.



By charting your baseline and your forecasts, you can:

- » **See how your actual results are doing.** A chart is almost always more revealing than a table of numbers.
- » **See how well your forecasts predict actual results.** Your eye is a good gauge of the quality of your forecasts.
- » **See how well a different variable — advertising dollars or the Consumer Price Index — predicts the sales of your product.**

Yes, an R squared or some other summary statistic can give you a concise estimate of how well your forecasts are working. But there's nothing, *nothing*, like a chart to tell you if you're forecasting results or if you're forecasting junk. Chapter 9 shows you how to set up charts with Excel.

## Forecasting with Advanced Tools

There's a lot to be said for using the Data Analysis add-in to create your forecasts. The add-in's tools are quick, they do the heavy lifting for you, and they're reasonably comprehensive, taking care of the math and some of the charting.

But there's nothing like doing it yourself. When you wave goodbye to the Data Analysis add-in, you establish and maintain control over what's going on with the forecast. If you have formulas in your worksheet cells — formulas that support your forecasts — you can change those formulas as your forecasting needs change. And you can change — or add to — the baseline and immediately see what the effect doing so has on your forecast. That's because the formulas are live: They react to changes in their inputs.

When the add-in's tools give you not formulas but static values instead, you can't easily experiment with the forecasts or see the effect of modifying the baseline. And the add-in's Regression tool gives you just the static values. The Exponential Smoothing tool is a little better, but it mixes formulas with static values. And the Moving Averages tool forces you to start from scratch if you want to change the number of records in the baseline that make up a moving average.

Suppose that you have the number 3 in cell A1 and the number 5 in cell A2. In cell A3 you can enter the sum of those two numbers, 8. But if you now change the number 3 in cell A1 to, say, 103, you still have 8 in A3. It's a constant — a number, not a formula. It doesn't react to what's in cell A1 or A2: You're still going to see the number 8 in cell A3.



On the other hand, suppose you have this in cell A3:

```
=A1 + A2
```

That's a formula, not a constant, and it tells Excel to add whatever's in A1 to whatever's in A2. So if you change what's in A1, or what's in A2, Excel recalculates the result and shows it — in this example — in A3.

The point to keep in mind is that the add-in's regression tool gives you numbers, not formulas. It calculates your forecast, and the underlying figures, and writes numbers onto your worksheet. That means, regardless of how you change the numbers in your baseline, you're still going to be looking at the same forecast as offered by the Regression tool.

But — and it's a big one — if you make the forecast yourself instead of relying on the add-in's tool, you can enter the formulas that the add-in denies you. Why is this important? By entering the formulas yourself, you have more control over what's going on with the forecast.

Relying on the add-in, which isn't a bad toolbox, and is one that you can generally trust, is perfectly okay. However, if you enter formulas, ones that react to changes in your baseline, you can make a change in the baseline and see what happens to the forecast. You can change this month's result from \$100,000 to \$75,000 and see whether your forecast for next month changes substantially. You can't do that with the add-in's Regression tool unless you start all over again, because it doesn't give you formulas. To a smaller degree, the same is true of the Exponential Smoothing tool.

But the more important reason, the reason for you to consider entering the formulas yourself, is that you're relying on your own knowledge of how and why forecasting works. In Part 4, I show you how to use functions like `LINEST` and `TREND` to do your regression-based forecasts. You also see how to use array formulas to get the most out of those Excel functions.

You don't need to enter all the formulas yourself to make good forecasts. The add-in includes reasonably good tools. But if you do enter the formulas yourself, not only can you be more confident that you know what's going on with your forecast, but you can also exercise more control over what your forecast says is going to happen. In a business as tricky and trappy as forecasting, the more control you have, the better.



## Chapter 2

# Forecasting: The Basic Issues

**U**nless you really enjoy playing with numbers, you need a good reason to bother with forecasting sales. In this chapter, I tell you some of the business reasons to forecast, beyond the fact that your Vice President of Sales makes you do it.

Like all specialties, forecasting uses terms that are unfamiliar to those who haven't yet been inducted into the secret society. This chapter introduces you to some of the important sales forecasting terminology.

If you're going to make a credible forecast, you need access to an archive of historical data that isn't necessarily easy to access. You'll often find it right there in an Excel workbook, but sometimes it isn't there; instead, it's in your company's accounting database, and someone will have to exhume it. In this chapter, you see some of the reasons to put yourself or your assistant through that task.

Excel offers several methods of forecasting. Each method works best — and some work *only* — if you set up a baseline using what Excel terms a *table*. Depending on the method you choose, that table may occupy only one column, or two (or more) columns. This chapter gives you an overview of those forecasting methods, along with a brief explanation of why you might use just one column of data for your baseline, or two or more columns, depending on your choice of forecasting method.

Excel is an ideal general-purpose analysis program to use for forecasting, in part because it has functions and tools that are intended to help you make your forecasts, and in part because you often store the necessary data in Excel anyway — so, it's right there, ready for you to use. In this chapter, you find out what's so great about using Excel to create your forecasts, and you find some groundwork on how best to put it to use in your own situation.

## Why Forecast?

People tend to think of the process of sales forecasting as a knee-jerk response to a frantic call for reassurance from some nervous, jumpy, excitable VP who's worried about having to dust off the résumé. And often, you have some reason to believe that's *exactly* what's going on.

But there are plenty of more productive reasons to go to the trouble of gathering up baseline data, getting it into the right shape to support a credible forecast, do the analysis, and then interpret it than just responding to a VP who's afraid the job is on the line. Here are a few of those reasons.

### To plan sales strategies

If you can use sales forecasts to get a handle on either future revenues, or unit sales, or both, you can help groups like Marketing, Product Management, and Production make decisions about activities such as promotion, pricing, and purchasing — each of which influences your company's sales results as well as its net income.

Suppose you take a look at quarterly sales results over a period of several years, and you see that during that time the sales of a particular product have been gently declining. (If the decline had been steep, you wouldn't have to look at a baseline — everyone from the sales force to the CEO would have been rattling your cage.) Your forecast indicates that the decline is likely to continue. Is the market for the product disappearing? That depends. You need to ask and answer some other questions first.

» **Is the product a commodity?** Some business analysts sneer at commodities — they're not very glamorous, after all — but commodities can be very profitable products if you dominate the market. If you don't dominate the market, maybe you shouldn't be in the market for that commodity. So, have your competitors been cutting into your market share, or is the total size of the market shrinking? If the problem is the competition, maybe you want to do

something to take back your share, even if that requires putting more resources into the product line — such as retooling its manufacture, putting more dollars into promotions, or cutting the price. But if the total market itself is shrinking, it may just be time to bail out.

- » **How old is the product?** Products do have life cycles. When products are bright and shiny, the sales revenues can grow sharply over a fairly short time frame. When products reach maturity, the sales usually flatten out. And then, as newer, better, fancier products arrive, the sales start to drop. Think streaming video versus DVD. Get Marketing and Product Management to assess whether the product is getting long in the tooth. If it is, it may be time to get out. Or, it may be smart to spruce up the product and differentiate it from the competition's versions, in order to squeeze some more profitable revenue out of it before you give up on it. Forecasting can inform that kind of decision, although it can't make it for you.
- » **How will Sales support the product?** If your company decides that it's not yet time to abandon the product, Sales Management needs to make some decisions about how to allocate its resources — that is, its sales reps. One way to do that, of course, is to take the product out of some reps' bags and replace it with another, more robust product. (Keep in mind that some reps *prefer* older products because they can use familiar sales strategies.)
- » **Is it possible that the decline in sales is due more to large-scale economic conditions than to problems with the product itself?** If so, you may decide to hang in and wait for the economy, consumer confidence, or the index of leading economic indicators to improve, instead of making a drastic decision to drop a product line.



REMEMBER

There's at least one good aspect to a product that's entering the final stage of its life cycle: You very likely have lots of historical data on its sales figures. And in general, the more historical data you have to base a forecast on, the more confidence you can place in that forecast.

## To size inventories

During the late 1980s, I worked for a Baby Bell — one of the companies that was spun off by the AT&T breakup. For a couple of years, I was in charge of managing resale equipment inventories at that Baby Bell.

My staff and I reduced the size of the equipment intended for sale to customers from a grotesque \$24 million to a more reasonable \$9 million in 18 months, without resorting to write-downs. We did it by forecasting sales by product line. This helped us tell which products we could expect to have high *turns ratios* (the speed with which the product line would sell) and we'd buy those in quantities that increased our discounts from our suppliers.

Until we were almost out of them, we refused to buy any products that our forecasts indicated would have low turns ratios. It didn't matter how piteous the pleadings of the sales managers who wanted them on hand for fast delivery just in case a customer decided to buy one and wanted it installed right now. (Getting a huge PBX out of warehouse storage in West Eyesocket, Connecticut, and shipping it to Broken Pelvis, Montana, can take longer than you may think. For one thing, you may have to pressure Connecticut's Regional VP into letting go of it. Today, VoIP software is rapidly replacing big electronic switches, but the principle remains the same: Expensive stuff can be hard to move.)

Plus, the annual carrying costs for equipment inventory in the late 1980s averaged around 15 percent of the cost of the equipment, including storage, cost of money, obsolescence, and so on. So by reducing the total inventory cost by \$15 million, we saved the company \$2.25 million each year. (That savings actually covered the cost of our salaries, by the way, with plenty left over.)



TIP

Simply reducing the size of inventory isn't the end of the story, though. Sales forecasting helps you plan just-in-time (JIT) inventory management, so you can time your purchases to correspond to when sales need to be fulfilled. The less time inventory spends in the warehouse, the less money you're paying to let it just sit there waiting to be sold.

## Talking the Talk: Basic Forecasting Lingo

You need to get a handle on the specialized terminology used in forecasting for a couple very practical reasons. One is that you may be asked to explain your forecasts to your boss or in a meeting of, for example, sales managers. In those situations you want to say things like, "We decided to use regression on the baseline because it turned out to be more accurate." You *don't* want to find yourself saying "Jeff found a formula in a book he has, and we used it on these numbers here. Seems to work okay."

Another good reason is that Excel uses many of these terms, as do other programs, and figuring out what's going on is a lot easier if you know what the terms mean. Okay, deep breath.

### Autoregressive integrated moving averages (ARIMA)

I mention autoregressive integrated moving averages (ARIMA) here not because this book is going to use it or even talk much about it. But if you're going to do

forecasting, some smart aleck will eventually ask you if you used ARIMA, and you should know how to reply. ARIMA is in part a forecasting method, and also a way of evaluating your baseline so that you can get quantitative evidence that supports using a regression approach, a moving-average approach, or a combination of both. Unless you really take to this forecasting stuff, you'll usually do just fine without it, even though it's an excellent, if complex, diagnostic tool.

By the way, your answer to the smart aleck should be, "No. I've been working with this baseline for so long now that I know I get my best results with exponential smoothing. Which, as you know, is one of the forms that ARIMA can take."

## Baseline

A *baseline* is a sequence of data arranged in chronological order. In terms of this book's basic topic, the forecasting of sales, some examples of baselines include total monthly revenues from January 2010 through December 2015, number of units sold weekly from January 1, 2015, through December 31, 2016, and total quarterly revenues from Q1 2007 through Q4 2016. Data arranged like this is sometimes called a *time series*, but in this book I use the term *baseline*.

## Correlation

A *correlation* coefficient expresses how strongly two variables are related. Its possible values range from  $-1.0$  to  $+1.0$ , but in practice you never find correlations so extreme. The closer a correlation coefficient is to  $\pm 1.0$ , the stronger the relationship between the two variables. A correlation of  $0.0$  means no relationship. So, you might find a correlation of  $+0.7$  (fairly strong) between the number of sales reps you have and the total revenue they bring in: The greater the number of reps, the more that gets sold. And you might find a correlation of  $-0.1$  (quite weak) between how much a rep sells and his telephone number.

A special type of correlation is the *autocorrelation*, which calculates the strength of the relationship between one observation in a baseline and an earlier observation (often, but not always, the relationship between two consecutive observations). The autocorrelation tells you the strength of the relationship between what came before and what came after. This in turn helps you decide what kind of forecasting technique to use. Here's an example of how to calculate an autocorrelation that might make the concept a little clearer:

```
=CORREL (A2 : A50, A1 : A49)
```

This Excel formula uses the CORREL function to show how strong (or how weak) a relationship there is between whatever values are in A2:A50 and those in A1:A49. The most useful autocorrelations involve baselines that are sorted in chronological

order. (This sort of autocorrelation is not quite the same as the autocorrelations calculated in ARIMA models.)

## Cycle

A *cycle* is similar to a seasonal pattern (see the “Seasonality” section, later in this chapter), but you don’t consider it in the same way as you do seasonality. The upswing might span several years, and the downswing might do the same. Furthermore, one full cycle might take four years to complete, and the next one just two years. A good example is the business cycle: Recessions chase booms, and you never know just how long each is going to last. In contrast, yearly seasons have the same length, or nearly so.

## Damping factor

The *damping factor* is a fraction between 0.0 and 1.0 that you use in exponential smoothing to determine how much of the error in the prior forecast will be used in calculating the next forecast.



TECHNICAL  
STUFF

Actually, the use of the term *damping factor* is a little unusual. Most texts on exponential smoothing refer to the *smoothing constant*. The damping factor is 1.0 minus the smoothing constant. It really doesn’t matter which term you use; you merely adjust the formula accordingly. This book uses *damping factor* where necessary because it’s the term that Excel’s Data Analysis add-in uses.

## Exponential smoothing

Stupid term, even if technically accurate. Using *exponential smoothing*, you compare your prior forecast to the prior *actual* (in this context, an *actual* is the sales result that Accounting tells you — after the fact — that you generated). Then you use the error — that is, the difference between the prior forecast and the prior actual — to adjust the next forecast and, you hope, make it more accurate than if you hadn’t taken the prior error into account. In Chapter 15, I show you how really intuitive an idea this is, despite its pretentious name.

## Forecast period

The *forecast period* is the length of time that’s represented by each observation in your baseline. The term is used because your forecast usually represents the same length of time as each baseline observation. If your baseline consists of monthly sales revenues, your forecast is usually for the upcoming month. If the baseline consists of quarterly sales, your forecast is usually for the next quarter. Using the



regression approach, you can make forecasts farther into the future than just one forecast period, but the farther your forecast gets from the most recent actual observation, the thinner the ice.

## Moving average

You've probably run into the concept of moving averages somewhere along the line. The idea is that averaging causes noise in the baseline to cancel out, leaving you with a better idea of the *signal* (what's really going on over time, unsullied by the inevitable random errors). It's an *average* because it's the average of some number of consecutive observations, such as the average of the sales in January, February, and March. It's *moving* because the time periods that are averaged move forward in time — so, the first moving average could include January, February, and March; the second moving average could include February, March, and April; and so on.

There's no requirement that each moving average include three values — it could be two, or four, or five, or conceivably even more. (Chapter 13 fills you in on the effects of choosing more or fewer periods to average.)

## Predictor variable

You generally find this term in use when you're forecasting with regression. The *predictor variable* is the variable you use to estimate a future value of the variable you want to forecast. For example, you may find a dependable relationship between unit sales price and sales volume. If you know how much your company intends to charge per unit during the next quarter, you can use that relationship to forecast the sales volume for next quarter. In this example, unit sales price is the predictor variable.

## Regression

If you use the *regression* approach to sales forecasting, it's because you've found a dependable relationship between sales revenues and one or more predictor variables. You use that relationship, plus your knowledge of future values of the predictor variables, to create your forecast.

How would you know those future values of the predictor variables? If you're going to use unit price as a predictor, one good way is to find out from Product Management how much it intends to charge per unit during each of the next, say, four quarters. Another way involves dates: It's entirely possible, and even common, to use dates (such as months within years) as a predictor variable. Even I can figure out what the next date value is in a baseline that at present ends at November 2015.

## Seasonality

During the span of a year, your baseline might rise and fall on a seasonal basis. Perhaps you sell a product whose sales rise during warm weather and fall during cold. If you can see roughly the same pattern occur within each year over a several-year period, you know you're looking at *seasonality*. You can take advantage of that knowledge to improve your forecasts. It's useful to distinguish seasons from cycles. You never know how long a given cycle will last. But each of four seasons in a year is three months long.

## Trend

A *trend* is the tendency of the level of a baseline to rise or fall over time. A rising revenue trend is, of course, good news for sales reps and sales management, to say nothing of the rest of the company. A falling baseline of sales, although seldom good news, can inform Marketing and Product Management that they need to make and act on some decisions, perhaps painful ones. Regardless of the direction of the trend, the fact that a trend exists can cause problems for your forecasts in some contexts — but there are ways of dealing with those problems. Chapter 17 shows you some of those ways.

# Understanding the Baseline

A *baseline* is a series of observations — more to the point in this book, a *revenue stream* — that you use to form a forecast. There are three typical forecasts, depending on what the baseline looks like:

- » **If the baseline has held steady**, your best forecast will probably be close to the average of all the sales amounts in the baseline.
- » **If the baseline has been rising**, your forecast will likely be higher than the most recent sales amount.
- » **If the baseline has been falling**, your forecast will probably be lower than the most recent sales amount.

*Note:* Those weasely words *likely* and *probably* are there because when there's a seasonal aspect to the sales that doesn't yet appear in your baseline, the next season might kick in at the same point as your forecast and reverse what you'd expect otherwise.



REMEMBER

Why is a baseline important? Because it elevates your forecast above the status of a guess. When you use a baseline, you recognize that — absent special knowledge such as the fact that your per-unit price is about to change drastically — your best guide to what happens next is often what happened before.

There's another weasel word: *often*. You'll have plenty of opportunities to use one variable, such as the total of sales estimates from individual sales representatives, to forecast the variable you're really interested in, sales revenues. In that case, you might get a more accurate forecast by using Excel to figure the formula that relates the two variables, and then use that formula to forecast the next value of sales revenues.

Depending on the strength of the relationship between the two variables, that formula can be a better guide than looking solely to the baseline of sales history. It's still a baseline, though: In this case, the baseline consists of two or more variables, not just one.

## Charting the baseline

The eye is a great guide to what's going on in your baseline. You can take advantage of that by making a chart that shows the baseline. There are a couple of possibilities:

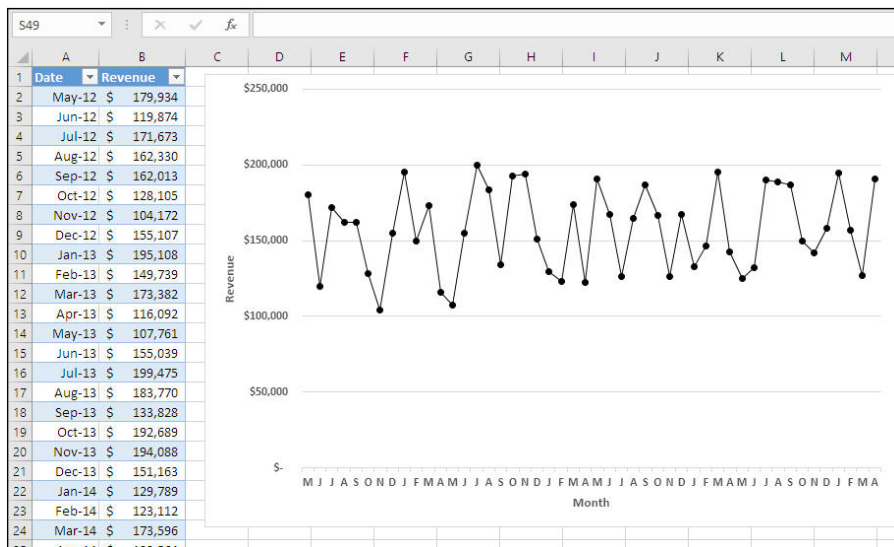
- » **If you're making your forecast solely on the basis of previous sales revenues**, a good choice is a Line chart, like the one shown in Figure 2-1. You can see that the revenues are flat over time, even though they jump around some. The baseline's pattern in the chart is a clue to the type of forecast to use: In Figure 2-1, that type could be exponential smoothing.
- » **If you're using another variable** — such as the total of the sales estimates provided by individual sales reps — you'd probably use an XY (Scatter) chart, like the one shown in Figure 2-2. Notice that the actuals track fairly well against the sum of the individual estimates, which may convince you to use the regression approach to forecasting the next period, especially because you can get your hands on the next estimate from the sales force to forecast from.



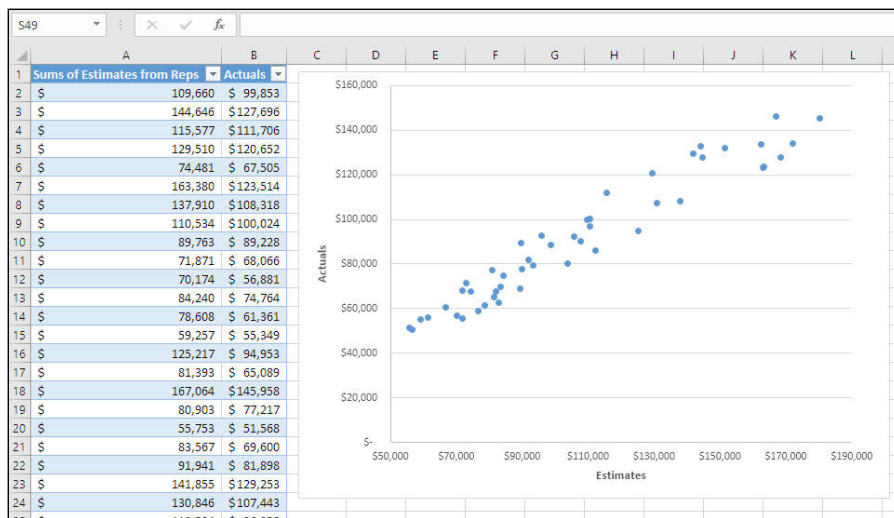
TIP

If you're going to base your next forecast on information from individual sales reps, don't make your forecast periods too short. If you do, you'll have the reps spending more time making estimates than making sales, which means their commissions decline, and the next thing you know they're working for your competition — and you can flush your forecast down the toilet.

**FIGURE 2-1:**  
The Line chart is ideal for just one variable, such as sales revenues.



**FIGURE 2-2:**  
In this case, a positive relationship exists between the sum of individual estimates and the actual results.

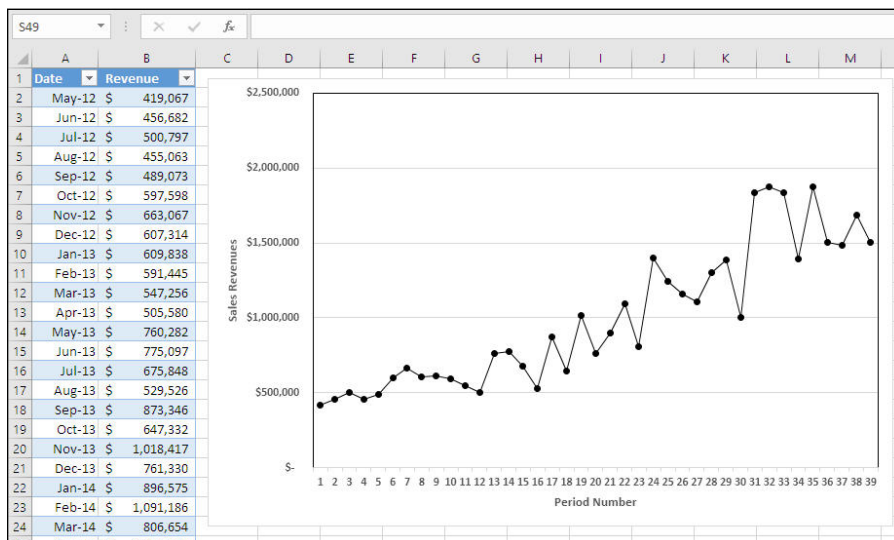


## Looking for trends

Trends are important in sales forecasting. For one thing, knowing if there's a trend in your baseline is critical to knowing more about what's going on in the product line. For another, the presence of a trend sometimes tells you that you have to do more preparation. If you've decided to use exponential smoothing, for example, you may want to remove the trend first (see Chapter 17 for more information on trends).

If you pored over the baseline shown in column A of Figure 2-3, it wouldn't take you long to conclude that there's an upward trend in the sales revenues. But if you exert the tiny bit of effort needed to chart the baseline, also shown in Figure 2-3, not only do you immediately see the trend, but you get a good intuitive idea of where the sales are headed and how fast they're getting there.

**FIGURE 2-3:**  
You could either detrend this series and use simple exponential smoothing, or forecast sales revenues using the period number as the predictor.



WARNING

Be careful when you see a trend such as the one shown in Figure 2-3. If these are weekly results, it may just be the first part of a seasonal pattern (or a cycle) that's about to head back down. Notice that the final seven periods look as though the results may be getting ready to do just that.

## Setting Up Your Forecast

The most straightforward way of getting a forecast is to lay out your baseline on a worksheet in a table configuration (see Chapter 6) and then call on the Data Analysis add-in to generate a forecast for you. That add-in accompanies Microsoft Office. You can find information about installing the add-in in Chapter 7.

The add-in and its tools are good news and bad news — more good than bad, actually. It hasn't changed substantially since Excel 1995, except that now the code is written using Visual Basic rather than the old weird Excel 4.0 macro language. It can be quirky, as you'll see if you decide to use it. And I think you should decide to use it, because, despite its quirks, it can save you some time. It can serve

as a reasonably good springboard for learning how to do it all yourself. And it can spare you the errors that inevitably occur when you roll your own forecasts (at least, they inevitably occur when I roll my own).

The add-in has 19 different numeric and statistical analysis tools. If you lay out your data in the right way, you can point one of its tools at your data and get a fairly complete and usually correct analysis — including autocorrelation analyses, moving-average forecasts, exponential-smoothing forecasts, and regression forecasts. It does the hard work for you, and because it's all precoded, you don't need to worry so much about, say, getting a formula wrong.

## Smoothing data

If you decide to use exponential smoothing to create your forecast (I help you make that decision in Chapter 10), all you'll need is your baseline of historic sales revenues. Each observation in the baseline should be from the same sort of forecast period — as often as not, revenue totals on a monthly basis.

You need no variable other than your sales results because, using smoothing, you're going to use one period's result to forecast the next — which is one reason you'll use the Data Analysis add-in's Correlation tool to determine the amount of autocorrelation in the baseline before you do the forecast. Substantial autocorrelation will tend to lead you toward using the Exponential Smoothing tool as your forecasting method — and it will help you determine what damping factor (or, equivalently, what smoothing constant) to use in developing your forecast.

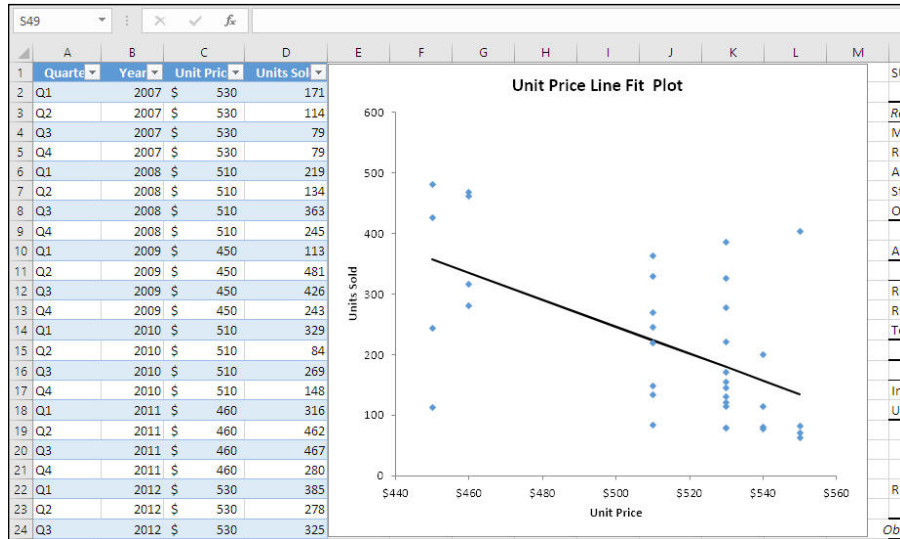
## Regression: It's all about relationships

If you have available some variable in addition to sales revenues or units sold, and you suspect that it's strongly related to the sales results, you should take a closer look at the relationship.

Suppose you can lay your hands on historical data that shows — by year and month, say — the unit price that you've charged and the number of units you've sold. If you're interested in forecasting the number of units you'll sell next month, the Data Analysis add-in's Regression tool can ease your task, as shown in Figure 2-4.

(In Figure 2-4, I modified the appearance of the chart as the Regression tool creates it to make it easier to gauge the relationship between price and volume. You can see how to do this in Chapter 11.)

**FIGURE 2-4:**  
The chart gives you a visual of what's going on between the two variables: Unit Price and Units Sold.



With this baseline, including unit price and units sold, your interest doesn't focus on revenues. After all, it's pretty clear from the chart that the higher the unit price, the fewer the units sold — and that will tend to minimize the variation in quarterly revenue. Instead, this analysis speaks to production. If you know how you'll set your unit price for next quarter, you can use the Regression tool to forecast the number of units you'll sell next quarter. That forecast might well inform your Production department about how to allocate its resources.



By the way, Excel terms the solid line shown in Figure 2-4 a *trendline*. When you see a trendline run from the upper left to the lower right, as in Figure 2-4, you know that the correlation between the two variables is negative (and in this case, the correlation between unit price and units sold is  $-0.57$ ). A *negative correlation* means that the higher the level of one of the variables, the lower the corresponding value of the other variable. If the trendline runs from the lower left to the upper right, you know that the correlation is positive. A *positive correlation* means that lower values on one variable are associated with lower values on the other, and that higher values on one are associated with higher values on the other.

## Using Your Revenue and Cost Data

If your ecological niche is in your company's sales management chain, it's likely that you keep a record in Excel of the company's sales results, or you have someone keep it for you. If your products are doing well, you like to fire up Excel, look at those results, and sigh contentedly. If your products aren't performing as they

should, you reflexively fire up Excel, look at those results, and wonder how to fix what's gone wrong. Either way, the data is probably in an Excel workbook already, and that makes doing your forecasting easy.

The novelist Rex Stout once had a character complain to his boss, with heavy sarcasm, “When I meekly mention that the science of bookkeeping has two main branches, first addition, and second subtraction . . .” Your interest may very well focus on the addition branch — that is, sales revenues. But profits are really what it's about, and you may want to take a look at profit forecasts. In that case, you also need to look at the subtraction branch — that is, cost of sales.

Building revenue baselines is pretty straightforward. You decide on your *forecast period* (the length of time that's represented by each observation in your baseline) and if necessary use Excel to figure the total revenue for each period in your baseline. Now you're ready to forecast how much you'll bring in during the upcoming period.

Figuring costs, particularly cost of sales, is trickier. Should you include the cost of goods sold? Of course, you'll include the costs of sales reps' salaries and commissions, of product promotions, of leave-behinds, of trash-and-trinkets, of travel and entertainment. What about indirect costs?

One way to proceed is to subtract those costs, for each period in your baseline, from the revenues for the same period to get a profit estimate for the period. This creates a new, profit baseline that you can use to forecast the next period's profit.

But what about opportunity costs? When you spend money to support the sales of one product line, you're diverting resources from another product line. That's an opportunity cost: You had the opportunity to spend that money in support of Gidgets rather than Widgets, and you might have created more revenue and more profit if you'd used the money to help sell Widgets.

No rule of accounting tells you whether to include opportunity costs in your profit calculations. Here, you're not *doing* accounting, you're *using* it to help lay your plans.

Figure 2-5 illustrates how forecasting can help you plan how to support product lines.

In Figure 2-5, the data on Widgets and Gidgets is pretty straightforward. The figure shows the actual revenue and the actual direct costs of supporting each product line during each month from January 2014 through July 2016, in row 3 through row 33. The worksheet gets the profit figures simply by subtracting the costs from the revenues.



F35									
	A	B	C	D	E	F	G	H	
1	Sales Date	Widgets			Gidgets				
2		Revenue	Cost	Profit	Revenue	Cost	Profit		
3	Jan-14	\$ 42,597	\$ 39,928	\$ 2,669	\$ 32,074	\$ 28,712	\$ 3,362		
4	Feb-14	\$ 42,127	\$ 33,368	\$ 8,759	\$ 32,869	\$ 27,071	\$ 5,798		
5	Mar-14	\$ 42,149	\$ 36,961	\$ 5,188	\$ 30,969	\$ 28,633	\$ 2,336		
6	Apr-14	\$ 49,931	\$ 39,494	\$ 10,437	\$ 30,437	\$ 26,565	\$ 3,872		
7	May-14	\$ 32,456	\$ 34,591	\$ (2,135)	\$ 34,405	\$ 27,901	\$ 6,504		
8	Jun-14	\$ 32,418	\$ 31,241	\$ 1,177	\$ 33,146	\$ 27,633	\$ 5,513		
9	Jul-14	\$ 41,145	\$ 39,917	\$ 1,228	\$ 34,046	\$ 28,943	\$ 5,103		
10	Aug-14	\$ 42,301	\$ 36,917	\$ 5,384	\$ 33,644	\$ 25,359	\$ 8,285		
11	Sep-14	\$ 46,239	\$ 34,729	\$ 11,510	\$ 31,548	\$ 26,885	\$ 4,663		
12	Oct-14	\$ 44,551	\$ 31,316	\$ 13,235	\$ 30,885	\$ 25,249	\$ 5,636		
13	Nov-14	\$ 30,346	\$ 34,504	\$ (4,158)	\$ 33,456	\$ 25,091	\$ 8,365		
14	Dec-14	\$ 34,663	\$ 38,756	\$ (4,093)	\$ 34,772	\$ 25,515	\$ 9,257		
15	Jan-15	\$ 49,297	\$ 35,631	\$ 13,666	\$ 34,757	\$ 27,213	\$ 7,544		
16	Feb-15	\$ 43,048	\$ 39,893	\$ 3,155	\$ 33,979	\$ 26,701	\$ 7,278		
17	Mar-15	\$ 30,915	\$ 28,513	\$ 2,402	\$ 33,116	\$ 28,501	\$ 4,615		
18	Apr-15	\$ 32,256	\$ 32,969	\$ (713)	\$ 34,340	\$ 25,751	\$ 8,589		
19	May-15	\$ 42,774	\$ 34,884	\$ 7,890	\$ 34,175	\$ 25,519	\$ 8,656		
20	Jun-15	\$ 42,456	\$ 33,729	\$ 8,727	\$ 32,810	\$ 25,292	\$ 7,518		
21	Jul-15	\$ 30,816	\$ 31,726	\$ (910)	\$ 33,538	\$ 28,948	\$ 4,590		
22	Aug-15	\$ 35,752	\$ 35,281	\$ 471	\$ 30,789	\$ 25,152	\$ 5,637		
23	Sep-15	\$ 42,581	\$ 31,034	\$ 11,547	\$ 33,970	\$ 26,771	\$ 7,199		
24	Oct-15	\$ 30,558	\$ 26,096	\$ 4,462	\$ 34,794	\$ 26,761	\$ 8,033		
25	Nov-15	\$ 35,397	\$ 34,742	\$ 655	\$ 32,328	\$ 25,428	\$ 6,900		
26	Dec-15	\$ 33,096	\$ 36,986	\$ (3,890)	\$ 34,564	\$ 26,332	\$ 8,232		
27	Jan-16	\$ 38,513	\$ 30,147	\$ 8,366	\$ 31,150	\$ 27,571	\$ 3,579		
28	Feb-16	\$ 37,458	\$ 37,797	\$ (339)	\$ 31,932	\$ 25,165	\$ 6,767		
29	Mar-16	\$ 45,450	\$ 30,604	\$ 14,846	\$ 32,344	\$ 26,023	\$ 6,321		
30	Apr-16	\$ 47,390	\$ 35,605	\$ 11,785	\$ 31,551	\$ 26,861	\$ 4,690		
31	May-16	\$ 30,946	\$ 32,511	\$ (1,565)	\$ 32,095	\$ 27,619	\$ 4,476		
32	Jun-16	\$ 49,998	\$ 37,388	\$ 12,610	\$ 31,014	\$ 26,570	\$ 4,444		
33	Jul-16	\$ 47,794	\$ 39,361	\$ 8,433	\$ 31,783	\$ 25,004	\$ 6,779		
35	Aug-16 Forecast	\$ 42,516	\$ 38,386	\$ 4,129	\$ 32,741	\$ 25,540	\$ 7,201		

**FIGURE 2-5:**  
The TREND worksheet function is based on linear regression — here, using the historical relationship between costs and revenues.

Row 35 shows forecasts for August 2016. Here’s how it gets them:

1. It forecasts the costs for August 2016 using exponential smoothing; see Chapter 15 for more information, but for those of you who are playing along at home, the smoothing constant has *not* yet been optimized by minimizing the mean square error.

The cost forecasts are shown in cells C35 and G35.

2. For each product, it forecasts the revenues for August 2016 by using the *regression* approach (where you use a dependable relationship between sales revenues and one or more predictor variables to make your forecast), in the guise of the TREND worksheet function.

Using information about the historic relationship between the costs and revenues for each product, it forecasts in cells B35 and F35 what the revenues would be, given the cost forecasts.

3. It forecasts the profits for August 2016 by subtracting the forecast cost from the forecast revenue.

Adding the forecast profits for both product lines results in a total profit for August 2016 of \$11,330.



TECHNICAL STUFF

Full disclosure: I'm using exponential smoothing for costs, and regression for revenues, simply to illustrate the methods. There's no special reason in this example to use two methods rather than one, or to choose those two particular methods.

Now, what if you took the opportunity costs of supporting Widgets into account, and instead poured them into Gidgets? In Figure 2-6, you can see the effect of abandoning Widgets and putting its costs — the resources your company spends supporting Widgets — into supporting Gidgets only.

1	A	B	C	D	E	F	G	H	I	J	K	L
2	Sales Date	Widgets			Gidgets			Support Gidgets Only				
3		Revenue	Cost	Profit	Revenue	Cost	Profit	Revenue	Cost	Profit		
3	Jan-14	\$ 42,597	\$ 39,928	\$ 2,669	\$ 32,074	\$ 28,712	\$ 3,362	\$ 76,677	\$ 68,640	\$ 8,037		
4	Feb-14	\$ 42,127	\$ 33,368	\$ 8,759	\$ 32,869	\$ 27,071	\$ 5,798	\$ 73,384	\$ 60,439	\$ 12,945		
5	Mar-14	\$ 42,149	\$ 36,961	\$ 5,188	\$ 30,969	\$ 28,633	\$ 2,336	\$ 70,945	\$ 65,594	\$ 5,351		
6	Apr-14	\$ 49,931	\$ 39,494	\$ 10,437	\$ 30,437	\$ 26,565	\$ 3,872	\$ 75,687	\$ 66,059	\$ 9,628		
7	May-14	\$ 32,456	\$ 34,591	\$ (2,135)	\$ 34,405	\$ 27,901	\$ 6,504	\$ 77,060	\$ 62,492	\$ 14,568		
8	Jun-14	\$ 32,418	\$ 31,241	\$ 1,177	\$ 33,146	\$ 27,633	\$ 5,513	\$ 70,620	\$ 58,874	\$ 11,746		
9	Jul-14	\$ 41,145	\$ 39,917	\$ 1,228	\$ 34,046	\$ 28,943	\$ 5,103	\$ 81,001	\$ 68,860	\$ 12,141		
10	Aug-14	\$ 42,301	\$ 36,917	\$ 5,384	\$ 33,644	\$ 25,359	\$ 8,285	\$ 82,622	\$ 62,276	\$ 20,346		
11	Sep-14	\$ 46,239	\$ 34,729	\$ 11,510	\$ 31,548	\$ 26,885	\$ 4,663	\$ 72,300	\$ 61,614	\$ 10,686		
12	Oct-14	\$ 44,551	\$ 31,316	\$ 13,235	\$ 30,885	\$ 25,249	\$ 5,636	\$ 69,191	\$ 56,565	\$ 12,626		
13	Nov-14	\$ 30,346	\$ 34,504	\$ (4,158)	\$ 33,456	\$ 25,091	\$ 8,365	\$ 79,463	\$ 59,595	\$ 19,868		
14	Dec-14	\$ 34,663	\$ 38,756	\$ (4,093)	\$ 34,772	\$ 25,515	\$ 9,257	\$ 87,589	\$ 64,271	\$ 23,318		
15	Jan-15	\$ 49,297	\$ 35,631	\$ 13,666	\$ 34,757	\$ 27,213	\$ 7,544	\$ 80,266	\$ 62,844	\$ 17,422		
16	Feb-15	\$ 43,048	\$ 39,893	\$ 3,155	\$ 33,979	\$ 26,701	\$ 7,278	\$ 84,746	\$ 66,594	\$ 18,152		
17	Mar-15	\$ 30,915	\$ 28,513	\$ 2,402	\$ 33,116	\$ 28,501	\$ 4,615	\$ 66,246	\$ 57,014	\$ 9,232		
18	Apr-15	\$ 32,256	\$ 32,969	\$ (713)	\$ 34,340	\$ 25,751	\$ 8,589	\$ 78,305	\$ 58,720	\$ 19,585		
19	May-15	\$ 42,774	\$ 34,884	\$ 7,890	\$ 34,175	\$ 25,519	\$ 8,656	\$ 80,892	\$ 60,403	\$ 20,489		
20	Jun-15	\$ 42,456	\$ 33,729	\$ 8,727	\$ 32,810	\$ 25,292	\$ 7,518	\$ 76,565	\$ 59,021	\$ 17,544		
21	Jul-15	\$ 30,816	\$ 31,726	\$ (910)	\$ 33,538	\$ 28,948	\$ 4,590	\$ 70,294	\$ 60,674	\$ 9,620		
22	Aug-15	\$ 35,752	\$ 35,281	\$ 471	\$ 30,789	\$ 25,152	\$ 5,637	\$ 73,977	\$ 60,433	\$ 13,544		
23	Sep-15	\$ 42,581	\$ 31,034	\$ 11,547	\$ 33,970	\$ 26,771	\$ 7,199	\$ 73,349	\$ 57,805	\$ 15,544		
24	Oct-15	\$ 30,558	\$ 26,096	\$ 4,462	\$ 34,794	\$ 26,761	\$ 8,033	\$ 68,723	\$ 52,857	\$ 15,866		
25	Nov-15	\$ 35,397	\$ 34,742	\$ 655	\$ 32,328	\$ 25,428	\$ 6,900	\$ 76,497	\$ 60,170	\$ 16,327		
26	Dec-15	\$ 33,096	\$ 36,986	\$ (3,890)	\$ 34,564	\$ 26,332	\$ 8,232	\$ 83,113	\$ 63,318	\$ 19,795		
27	Jan-16	\$ 38,513	\$ 30,147	\$ 8,366	\$ 31,150	\$ 27,571	\$ 3,579	\$ 65,210	\$ 57,718	\$ 7,492		
28	Feb-16	\$ 37,458	\$ 37,797	\$ (339)	\$ 31,932	\$ 25,165	\$ 6,767	\$ 79,893	\$ 62,962	\$ 16,931		
29	Mar-16	\$ 45,450	\$ 30,604	\$ 14,846	\$ 32,344	\$ 26,023	\$ 6,321	\$ 70,382	\$ 56,627	\$ 13,755		
30	Apr-16	\$ 47,390	\$ 35,605	\$ 11,785	\$ 31,551	\$ 26,861	\$ 4,690	\$ 73,373	\$ 62,466	\$ 10,907		
31	May-16	\$ 30,946	\$ 32,511	\$ (1,565)	\$ 32,095	\$ 27,619	\$ 4,476	\$ 69,875	\$ 60,130	\$ 9,745		
32	Jun-16	\$ 49,998	\$ 37,388	\$ 12,610	\$ 31,014	\$ 26,570	\$ 4,444	\$ 74,655	\$ 63,958	\$ 10,697		
33	Jul-16	\$ 47,794	\$ 39,361	\$ 8,433	\$ 31,783	\$ 25,004	\$ 6,779	\$ 81,815	\$ 64,365	\$ 17,450		
35	Aug-16 Forecast	\$ 42,516	\$ 38,386	\$ 4,129	\$ 32,741	\$ 25,540	\$ 7,201	\$ 77,931	\$ 63,927	\$ 14,004		

**FIGURE 2-6:**  
What happens when you abandon Widgets and put its costs into supporting Gidgets.

In Figure 2-6, columns A through H are identical to those in Figure 2-5. Columns J through K show the effect of taking the resources away from Widgets and using them to support Gidgets. The following steps show how to get those projections:

1. In cell K3, enter  $=C3+G3$  and copy and paste the formula into cells K4 through K33, and into cell K35.

Column K now has the sum of the actual costs for the two product lines from January 2014 through July 2016, plus the sum of the forecast costs in K35.

2. In cell J3, enter  $=(F3/G3)*K3$  and copy and paste the formula into cells J4 through J33.

This formula gets the ratio of revenue to cost for Gidgets in January 2014, and multiplies it by the total costs shown in cell K3. The effect is to apply one measure of gross margin to a higher measure of costs, and estimate what the revenue for Gidgets would be in that case.

**3. In cell J35, enter this formula** =TREND(J3:J33,K3:K33,K35).

This forecasts the revenues for Gidgets in August 2016, given the relationship between the projections of revenues and costs in J3 through K33, if you decided to support Gidgets only.

**4. To get a forecast of profit for Gidgets only in August 2016, enter** =J35-K35 **in cell L35.**

Notice that the sum of the profit in August 2016 for Widgets and Gidgets is \$4,128 + \$7,201 = \$11,330. But if you committed your Widget resources to Gidgets, your profit for August 2016 would be \$14,004 — \$2,674 more. In raw dollars, that doesn't seem like much, but it's a 24 percent increase. Generations of European casino owners have grown wealthy on much smaller advantages.

The reason, of course, is that the gross margin on Gidgets is larger than that on Widgets, even though your revenue on Widgets is almost 30 percent greater than on Gidgets. To summarize:

- » In column K, you act as though you had committed all your resources to Gidgets only, from January 2014 through July 2016. The effect is to remove all support from Widgets and add it to the support given to Gidgets.
- » In column J, you estimate the revenues you'd earn if you supported Gidgets only, using the historical margin for Gidgets.
- » Using the TREND function, you regress the revenue estimates in J3 through J33 onto the costs in K3 through K33, and apply the result to the estimated cost in K35. Subtract K35 from J35 to get a forecast of profit if you recognized your opportunity costs and supported Gidgets only.

Using two different scenarios — Widgets with Gidgets, and Gidgets alone — makes this example a little more difficult to follow. But it's a realistic illustration of how you can use the basic forecasting function TREND to help make an informed decision about resource allocation.

Of course, other considerations would factor into a decision to shift resources from one product line to another — an analysis of the nature of the errors in the forecasts (often termed the *residuals*), sunk costs, possible retooling to support added manufacturing capacity in a product line, the necessity of ongoing support for customers who have invested in Widgets, and so on. But one of the criteria is almost always financial estimates, and if you can forecast the financials with confidence, you're ahead of the game.



## IN THIS CHAPTER

Putting your forecast in context with qualitative data

Avoiding common errors in sales forecasting

Understanding the effects of seasons and trends

## Chapter 3

# Understanding Baselines

You build your sales forecast on something called a *baseline* — that is, data that describes your level of sales, usually in prior months, quarters, or years. But creating a numeric forecast without looking at the context isn't a good idea. You need to make sure you have a handle on product management's plans, marketing's promotional budget, sales management's intentions for hiring (or firing), and so on.

Even with a good context and a good baseline, several common errors can send your forecast reeling off course. Recognizing and avoiding these errors is easy if you know what they are, and in this chapter I point them out for you.

Your baseline will often reflect both an ongoing trend (sales have been heading generally up or down) and seasons (sales reliably spike or drop at certain times of the year). In this chapter, I call out some of the reasons that context, common errors, trends, and seasonality contribute to good forecasts — and bad ones.

# Using Qualitative Data

*Qualitative data* is information that helps you understand the background for quantitative data. Of course, that begs the question: What's quantitative data? I want to focus on this issue early, because it's an important one, and one that makes a real difference to the value of your sales forecasts.

*Quantitative data* is numeric data — the number of units your team sold during the prior quarter, or the revenue that your team brought in during March. With quantitative data, you can use Excel to calculate the number of units sold per month, or the fewest, or the most. You can use Excel to figure a moving average of the revenue your sales team has earned, or its minimum revenue, or the percentage of annual revenue earned during October.

In contrast, *qualitative data* doesn't have an average, a minimum, or a maximum. It's information that helps you *understand* quantitative data. It puts the numbers into a context. It helps to protect you against making really dumb mistakes. I've made my share of dumb mistakes in forecasts, and they've often happened because I haven't paid enough attention to the qualitative data — to my regret.

The right mindset can help you keep all the numbers in perspective. Knowing what questions to ask about your company's direction is key, of course. And you can better decide how to structure your numbers if you understand how your company wants to use your forecast. Here's a closer look at those issues.

## Asking the right questions

Suppose that your VP of Sales asks you to forecast how many cars your agency will sell during the next year. If your agency sells mostly Fords, it's reasonable to take a whack at a forecast. If, up until last year, your agency sold mostly Duesenbergs, making a forecast is unreasonable. You can't sell any Duesenbergs because nobody's making them anymore.

That example is admittedly extreme, but it's not entirely stupid. You need to know what your company is going to bring to market during the time period that you want to forecast into. Otherwise, your sales history — your baseline — just isn't relevant. And you can't make an accurate forecast that's based on an irrelevant baseline.



TIP

Here are some questions you should ask before you even start thinking about putting a baseline together:

» **How many salespeople will your company make available to you?** Will you have more feet on the street than you did last year? Fewer? About the same?

The size of the sales force makes a difference. To make a decent forecast, you need to know what sales resources you're going to have available.

- » **Will the commission levels change during the forecast period?** Is your company incentivizing its sales force as it has during, say, the last 12 months? If so, you don't need to worry about this in making forecasts. But if the business model has changed and commission rates are going to drop because the competition has dropped — or rates are going up because the competition has stiffened — your forecast needs to take that into account.
- » **Will the product pricing change during this forecast period?** Will your product line's prices jump? If so, you probably need to build some pessimism into your forecast of units sold. Will they drop? Then you can be optimistic. (Keep in mind that pricing usually affects units sold more than it does revenue.)

You can't use forecasting to answer questions like these. And yet their answers — which qualify as qualitative data — are critical to making good forecasts. You can have a lengthy, well-behaved baseline, which is really key to a good forecast. And then you can get completely fooled if your company changes its product line, or reduces its sales force, or changes its commission structure so much that the sales force walks, or lowers its prices so far that the market can't keep its collective hands off the product line. Any of these is going to make your forecast look like you shrugged and rolled a couple of dice. Albert Einstein said that God doesn't play dice; Stephen Hawking says that God does. In either case, you don't want to be thought of as a high roller.



REMEMBER

You can't depend entirely on a baseline to make a sales forecast. You need to pay attention to what your company is doing in its marketing, its pricing, its management of people, its response to the competition, in order to make a good sales forecast.

This book shows you how to make good forecasts *under the conditions in place when you got your baseline*. If those conditions are still in place, your forecast can be an accurate one. If not, it can't. So understanding as much as possible about the conditions that will be in place during your forecast period is important.

## Keeping your eye on the ball: The purpose of your forecast



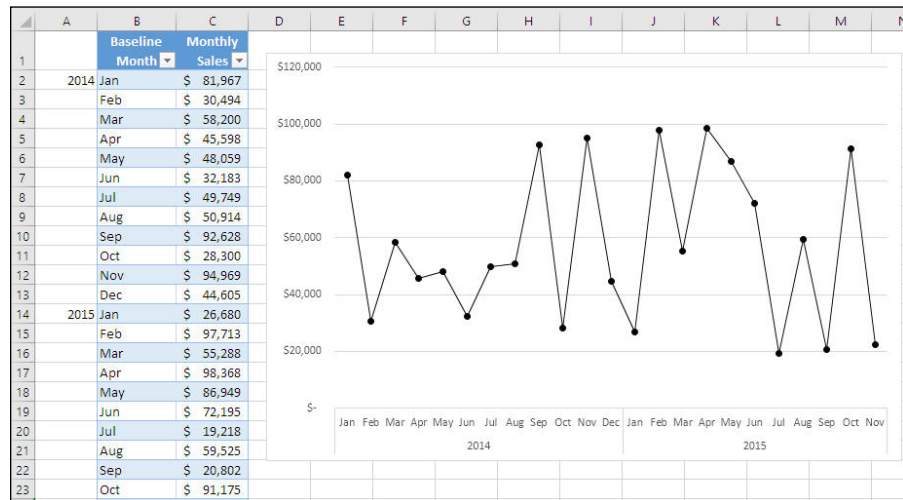
TIP

Set up your baseline to reflect the period you want to forecast into. That is, if you want to forecast one month's sales, your baseline should show your sales history in months. If the purpose of the forecast is to help guide financial projections such as earnings estimates, you probably want to forecast a quarter's results, and your



baseline should be organized into quarters. Chapter 8 shows you an easy way to turn data on individual sales into summaries for the time period you're after.

Figure 3-1 shows an example of a useful baseline.



**FIGURE 3-1:** The forecast is for the next month, so the baseline provides monthly sales history.



TIP

You could easily create the list of month names in column B of Figure 3-1. You would select cell B2 and type **January**, or **Jan**, as shown in the figure. Press Enter, and if necessary reselect cell B2 (or press Ctrl+Enter to leave cell B2 selected when you enter its value). Notice the small black square in the lower-right corner of the cell — it's called the *fill handle*. Move your mouse pointer over the fill handle. You'll see the pointer change to crosshairs. Now, although you can still see the crosshairs, press the mouse button, continue holding it down, and drag down as far as you want. Excel fills in the names of the months for you. This also works for days of the week.



TIP

There are some rules of thumb about building a baseline that you'll find it useful to keep in mind.

» **Use time periods of the same length in your baseline.** Using one period covering February 1 through February 14, and the next period covering February 15 through March 31 is peculiar. I've seen it done, though, just because it turned out to be convenient to put the data together that way. But that throws things off, because the apparent February revenues are an underestimate and the apparent March revenues an overestimate. Regardless of the forecast approach you use, that's going to be a problem. (You can safely ignore small differences, such as 28 days in February and 31 days in March.)



- » **Make sure the time periods in your baseline are in order, earliest to latest.** Several popular forecasting techniques, including two described in this book, rely on the relationship between one period's measurement and the next period's measurement. If your time periods are out of time order, your forecast will be out of whack. Often, your raw measures won't be in chronological order, and for various reasons you'll want to summarize them with a pivot table — which you can easily put into date order. In fact, pivot table's put summarized data into chronological order by default. See the figures in Chapter 8 for several examples.
- » **Account for all time periods in the baseline.** If your baseline starts in January 2015, you can't leave out February 2015, even if the data is missing. If the remaining months are in place, skip January 2015 and start with March 2015. Why? Because you want to make sure you're getting the relationship right between one period and the next.

## Recovering from Mistakes in Sales Forecasting

Forecasting can be tricky business. There's an old line, "It's hard to make predictions, especially about the future," attributed to people from Yogi Berra to Niels Bohr to Mark Twain.

And forecasting *is* tricky. You can do everything right and wind up with a forecast that completely misses the mark. It's not pure math. Human factors, the economy, the weather, technology — they all conspire to make your forecast look bad. This section discusses some reasons for mistakes that are beyond your control, and some that you should be able to get your arms around.

### Getting over it

In the forecasting dodge, you have to get used to being wrong. The best you can do is get close. More often than not you're going to miss the target. Fortunately, in sales forecasting, close is usually all that's needed. You just can't tell what's going to happen in the marketplace tomorrow, next month, next year. The best you can do is to act on these recommendations. It will help to get your management used to that idea. Then, they won't be too surprised when the forecast is off base.

And it will be. You can use the past as a guide to the future, but it won't always be a *reliable* guide. Because the future doesn't always respond to the past, your forecasts will sometimes be, well, wrong.

The problem is that the market doesn't stand still, for reasons like these:

- » Customers make new choices.
- » Product lines change.
- » Marketing strategies change.
- » Pricing strategies change.

Given all that, you just can't expect to nail your forecasts time and again.

But — and this is a big but — you usually have some lead time. Market conditions tend not to change suddenly. Customers don't all shift to ordering exclusively Hewlett-Packard computers on Tuesday, when they've been ordering Hewlett-Packards *and* Dells through Monday.

These things happen more gradually, and that's one reason that your baseline is so important. The forecasting tools that this book describes take that into account. They take note of the fact that one product's market share is gently declining, while another's is gently rising.

## Using revenue targets as forecasts

Here's how sales forecasts frequently come about: The Sales VP at corporate needs to tell the CFO what the revenues are going to be for Q2 2017. As those of us who have been in sales all know:

Little bugs have littler bugs  
Upon their backs to bite 'em,  
And littler bugs have littler still,  
And so on, ad infinitum.

So the Sales VP gets after the regional sales directors, who get after the district sales managers, and so on, ad infinitum, for Q2 2017 sales forecasts.

Now, suppose I'm a district sales director or a branch sales manager, and I'm supposed to come up with a sales forecast for Q2 2017. If I'm like some of the sales managers I've worked with in the past, here's how I do it:

**1. I check my quota for the second quarter.**

Turns out that's \$1,500,000.

**2. I phone in my forecast, which coincidentally is also \$1,500,000.**

An experienced sales manager would also build in a fudge factor.

Now that's a sorry way of forecasting. It's bad business and it chases its own tail. One major purpose of forecasting is to set sales quotas on a regional, branch, and personal basis. And here companies are rolling up quotas to set forecasts.

So you shouldn't take a quota and pretend it's a forecast. Of course, people do it all the time, but that doesn't make it a good idea.



TIP

A good idea is to look at the qualitative aspects of your product line, your sales force, your market, and your competition, to make sure you're on point for your forecast period. Then, if you still feel comfortable with those, take your baseline and extend it using one of the methods you'll read about starting with Chapter 10.

## Recognizing Trends and Seasons

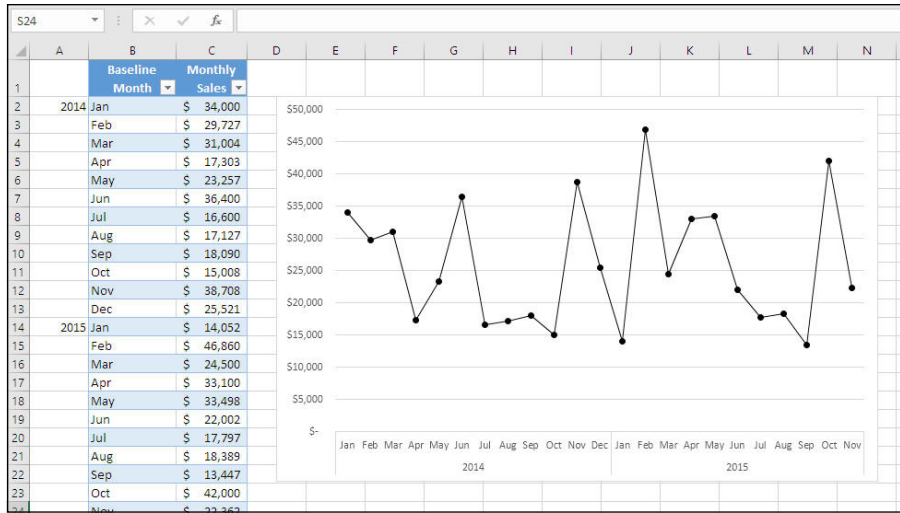
Baselines are often *stationary*: They tend to stay centered around an average value, even though they don't stick right to that average. Figure 3-2 has an example of a stationary baseline.

But baselines also often move up or down — the sales figures for a product generally rise or drop with some fluctuation in the direction, but overall you can see what's happening. These are baselines that have *trend*. You can see a baseline with an upward trend in Figure 3-3.

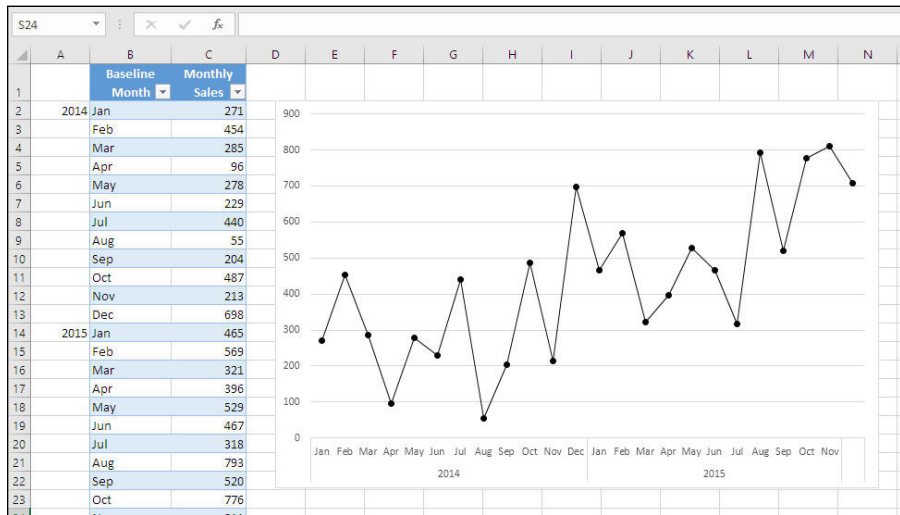
There are also baselines with *seasonality*. These baselines tend to move up and down regularly over time, usually in ways that correspond to seasons. Sales of fruit rise in the spring and summer, and drop back down in the fall and winter (unless you live in the southern hemisphere). Figure 3-4 shows a baseline of seasonal sales.

The next two sections help you understand why recognizing trends and seasons is important.

**FIGURE 3-2:**  
There's no clear trend in this baseline. It sticks close to home.



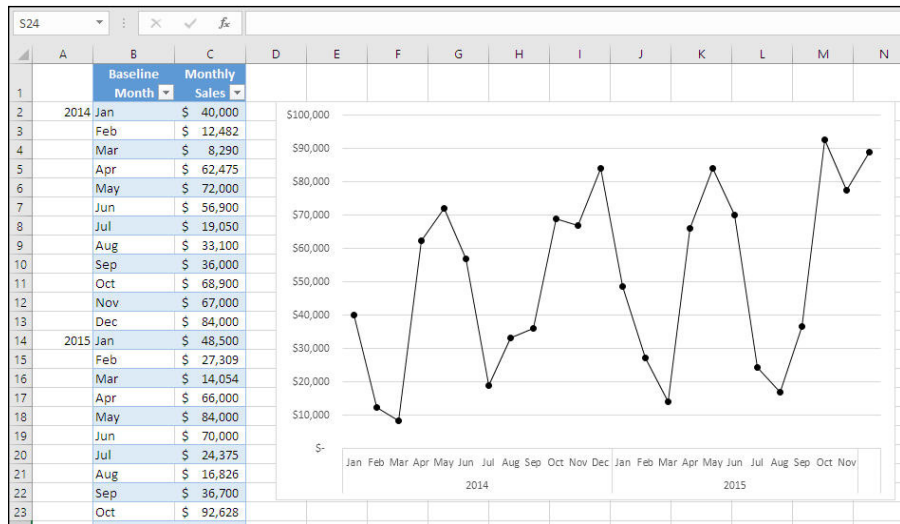
**FIGURE 3-3:**  
This baseline has trend. It wanders some, but you can see that the direction is generally up.



## Identifying trends

A baseline with a trend generally heads up (refer to Figure 3–3) or down. Chapter 4 discusses how trends can make forecasting trickier, as well as the steps you can take to make the forecast more accurate. For the time being, though, understanding what trends are about is a good idea.

In sales, trends tend to follow changes in customers' behavior. For good or ill, trends are an economic fact. In looking for trends, bear in mind the following:



**FIGURE 3-4:** The size of the season's effect on sales varies, but it's there quarter after quarter.

- » **People stop using certain products or services.** There are lots of ways that society encourages people to purchase some products — an upward trend — and discourages them from purchasing other products — a downward trend. Over time, upward trends often turn into downward trends due to changing market conditions. For example, someone who sells tobacco, whether retail or wholesale, probably wants to see an upward trend — but as consumers have more information about the dangers of smoking, fewer of them will buy the product. (Some will always buy, but not in the numbers seen decades ago.)
- » **People want the newer, faster versions.** Wi-Fi? Cable? Your very own fiber loop? Doesn't matter. People are impatient and they want to get stuff to and from the Internet faster than they used to. They go from 1,200 bps (yes, people used to send and receive at 1,200 bps, and slower yet) to 56,000 to whatever multimegabit rate your phone or cable company offers. The number of people subscribing to higher-speed communications increases, as a trend, over time. The same is true of many other technological improvements.
- » **People spend more dollars, but they may not spend more constant dollars.** Here's the deal: It costs more to buy a car in 2016 than it did in 1996. Blame it on inflation. Or blame it on the bossa nova — the fact is that things cost more than they used to. There are ways to deal with this, such as converting prices to constant dollars, but unless you do so you're going to be looking at a trend, and a meaningless one at that. Chapters 16 and (especially) 17 show you some good ways to deal with trend in a forecasting context.
- » **People spend more for the things they want.** For example, people generally pay whatever gasoline costs, even if the cost rises at a high rate. There are

lots of reasons for the increasing cost of gasoline, ranging from South American politics to thirsty SUVs to exploding economies in the Far East. Shrinking or static supply, blended with increased demand, creates upward revenue trends; as we've seen in 2015 and 2016, expanding supply, and static or decreased demand, creates downward trends.



TECHNICAL  
STUFF

One of the problems with a trend is that there's a mathematical relationship between one figure and the next in the baseline. The two main approaches to forecasting that this book covers — smoothing and regression — deal with those relationships differently. You tend not to worry about the relationship between one figure and the next if you're using the regression method of sales forecasting. If you're using smoothing, you sometimes want to start by removing the trend from the baseline, and reintroducing it later. You can find more about removing trends in Chapter 17.

## Understanding seasonality

A *seasonal* baseline is one that rises and falls regularly. For example, one that has higher sales revenue during the summer and lower sales revenue during the winter (such as Speedo swimsuits), or higher during the first and third quarters, and lower during the second and fourth quarters (such as a line of textbooks for a course that is offered every other quarter).

A seasonal baseline can be a special case of a *cyclical* baseline. Cyclical baselines rise and fall but not necessarily on a regular basis. A good example is the business cycle as it's related to recessions. Recessions come and go, but nothing requires them to follow the calendar. The U.S. economy contracted, big time, in the late 1860s, the early and mid-1880s, the 1910s, during the Great Depression of the late 1920s and early 1930s, and during the Great Recession of 2007 through 2009. But there is nothing regular about when these contractions occurred. They're *cyclical*, not *seasonal*.

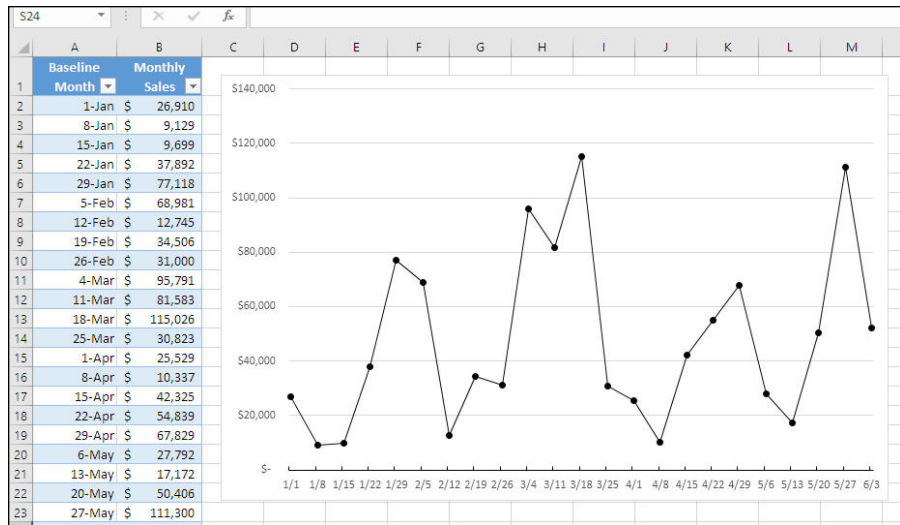
Contrast that with a baseline that rises and falls along with a calendar grouping. Sales that depend on the season of the year are both cyclical *and* seasonal. They follow a cycle, and it's a regular, seasonal cycle. Depending on the product and the time of year, the seasonal cycle might rise and fall every 3 months, or every 6 months, or even every 12 months.



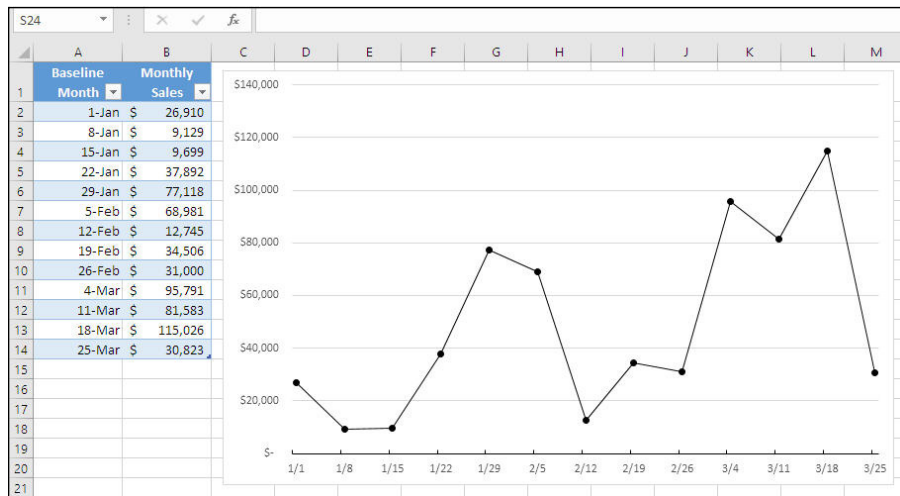
WARNING

A trap may be lurking. Suppose that you're dealing with a cyclical or seasonal series that looks like the one in Figure 3-5. How serious the trap is depends on how long your baseline is, and on how far out you want to forecast. Suppose you build your baseline on a weekly basis, from January 1 through March 12. It could look like the one in Figure 3-6, which shows a subset of the baseline in Figure 3-5.

**FIGURE 3-5:** You can get fooled into thinking you're dealing with a trend in your baseline, when a longer look would show you that you have a seasonal baseline.



**FIGURE 3-6:** Here, your baseline is limited to the upward trend of the seasonal series shown in Figure 3-5.



If you want to forecast into March 19, you're probably okay — although you don't yet know it on March 12, the series is still on its way up. But if on March 19 you want to forecast beyond into March 26, you're going to have a problem — although you don't know it yet, the series is starting down because of its seasonality.



**TIP**

Generally, the longer your baseline, the better your forecast. If your baseline is as shown in Figure 3-5, where you have six months' worth of data to forecast from, then you can tell the trend is seasonal and allow for that in your forecast. If the baseline extends from January 1 through March 19, though, you're going to get fooled.





## Chapter 4

# Predicting the Future: Why Forecasting Works

**M**any baselines of sales history rise during the early stages of a product's release, and fall as technology and fashions move on. During a product's middle age, the sales flatten out. Your forecasts' accuracy improves if you understand the nature of those trends and whether you should detrend the baseline. Excel offers methods of testing for trends, to help you decide whether you're looking at something real or just random variation.

Relationships in your sales baselines are the key to any forecast, whether the relationship is between one month's results and the next month's, or between historical sales results and some other variable such as advertising costs. In this chapter, I show you how to use Excel to quantify those relationships using Excel's worksheet functions and the Data Analysis add-in.

# Understanding Trends

Unfortunately, many decision-makers have no faith in sales forecasting. Their image of a forecaster is a combination of the meteorologist on Channel 7 and someone gazing into a crystal ball, probably wearing a pointy hat decorated with moons and stars.

But quantitative forecasting works for reasons that are sound, mathematical, and logical, and you can find plenty of examples of forecasts working in practice. If someone looks at you suspiciously when you trot out your sales forecast, you'll want to understand the reasons that forecasts aren't some kind of magic. One of those reasons is that many baselines — particularly baselines of sales revenues — involve *trends*.

A trend is easy to define, if not always easy to manage. A trend has two main characteristics:

- » **It goes up (good, if you're measuring revenues) or it goes down (not good for revenues).** It may fluctuate — for example, you'll see some temporary downturns in a baseline that's trending up — but in the main it's going in one direction. If you see many consecutive increases followed by many consecutive decreases, you're probably dealing with a seasonal or cyclic baseline, and certainly not a trend.
- » **The trend lasts much longer than the forecast period.** Suppose you use the results of four calendar quarters to forecast a fifth quarter. You may find that the trend established for the first four is substantially different from the one you'd get when the fifth quarter's actuals are in. But a trend figured on, say, 20 calendar quarters is unlikely to change much in the 21st quarter, unless a sudden and major sea change occurs in the market.

What causes trends? There are as many reasons for trends as you care to think up. Just a few examples:

- » **Products go out of style.** Smoked a cigarette, pipe, or cigar lately? If you have, you're out of step. The society you belong to is frowning and wagging its finger at you. You're having difficulty lighting up after a meal in a restaurant. The trend, my friend, is down: People don't want to smell your smoke anymore (mine either, and thank heaven for the nicotine patch).
- » **Inflation sets in.** As this book is written, the United States has had very low inflation for several years. During the 1970s and into the 1980s (when high interest rates finally slowed it down) there was serious inflation in the U.S. economy. The inflation caused prices — and therefore revenues — to trend upward.

- » **Technology improves productivity.** When your parents or grandparents or even great-grandparents were small children, they might have set aside one day a week to do the wash, outdoors in a big metal drum filled with soapy water that they'd churn with a big stick. Then along came washing machines. Washing machines were initially very expensive compared to a metal drum and a stick, and the demand for washing machines was, therefore, fairly low. But as economies of scale kicked in, and unit prices came down, what was once only for the wealthy became the norm, and the increase in revenues far outweighed the decrease in unit cost, until washing machines became commodities — then revenues flattened out. But there was a significant trend for decades because of increased productivity and demand.
- » **Products become more popular.** There are more cars, trucks, and SUVs on the road than there were last year, and last year there were more than the year before, and so on, all the way back to the Model T and even earlier. And each year that a census has been taken, the population of the United States has increased. You get more people, you get more people wanting to drive, you get more cars and trucks and SUVs.

The rest of this chapter gets into some of the effects of trends (and there are in reality many, many more baselines that have trends than baselines that are stationary). Trends are a principal reason that forecasting works. But if you can tell that a baseline is stationary — trending neither up nor down — you can do every bit as well as you can with a trend, if you handle things correctly.

## Watching revenues go up — and down

One reason that forecasting skeptics are, well, skeptical is that they tend not to understand how what happens in baselines tells us about the future.

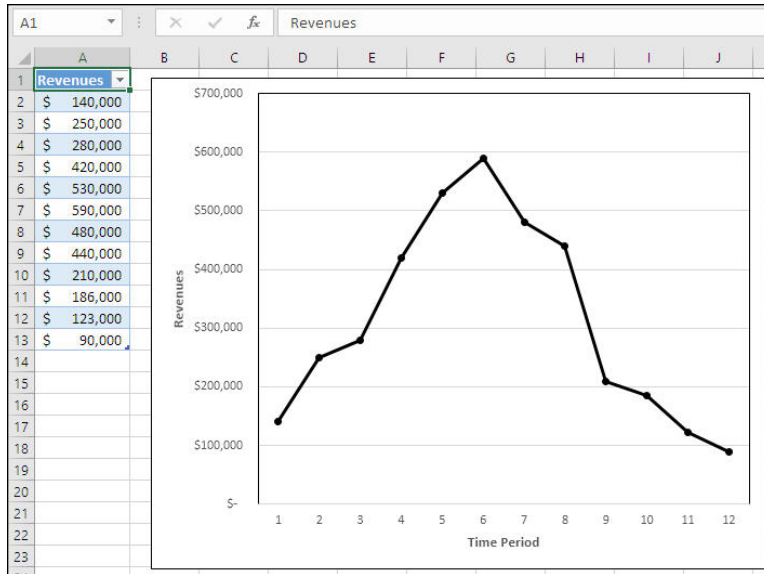
Consider rain clouds. They don't just appear overhead all of a sudden and start pouring water down on you. They form gradually, or the prevailing winds bring them over your area. In either case, if you're watching the sky, you have advance warning that you're going to get wet.

It's the same with baselines. They don't go screaming up and then suddenly, with no warning, go screaming down. Unless the forecast period you're using is shorter than a year (not recommended for most sales forecasting), you'll seldom see a baseline that looks like the one shown in Figure 4-1.

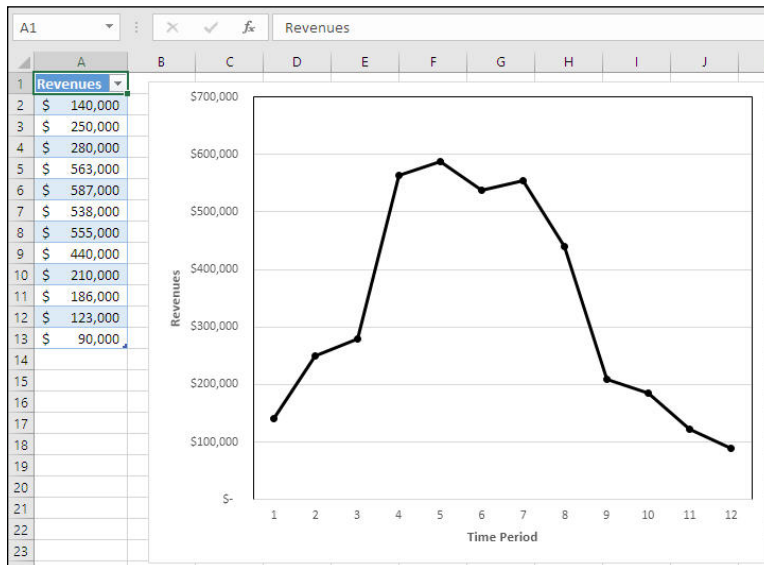
Of course, the situation shown in Figure 4-1 *could* occur, especially if some convulsion occurred in, say, the company's production facilities. But you'd know about that before it impacts sales.

Much more likely is the situation shown in Figure 4-2. Compare the changes in the trend's direction in Figures 4-1 and 4-2. The change in Figure 4-1 is abrupt and dramatic. Unless repeated due to seasonality, sales revenues just don't behave that way in the normal course of events. As in Figure 4-2, revenues slow, flatten, and finally decline. The point is that you have time to notice what's going on, and so does your forecasting technique.

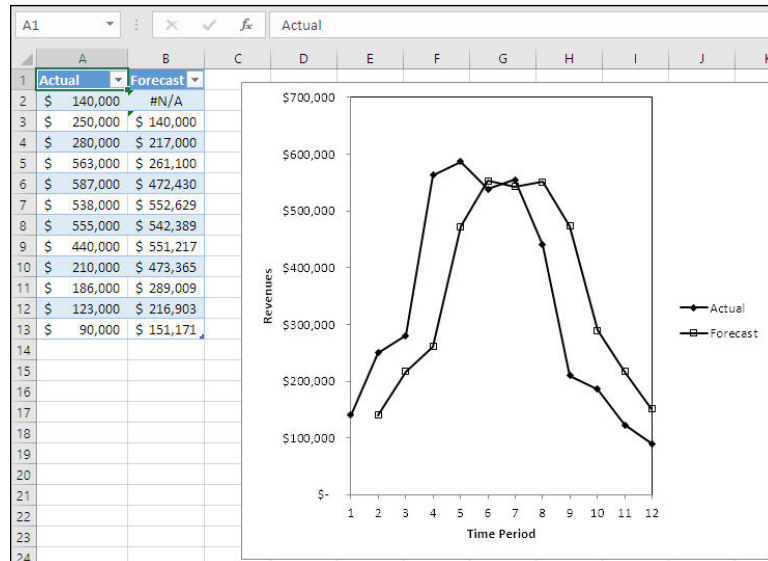
**FIGURE 4-1:**  
A realistic forecaster or user of forecasts doesn't worry about this sort of nonsense.



**FIGURE 4-2:**  
When a trend turns, it's usually due to seasonality or to movement through the product's life cycle.



If you're using a single-variable method, such as moving averages or exponential smoothing, your forecast builds recent sales data into your forecast. By the time the sixth data point has entered the baseline, the forecast has built in the fourth and fifth data points, and if the actuals had been shooting up, the forecast would push up in response. If the more recent actuals are beginning to fall off, it's not long before the forecast notices and starts to drop. Figure 4-3 shows the baseline in Figure 4-2 with a forecast line overlaid.

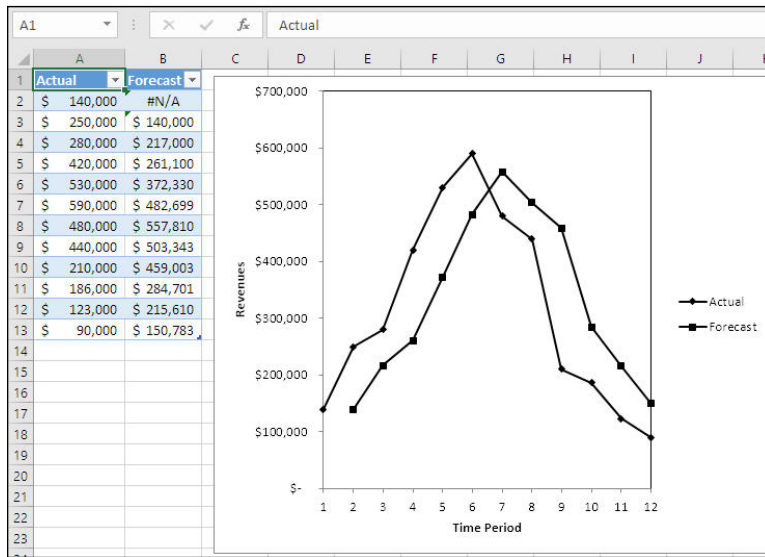


**FIGURE 4-3:** Smoothing and moving average forecasts typically lag a bit behind the actuals.

As I've said, it's unusual for baselines to resemble the one shown in Figure 4-1, but what if you do run into one and you use smoothing on it? Figure 4-4 has the Figure 4-1 data, including a forecast line from exponential smoothing.

So, when the worm turns at period 6 and drops \$110,000 between periods 6 and 7, the forecast for period 7 is off by \$77,000. And because this sort of forecast always lags behind a bit, it continues to be an overestimate through the end of the baseline.

There's nothing magical about quantitative forecasting. The process doesn't just pull numbers out of the worksheet like rabbits from a hat. It looks at what has gone before and makes its best estimate as to what the history suggests will happen next. Even if an abrupt change occurs in the series, the forecast often catches up before long.

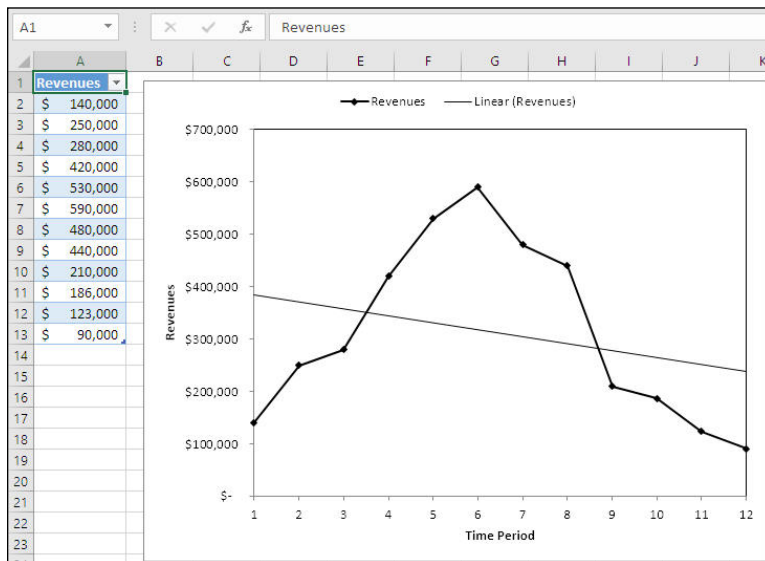


**FIGURE 4-4:**  
The forecast catches on to the change in direction pretty quickly.



**WARNING**

Standard regression forecasts don't work so well — in fact, they don't work at all — in catching up with changes in the baseline's direction: They provide straight-line (linear) projections, as shown in Figure 4-5. Autoregression forecasts, in which you regress a baseline on itself, are an exception. They catch up just like moving averages and smoothing.



**FIGURE 4-5:**  
A straight-line, regression forecast doesn't help you account for seasons or cycles.

# Testing for trends

How do you know whether a trend is real? If you see a baseline that looks like it's drifting up or down, does that represent a real trend or is it just random variation? To answer those questions, we have to get into probability and statistics. Fortunately, we don't have to get into them too far — wrist-deep, maybe.

The basic train of thought goes like this:

- 1. I use Excel to tell me what the correlation is between sales revenues and their associated time periods.**

It doesn't matter if I represent that time period as January 2011, February 2011, March 2011 . . . December 2016, or as 1, 2, 3 . . . 72.

- 2. If there's no relationship, as measured by the correlation, between revenues and time period, there's no trend and I don't need to worry about it.**
- 3. If there *is* a relationship between revenues and time periods, I have to choose the best way to handle the trend.**
- 4. After Excel calculates the correlation, I have to decide whether it represents a real relationship between time period and revenue amount, or whether it's just a lucky shot.**

If the probability that it's just luck is less than 5 percent, I'll decide it's a real trend. (Nothing magic about 5 percent, either — it's conventional. Some people prefer to use 1 percent as their criterion — it's more conservative than 5 percent, and they feel a little safer.) This raises the issue of statistical significance: What level of probability do you require before you decide that something (here, a correlation) is the real McCoy?



TECHNICAL  
STUFF

There are various methods for testing the statistical significance of a correlation coefficient. Here are three popular methods:

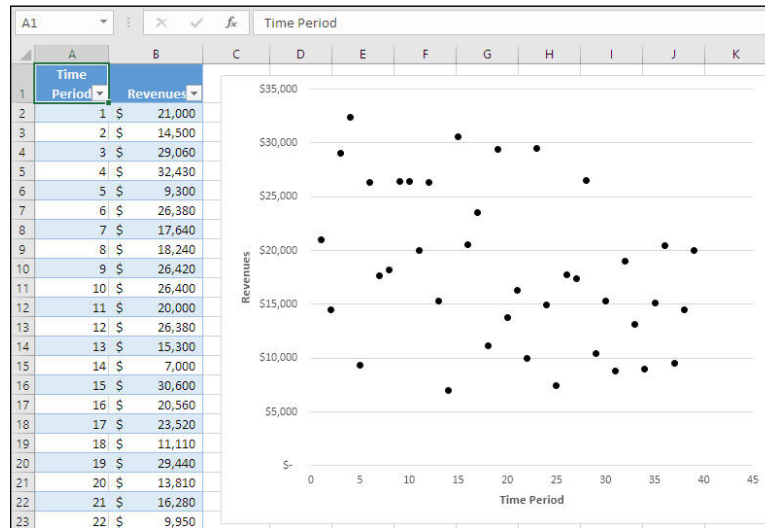
- » Test the correlation directly and compare the result to the normal distribution.
- » Test the correlation directly and compare the result to the *t-distribution* (the *t-distribution*, although similar to the normal curve, assumes that you're using a smallish sample rather than an infinitely large population).
- » Convert the correlation with the *Fisher transformation* (which converts a correlation coefficient to a value that fits in the normal curve) and compare the result to the normal distribution.

Other popular methods for testing the statistical significance of a correlation coefficient exist. Each returns a slightly different result. In practice, you'll almost

always make the same decision (the correlation is or is not significantly different from zero), regardless of the method you choose.

Let's run through an example of testing a correlation's statistical significance. Figure 4-6 shows the basic data: the time period in column A and the sales revenues in column B, along with the standard — and very useful — chart of revenues by time period to show you more of what's going on.

**FIGURE 4-6:**  
Notice that revenues gently decline over time in the chart, from upper left toward lower right, so the correlation is probably negative.



The numbers that identify the time periods are in cells A2 through A40. The sales revenue figures for each time period are in cells B2 through B40.

With the data on the worksheet, get the formulas in place:

1. **Select a blank cell — say, D2 — and enter** =CORREL(A2:A40,B2:B40).

In plain English, show the correlation between the period number in cells A2 through A40, and the sales revenues in cells B2 through B40. I don't have words to express how great this is. I don't want to make myself sound like a geezer, but when I took my first statistics course, calculating this correlation could take as long as 20 minutes, because we had to do it on a Burroughs adding machine that had a crank on its side. Now, with Excel, it takes all of 10 seconds.

The correlation is  $-0.38$ . Is that likely to be real or just random variation? In other words, if you extended the periods maybe another 40 or so back in time, would you get a correlation similar to the one you have here? A negative



correlation somewhere between, say,  $-0.25$  and  $-0.5$ ? Or would you get a correlation somewhere between  $-0.1$  and  $+0.1$  (that is, no real relationship)?

2. **To find out whether the correlation is real or random, enter  $=D2/(1/\text{SQRT}(\text{COUNT}(A2:A40)-1))$  in, say, cell D4.**

The formula assumes that the CORREL formula from Step 1 is in cell D2.

This returns something called a *z-value* or *z-statistic*. It tells you where the correlation in D2 lies on a normal curve. A negative *z-value* lies below the normal distribution's average value.

3. **Suppose you could repeat this test 100 times, each time with a different set of data. To find out how many of those data sets would have z-values as large as the one in cell D4, enter  $=\text{NORMSDIST}(D4)$  in cell D6.**

This formula returns the value 0.93 percent. That's less than 1 percent. In plain English, this means that if the true correlation were 0, you'd expect to get a calculated correlation as far from 0 as  $-0.38$  in only 1 of those hypothetical 100 data sets. It's more rational to assume that the correlation between revenue and time period is truly nonzero than it is to assume that it's the one out of 100 that's an aberration.

When you've taken the three steps just described, the worksheet looks like the one shown in Figure 4-7. It has three formulas not yet displayed in Figure 4-6, which taken together tell you whether the trend in the sales data is more likely to be a real one or is more likely to be a ghost. You're converting the correlation to a *z-value*, which is one that you can use in conjunction with the normal curve.

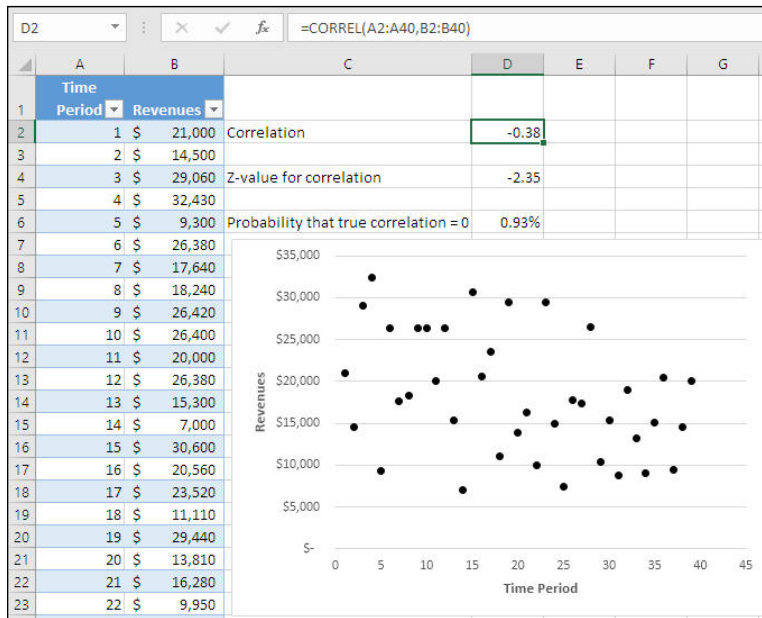


When you're testing correlations for statistical significance, as Steps 1 through 3 did, keep in mind that two issues determine the probability you calculate: the size of the correlation and the number of data points that go into it. Here you have a correlation of moderate size *and* a reasonably large number of data points, so getting a statistically significant result isn't surprising.

If you conclude that the trend the correlation measures is real (and when the probability is less than 1 percent that the correlation is a ghost, you probably should accept that conclusion), you have two more questions to ask yourself:

- » **Should I use a forecasting approach that handles trends well?** You'd think that if you detected a trend, you should use a forecasting approach that handles trends well. That's often true, but not necessarily. Suppose that instead of using time period as one of the variables in your correlation analysis, you used something such as sales revenues made by the *competition*.

**FIGURE 4-7:**  
Because you've calculated the correlation's z-value in cell D4, you can compare it to the normal standard distribution (NORM.S.DIST) in cell D6.



If the competition's revenues are slipping as yours are (or if both sets of revenues are growing), you'll find a likely significant correlation between your revenues and the competition's. But it's quite possible — even likely — that there's no real, causal relationship between their revenues and yours. It may be that both yours and theirs are correlated with the real causal factor: The size of the overall market is changing. In that case, you would probably be much better off using a measure of overall market size as your predictor variable. In this scenario, market size has a direct, causal relationship to your revenue, whereas your competition's revenue has only an indirect relationship to your revenue.

» **Should I detrend the data?** A hidden variable, such as a consistent change in the overall size of a market, can lead you to believe that a predictor variable and the variable you want to forecast are directly related, when in fact they're not. Or the predictor and the forecast may change in similar ways because they're both related to *time*.

The way to handle this sort of situation is to detrend both variables first by means of a transformation. (I show you how to do that in Chapter 17.)

Or you may prefer to make your forecast using an approach that doesn't necessarily handle trends well, such as moving averages or simple exponential smoothing. One reason for doing this is that you may find the regression approach with your data set isn't as accurate a forecaster as moving averages or smoothing. Again, see if you can transform the data to remove the trend.

# Matchmaker, Matchmaker: Finding Relationships in the Data

When you make a *quantitative forecast* (a forecast that uses a numeric baseline rather than something like expert opinions), you're always looking for relationships. Suppose you're considering using regression to forecast. You can get your hands on several possible predictor variables, any one (or any combination) of which might give you your best forecast.

In the sales arena, this means looking for relationships between sales and some other variables like size of sales force, time period, or unit price. (Expert opinions, as long as they come from a real expert, are valuable, too — even if you use them only to provide a context for your quantitative forecast.)

The relationship between sales revenue for one time period and a prior time period is also frequently of interest. This is called an *autocorrelation* and is close conceptually to *autoregression*. Calculating an autocorrelation can help you make many decisions, including the following:

- » Which forecasting method to use
- » Whether you'd be misled by a moving-average forecast
- » How to structure an exponential-smoothing forecast
- » Whether to detrend a baseline

Especially if you have a sizable number of possible predictor variables, calculating the relationships one by one can be a real pain. For that, you'll want to use the Data Analysis add-in.

One of the tools you'll find in the Data Analysis add-in is the Correlation tool. (For more information about installing the add-in into Excel, see Chapter 7.) If you set up your baseline as an Excel table, the Correlation tool takes most of the agony out of calculating several correlations.

Figure 4-8 begins an example that this section works through. It shows:

- » Sales revenues (the variable you want to forecast)
- » Time period
- » Unit price
- » Size of sales force

	Sales (\$000)	Time Period	Unit Price	Sales Force	Advertising (\$000)	Revenue Estimates (\$000)
10	\$ 251,897	1	240	52	\$ 3,149	\$ 121,806
11	\$ 154,426	2	265	56	\$ 19,303	\$ 160,487
12	\$ 237,450	3	255	59	\$ 23,745	\$ 215,323
13	\$ 201,980	4	281	58	\$ 50,495	\$ 251,289
14	\$ 271,298	5	257	54	\$ 13,565	\$ 132,589
15	\$ 260,171	6	247	54	\$ 65,043	\$ 158,200
16	\$ 243,231	7	275	51	\$ 24,323	\$ 104,498
17	\$ 281,218	8	265	53	\$ 31,637	\$ 330,361
18	\$ 330,765	9	260	59	\$ 66,153	\$ 238,259
19	\$ 289,533	10	273	50	\$ 21,715	\$ 361,302
20	\$ 332,443	11	253	53	\$ 12,467	\$ 350,477
21	\$ 241,126	12	262	55	\$ 12,056	\$ 118,257
22	\$ 153,515	13	278	41	\$ 23,027	\$ 184,314
23	\$ 119,380	14	276	22	\$ 4,477	\$ 131,311
24	\$ 200,280	15	267	33	\$ 40,056	\$ 204,771

**FIGURE 4-8:**  
This is too much data to calculate conveniently with worksheet functions.

- » Advertising dollars
- » Total of sales managers' revenue estimates

Your goal is to decide which (if any) of the last five variables to consider as predictor variables in a regression forecast of sales revenue. To begin that work, calculate each of the correlation coefficients. I show you how to do that in the next section.

## Choosing the predictors

One of the goals of the regression approach is *parsimony* — a highfalutin way of saying that you don't want to use more predictor variables than are needed, when you *are* using more than one. Suppose that one of your available predictor variables is unit price and another is units sold. These two variables tend to be strongly related: The lower the price, the more units you'll sell.

So, in a regression situation, you'd want to use one or the other, but probably not both, as a predictor. If two variables such as unit price and units sold are strongly correlated, they tend to be redundant. Suppose you use time period and unit price together to forecast revenue. In that case, you normally wouldn't want to add units sold to the forecast equation, because it would add little information to what's already provided by unit price.



TECHNICAL  
STUFF

And adding units sold to the equation would come at a cost: some technicalities about degrees of freedom, something awful called multicollinearity, and some more-practical stuff about a misleadingly high Multiple R (see Chapter 11 for more information on all of this).

Your initial correlation analysis should look at the strength of the relationship between each predictor variable and sales revenues, because if the relationship is weak, you don't want it in your forecast equation. But you should also look at the strength of the relationship between each pair of predictors, because if that's too strong, you'll wind up with more variables than you need or want in the equation.

Instead of starting with a calculation of all 15 correlations, you can use Excel's LINEST worksheet function. That function returns results that enable you to evaluate the relationships among the predictor variables and between each predictor variable and the predicted variable (here, sales revenue). We look more closely at that approach in Chapter 12. For the present, thinking in terms of individual correlations helps set the stage for using the more sophisticated analysis you can get from LINEST.

There are six variables in the baseline list. That means you need to calculate 15 correlation coefficients. If you use the worksheet function CORREL to calculate them, you waste your time and risk making errors. It's tedious, exacting work. And because I've done that myself — that is, I've calculated the correlations one by one using CORREL — I can tell you it leads to lower-back pain later on.



TIP

If you want to know how many different correlations you'll need for any given number of variables, use this formula, where NC is the number of correlations and NV is the number of variables:

$$NC = NV * (NV - 1) / 2$$

Here's how to get what's called a *correlation matrix* (which is a table of correlation coefficients) with the Data Analysis add-in's Correlation tool, using the data shown in Figure 4-8:

- 1. With the Data Analysis add-in installed in Excel, click the Data tab on the Ribbon and then choose Data Analysis in the Analyze group.**

The Data Analysis dialog box appears.

- 2. In the Analysis Tools list box, choose Correlation, and click OK.**

The Correlation dialog box appears.

- 3. Click in the Input Range box and drag through your entire input range, including the labels at the top of each column.**

4. **Make sure that the Columns radio button is selected.**
5. **Select the Labels in First Row check box.**
6. **Click the Output Range button.**

The Correlation tool might make the Input Range box active again when you click the Output Range button. If it does, click in the Output Range box and drag across any cell address that might already appear there before you enter the cell where you want the output to start (such as A1).

7. **Click OK.**

The Correlation tool takes over and produces the triangular matrix of correlations shown in Figure 4-9.

	Sales (\$000)	Time Period	Unit Price	Sales Force	Advertising (\$000)	Revenue Estimates (\$000)
1	1					
2	0.56	1				
3	-0.32	-0.26	1			
4	0.17	-0.40	0.21	1		
5	0.57	0.45	-0.04	0.16	1	
6	0.40	0.07	0.12	0.38	0.34	1
7						
8						
9	\$ 251,897	1	240	52	\$ 3,149	\$ 121,806
10	\$ 154,426	2	265	56	\$ 19,303	\$ 160,487
11	\$ 237,450	3	255	59	\$ 23,745	\$ 215,323
12	\$ 201,980	4	281	58	\$ 50,495	\$ 251,289
13	\$ 271,298	5	257	54	\$ 13,565	\$ 132,589
14	\$ 260,171	6	247	54	\$ 65,043	\$ 158,200
15	\$ 243,231	7	275	51	\$ 24,323	\$ 104,498
16	\$ 281,218	8	265	53	\$ 31,637	\$ 330,361
17	\$ 330,765	9	260	59	\$ 66,153	\$ 238,259
18	\$ 289,533	10	273	50	\$ 21,715	\$ 361,302
19	\$ 332,443	11	253	53	\$ 12,467	\$ 350,477
20	\$ 241,126	12	262	55	\$ 12,056	\$ 118,257
21	\$ 153,515	13	278	41	\$ 23,027	\$ 184,314
22						

**FIGURE 4-9:**  
The correlations of 1.00 appear because the correlation of a variable with itself is always 1.00.



TIP

For correlations, the Correlation tool gives you 15 decimal places. That's too much for present purposes and it makes the matrix harder to read. To change the number of decimal places:

1. **Select the full matrix.**
2. **Click the Ribbon's Home tab.**
3. **Click Format in the Cells group.**

4. **Select the Number tab.**
5. **In the Category drop-down list, select Number.**
6. **Accept the default of two decimal places (or change it if you want) and click OK.**

This changes the 1 integers in the main diagonal to show as 1.00. You can select them one by one and reduce the decimal places to zero, but that's a little pointless.



REMEMBER

The principal goal of analyzing the correlations is to start the process of eliminating unnecessary predictor variables and identifying potentially useful ones.

## Analyzing the correlations

What does all this preliminary correlation analysis tell you? The first thing I can tell by looking at the correlation matrix is that I want to start by using Time Period and Advertising as predictors, at least for a first run. They have respectable correlations with sales: 0.56 and 0.57 in cells B3 and B6 of Figure 4-9. I might also use Revenue Estimates, which has a correlation of 0.40 with sales (cell B7).

And I probably want to eliminate Unit Price and Sales Force as predictor variables — they both have relatively low correlations with sales. Further, Sales Force has a moderate correlation with Time Period, which I've already tentatively decided to keep. So, Sales Force might not add much information beyond that supplied by Time Period.

My next step would be to use the Data Analysis add-in's Regression tool to create a forecast. (See Chapter 11 and Chapter 16 for more information on using the Regression tool.) And I'd start by using Time Period, Advertising, and Revenue Estimates as my predictor variables. (Bear in mind that when you use more than one predictor variable, you still wind up with one forecasting equation. For each predictor variable that you add, Regression just adds another factor to the equation.)

I'm not utterly confident of my judgments about the correlations. So I'd continue by using the Regression tool two or three more times: once adding Unit Price to the predictors, once adding Sales Force. And depending on those results, I might add them both in to the predictors and use the Regression tool on all the possible predictors. If I had a much shorter baseline, I'd be worried about that.



TIP

I like to have at least eight times as many time periods in the baseline as predictor variables in the regression equation — but that's just a rule of thumb.

Figure 4-10 shows the results of using the Data Analysis add-in's Regression tool with Time Period, Advertising, and Revenue Estimates as the predictor variables.

	A	B	C	D	E	F	G	H	I	J
1	Sales (\$000)	Time Period	Advertising (\$000)	Revenue Estimates (\$000)		SUMMARY OUTPUT				
2	\$ 251,897	1	\$ 3,149	\$ 121,806						
3	\$ 154,426	2	\$ 19,303	\$ 160,487		Regression Statistics				
4	\$ 237,450	3	\$ 23,745	\$ 215,323		Multiple R	0.71			
5	\$ 201,980	4	\$ 50,495	\$ 251,289		R Square	0.51			
6	\$ 271,298	5	\$ 13,565	\$ 132,589		Adjusted R Square	0.46			
7	\$ 260,171	6	\$ 65,043	\$ 158,200		Standard Error	42863.57			
8	\$ 243,231	7	\$ 24,323	\$ 104,498		Observations	40			
9	\$ 281,218	8	\$ 31,637	\$ 330,361		ANOVA				
10	\$ 330,765	9	\$ 66,153	\$ 238,259						
11	\$ 289,533	10	\$ 21,715	\$ 361,302			df	SS	MS	F
12	\$ 332,443	11	\$ 12,467	\$ 350,477		Regression	3	67639514632	22546504877	12.27
13	\$ 241,126	12	\$ 12,056	\$ 118,257		Residual	36	66142269278	1837285258	
14	\$ 153,515	13	\$ 23,027	\$ 184,314		Total	39	133781783910		
15	\$ 119,380	14	\$ 4,477	\$ 131,311						
16	\$ 200,280	15	\$ 40,056	\$ 204,771						
17	\$ 141,599	16	\$ 5,310	\$ 136,308		Intercept	156152.4	22546.21	6.93	0.00
18	\$ 210,875	17	\$ 18,452	\$ 240,851		Time Period	2044.658	659.35	3.10	0.00
19	\$ 285,605	18	\$ 42,841	\$ 221,149		Advertising (\$000)	0.836446	0.39	2.12	0.04
20	\$ 298,577	19	\$ 11,197	\$ 106,862		Revenue Estimates (\$000)	0.217649	0.10	2.16	0.04
21	\$ 259,678	20	\$ 9,738	\$ 161,538						
22	\$ 255,948	21	\$ 22,395	\$ 278,345						
23	\$ 305,582	22	\$ 37,948	\$ 149,876						

**FIGURE 4-10:** Either the Multiple R or the R Square figure tells you that this regression forecast will be reasonably reliable.

A lot of information is wrapped up in this analysis. At the outset, the most important item to look at is the R-squared number in cell G5. It can range from 0 to 1.0, and the closer it is to 1.0, the more accurate you can expect your forecast to be. An R-squared of 0.5 isn't bad. That means that 50 percent of the variability in sales can be predicted using the regression forecast equation.



TECHNICAL STUFF

Depending on where you look, you'll find that Excel and Excel-related documents use the terms *R Square* and *R-squared*, and even  $R^2$ . They're the same thing.

In Figure 4-10, you find the coefficients for the forecast equation in cells G17 through G20. Rounding them a little, your equation would be:

$$\text{Sales} = 156152 + (2044 * \text{Time Period}) + (0.8 * \text{Advertising}) + (\text{Revenue Estimates} * 0.22)$$

Just plug the next period's values of Time Period, Advertising, and Revenue Estimates into the equation to get your Sales forecast for the next period.



REMEMBER

The correlation analysis just gives you a place to start with your choice of predictor variables. With 5 possible predictors, there are 31 different combinations including 1, 2, 3, 4, or 5 predictor variables, and you need a place to start. An analysis of the relationships between the variables is a good one.





# **Organizing the Data**

### **IN THIS PART . . .**

Part 2 gets into how to set up your data as the basis for a forecast. Understanding your product line, your company's sales strategy, and the marketplace that you and your people are selling into is really important. It's equally important to have a sales history that you can use to make your numeric forecast. Part 2 shows you how to set up that sales history, so Excel can take best advantage of those numbers.

Knowing when you need to pay attention to order — and when you can ignore it

Recognizing the importance of time periods

Making your time periods equal

## Chapter 5

# Choosing Your Data: How to Get a Good Baseline

In most cases, you'll get the most out of your historical baseline of sales data if you put it in chronological order. And because you're going to forecast into the future, the order should be ascending chronological order. This chapter shows you the easiest way to arrange that order for your baseline.

Your forecast will be for a particular period of time. To some extent, your baseline's time periods determine the length of time your forecast will cover. For example, if your baseline's time periods are years, forecasting revenues for the next month is tough. On the other hand, if you need to forecast a year, you may have to jiggle your baseline some. In either case, the length of each time period in your baseline is important.

When you forecast, you're trying to separate the *signal* (the regular, dependable component of your baseline) from the *noise* (the irregularities that come from unpredictable events, like sales reps being out sick, random changes in your customers' buying patterns, and so on). To help with this separation, you want to impose some order on the chaos. One of the ways you do this is to use equally spaced time periods of nearly equal length.

# Early to Bed: Getting Your Figures in Order

One of the characteristics of a useful baseline is that the data is in a rational *order*. For some purposes, you may have a table of monthly sales revenues that's sorted in order of magnitude. That is, you might show the largest monthly revenue first, and then the next largest, and so on. That would tend to highlight the most (and, at the bottom, the least) successful time periods. Or you might have the table sorted by sales rep, to gather each rep's results together.

It's even possible that you have in an Excel workbook a range of sales data that came back from a database query in some random sequence. The thing to keep in mind is that before you can create a sensible forecast, the data must be grouped and sorted in ascending order by date.

## Why order matters: Moving averages

When you forecast using moving averages, you're taking the average of several consecutive results — in this book, I look at sales results, but I could just as easily be tracking the number of traffic accidents over time. So, you may get the moving averages like this:

- » **First moving average:** The average of months January, February, and March
- » **Second moving average:** The average of months February, March, and April
- » **Third moving average:** The average of months March, April, and May

Notice that the moving averages each combine an equal number of months (three apiece) and that each consecutive moving average *begins* with the next consecutive month. Figure 5-1 has an example.

In Figure 5-1, columns C through F show the moving averages themselves, as well as where each moving average comes from. For example, the third moving average is 42,745 (in cell E6), and it's the average of the values in cells B4, B5, and B6.

Suppose you decide that each moving average will be based on three baseline values. The first moving average *must* be based on the first three, chronologically consecutive values. The baseline values *should* be in chronological order, as shown in Figure 5-1. It's possible to make the first moving average consist of January, February, and March even if the baseline is in some random order — but doing so is tedious and error prone. And it doesn't make a lot of sense. But if you sort the baseline in chronological order, you can use a simple copy and paste, or AutoFill, to create the moving averages.

**FIGURE 5-1:**  
The moving averages show how the baseline's level, or *trend*, is gradually increasing.

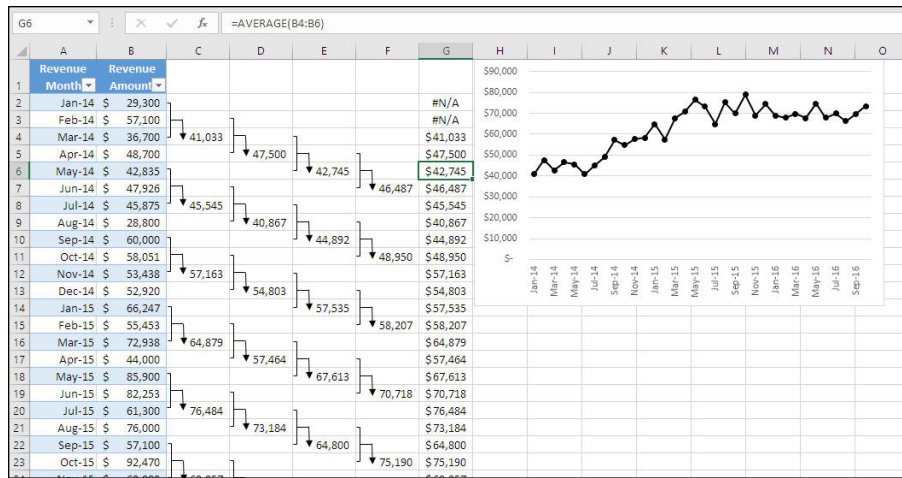


Figure 5-1 shows how the moving averages *should* look: The baseline in columns A and B is in order. There's one and only one record for each time period. The level of the baseline is gradually increasing over time, and the chart of moving averages reflects that increase.

Here's how easy it is to get the moving averages shown in Figure 5-1:

1. In cell G4, type `=AVERAGE(B2:B4)` and press Enter.
2. If necessary, reselect cell G4. Click the Ribbon's Home tab and choose Copy from the Edit group.
3. Select the range of cells G5:G37 and choose Paste from the Edit group.

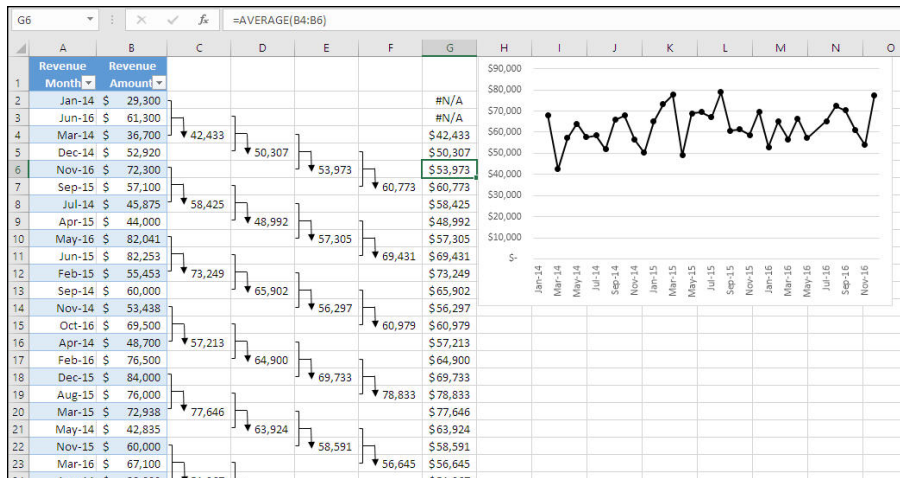
That's all there is to it. (AutoFill is even quicker on the worksheet, but to describe it takes more words on the printed page.)

On the other hand, Figure 5-2 shows what can happen when the data in your baseline is out of order.

In Figure 5-2, there's no rhyme or reason to the order in which the baseline data appear — and this is just the sort of thing that can happen if you've gotten the baseline data from monthly reports that have been stuffed into a file drawer, or even if you pulled them into your worksheet from a database that stores the monthly results in some other order.

The chart does show the moving averages in chronological order, but when the averages are based on a random sequence of months, that's not much help. Notice in Figure 5-2 that the moving averages in the chart form a line that shows no trend — but we know from Figure 5-1 that the trend is gently up.

**FIGURE 5-2:**  
Your moving averages can jump all over the place if you haven't tended to your baseline's order.

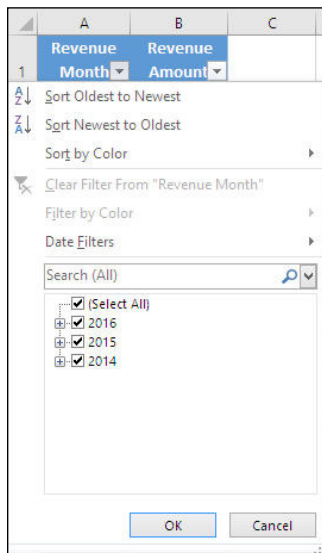


If you do get data in some sort of random order, as in columns A and B of Figure 5-2, the problem is easily fixed. Take these steps:

1. Click the drop-down arrow in cell A1.
2. Click on Sort Oldest to Newest as shown in Figure 5-3.

Your baseline will now be sorted into chronological order, and your moving averages will make sense. (I'm assuming that your table contains just one record per time period. If not, you'll probably want to start by moving your data into a pivot table and grouping the records. Chapter 8 shows you how that's done.)

**FIGURE 5-3:**  
Different data types have different sorting options.



## Why order matters: Exponential smoothing

The idea behind exponential smoothing is similar to the idea behind moving averages. In both cases, you forecast what's going to happen next on the basis of what happened before. For example, in a three-period moving average, you would get an April forecast by averaging January, February, and March actuals, or a fourth-quarter forecast by averaging the actuals from the first, second, and third quarters.

And with moving averages, you generally give each period that goes into an average the same weight:  $Q_4 \text{ forecast} = (Q_1 + Q_2 + Q_3) \div 3$ . In exponential smoothing, the farther back you go in the baseline, the less the impact of the actuals on the next forecast. For example, to get a forecast for July using exponential smoothing, these factors are plausible:

- » June exerts 100 percent influence on the July forecast.
- » May exerts 70 percent influence.
- » April exerts 50 percent influence.
- » March exerts 34 percent influence.
- » February exerts 24 percent influence.

So the farther back into the baseline you go, the less the influence an actual result exerts on the next forecast. This is an intuitively appealing approach: You've probably had experience with how, say, a fashion in clothing has less and less impact on your choices in apparel as more and more months go by. Leisure suits, for example. Jeans with holes in them. Spats.

The formula that you use to do exponential smoothing is deceptively simple. See Figure 5-4 for an example.

You can tell from looking at the formula in the Formula Bar that a forecast is a weighted average of the prior actual and the prior forecast. Again, after you've entered the formula once, near the top of the baseline, it's just a matter of copying and pasting it down to the end of the baseline.

But that's because the baseline is in chronological order, with the earliest actual results shown first. If the baseline were in some random order, as in Figure 5-2, you *could* do exponential smoothing, but it would take forever to get the formulas right.

C4			
Revenue		Revenue	
Month	Amount	Forecast	
Jan-14	\$ 29,300	#N/A	
Feb-14	\$ 57,100	\$ 29,300	
Mar-14	\$ 36,700	\$ 37,640	
Apr-14	\$ 48,700	\$ 37,358	
May-14	\$ 42,835	\$ 40,761	
Jun-14	\$ 47,926	\$ 41,383	
Jul-14	\$ 45,875	\$ 43,346	
Aug-14	\$ 28,800	\$ 44,105	
Sep-14	\$ 60,000	\$ 39,513	
Oct-14	\$ 58,051	\$ 45,659	
Nov-14	\$ 53,438	\$ 49,377	
Dec-14	\$ 52,920	\$ 50,595	
Jan-15	\$ 66,247	\$ 51,293	
Feb-15	\$ 55,453	\$ 55,779	
Mar-15	\$ 72,938	\$ 55,681	
Apr-15	\$ 44,000	\$ 60,858	
May-15	\$ 85,900	\$ 55,801	
Jun-15	\$ 82,253	\$ 64,831	
Jul-15	\$ 61,300	\$ 70,057	
Aug-15	\$ 76,000	\$ 67,430	
Sep-15	\$ 57,100	\$ 70,001	
Oct-15	\$ 92,470	\$ 66,131	

**FIGURE 5-4:** This becomes tedious in a hurry if your baseline isn't in chronological order.

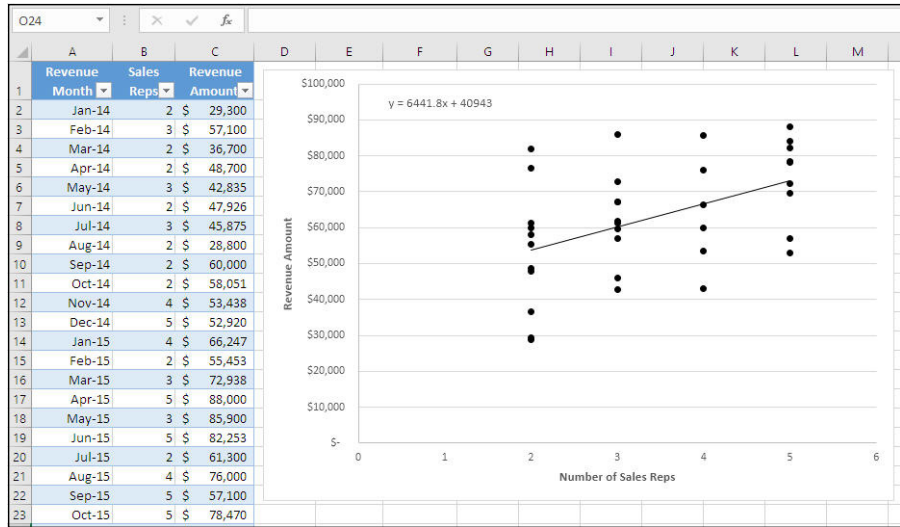
## Why order doesn't matter: Regression

With moving averages and exponential smoothing, the argument for putting the baseline in chronological order is based on the fact that you're using the forecast variable itself to make the next forecast. You're averaging, or otherwise combining, prior values of sales revenues so as to forecast the next revenue figure. Therefore, having the baseline in chronological order is helpful.

The same thing can happen with the regression approach, where you get an application such as Excel to look at your baseline and develop an equation that you use for your forecasts. Used this way, it's called *autoregression* because, once again, you're forecasting from past actuals. Autoregression needs the baseline to be in chronological order.

Another example of regression in forecasting is the use of a *different* variable to forecast sales — something such as the size of your sales force, your changing market share, unit price, even month and year during which your sales reps made those sales. Many variables can affect your sales revenues or number of units sold. And here, finally, it doesn't matter what order your baseline is in. Figure 5-5 shows the principal reason.





**FIGURE 5-5:** The chart, an XY (Scatter) chart, doesn't require that the baseline be in any particular order.

You're not making your forecast based on earlier values of your sales baseline. Your forecast is based on another variable entirely. If you're going to get more sales reps, your analysis might forecast higher revenues. This is due to the historic relationship between number of sales reps and amount of revenue. Regression analysis gives you an equation (shown in the chart in Figure 5-5). You pop the upcoming number of sales reps into the equation, and out pops your revenue forecast.

Because regression analysis doesn't require you to have your baseline in chronological order, you can get away with even a random order. But why would you? I suppose if that's how your data comes in and you only want to use regression, you might pass on sorting it first.

## Staying Inside the Lines: Why Time Periods Matter

Two terms in particular are important to this section:

- » **Baseline:** I use this term a little loosely. It's commonly used to mean historical data, going back some undefined distance from the present day, and consisting of at least one variable — the one that will be forecast. The baseline can have other measures, too, such as the dates that the forecast variable was measured, and other variables of interest like unit price or number of sales reps.

But *baseline* can also be used to mean just the forecast variable itself. You can almost always tell from the context which usage someone means.

» **Time period:** This is the period that you've split your baseline into. In sales forecasting, that's usually months and quarters, although some businesses recognize revenue weekly. (Doesn't matter, by the way: All sales figures are grist for this mill.)

In the following sections, I go into more detail on time periods.

## Deciding how far to forecast

When you're asked to make a forecast, one of the first things you need to consider is how far into the future you want to peer. Some forecasting techniques put you in a position to forecast farther out than do others. Figure 5-6 shows two techniques that let you forecast just one time period ahead.

	A	B	C	D	E
		Actual Revenue		Moving Averages	
1	Revenue Month	Amount			
2	Jun-14	\$ 47,926			
3	Jul-14	\$ 45,875		#N/A	
4	Aug-14	\$ 28,800		#N/A	
5	Sep-14	\$ 60,000		\$40,867	=AVERAGE(B2:B4)
6	Oct-14	\$ 58,051		\$44,892	
7	Nov-14	\$ 53,438		\$48,950	
8	Dec-14	\$ 52,290		\$57,163	
9	Jan-15	\$ 66,247		\$54,593	
10	Feb-15	\$ 55,453		\$57,325	
11	Mar-15	\$ 72,936		\$57,997	
12	Apr-15	\$ 44,000		\$64,879	
13	May-15	\$ 85,900		\$57,463	
14	Jun-15	\$ 82,253		\$67,612	
15	Jul-15	\$ 61,300		\$70,718	
16	Aug-15	\$ 76,000		\$76,484	
17	Sep-15	\$ 57,100		\$73,184	
18	Oct-15	\$ 92,470		\$64,800	
19	Nov-15	\$ 60,000		\$75,190	
20	Dec-15	\$ 84,000		\$69,857	
21	Jan-16	\$ 62,900		\$78,823	
22	Feb-16	\$ 76,500		\$68,967	
23	Mar-16	\$ 67,100		\$74,467	
24	Apr-16	\$ 59,700		\$68,833	
25	May-16	\$ 82,041		\$67,767	
26	Jun-16	\$ 61,300		\$69,614	
27	Jul-16	\$ 80,700		\$67,680	
28	Aug-16		Forecast: \$74,680		=AVERAGE(B25:B27)
29	Sep-16		Forecast: \$71,000		=AVERAGE(B26:B28)

**FIGURE 5-6:** Moving averages are usually limited to one-step-ahead forecasts.

Other chapters (such as Chapter 13 and Chapter 15) show you much more about how and why the formulas in Figure 5-6 make for good forecasts. For now, notice what happens when you stretch them too far: Like rubber bands, they break and snap back at you.

Look first at cell D5 in Figure 5-6. It's the average of cells B2, B3, and B4, and it's what the moving-average approach forecasts for September 2004. That is, the way this forecast is set up, the forecast for September is the average of June, July, and August. You can see the forecast of \$40,867 in cell D5, and the formula itself for illustration in cell E5.

The formula in D5 is copied and pasted down through cell D28, where it provides the "real" forecast for August 2016. I'm using "real" in the sense that I haven't yet seen an actual value for that month — my most recent actual value is for July 2016 — so August 2016 is past the end of the baseline and the forecast for that month is a real forecast. The formula itself appears in cell E28.

But if I copy and paste the formula one more row down, to try for a forecast for September 2016, I've stretched it too far. Now it's trying to average the actual results for June through August 2016, and I have no actual for August. Because of the way that Excel's AVERAGE works, it ignores cell B28 and the formula returns the average of B26 and B27.

The District Attorney will decline to prosecute if you're found shifting suddenly from a three-month moving average to a two-month moving average, but you really shouldn't. If you do, you're inviting an apple to mix with the oranges.

And if you take your forecast much farther down, it'll start returning the really nasty error value #DIV/0!. (That exclamation point isn't mine, it's Excel's, and it's meant to get your attention. Excel is yelling at you, "You're trying to divide by zero!")

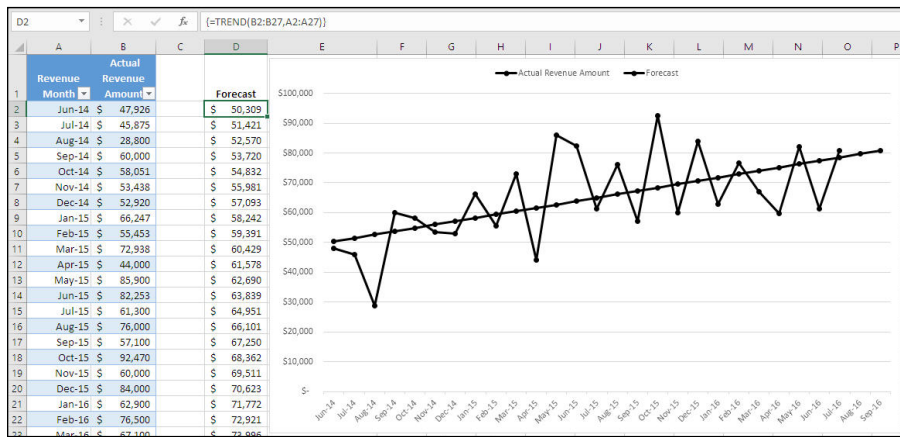
A similar situation occurs with exponential smoothing, and it's shown in Figure 5-7. The formula for smoothing is different from the formula for moving averages, but something similar happens when you get past the one-step-ahead forecast in cell D28.

Notice that the formula in cell D29 (the formula is shown in E29; the value that the formula returns appears in D29) uses the values in cells B28 and D28. But because we don't yet have an actual for August, the "forecast" for September 2016 is faulty: In fact, it's nothing more than the forecast for August multiplied by 0.7. Again, in this sort of exponential smoothing, you're limited to a one-step-ahead forecast.

Figure 5-8 shows a different situation, where the forecast is built using regression rather than moving averages or exponential smoothing.

	A	B	C	D	E
		Actual Revenue		Exponential Smoothing	
1	Revenue Month	Revenue Amount			
2	Jun-14	\$ 47,926		#N/A	
3	Jul-14	\$ 45,875		\$ 47,926	
4	Aug-14	\$ 28,800		\$ 47,311	
5	Sep-14	\$ 60,000		\$ 41,757	$=0.3*B4+0.7*D4$
6	Oct-14	\$ 58,051		\$ 47,230	
7	Nov-14	\$ 53,438		\$ 50,476	
8	Dec-14	\$ 52,290		\$ 51,365	
9	Jan-15	\$ 66,247		\$ 51,642	
10	Feb-15	\$ 55,453		\$ 56,024	
11	Mar-15	\$ 72,936		\$ 55,853	
12	Apr-15	\$ 44,000		\$ 60,978	
13	May-15	\$ 85,900		\$ 55,884	
14	Jun-15	\$ 82,253		\$ 64,889	
15	Jul-15	\$ 61,300		\$ 70,098	
16	Aug-15	\$ 76,000		\$ 67,459	
17	Sep-15	\$ 57,100		\$ 70,021	
18	Oct-15	\$ 92,470		\$ 66,145	
19	Nov-15	\$ 60,000		\$ 74,042	
20	Dec-15	\$ 84,000		\$ 69,830	
21	Jan-16	\$ 62,900		\$ 74,081	
22	Feb-16	\$ 76,500		\$ 70,727	
23	Mar-16	\$ 67,100		\$ 72,459	
24	Apr-16	\$ 59,700		\$ 70,851	
25	May-16	\$ 82,041		\$ 67,506	
26	Jun-16	\$ 61,300		\$ 71,866	
27	Jul-16	\$ 80,700		\$ 68,696	
28	Aug-16		Forecast: \$ 72,297	$=0.3*B27+0.7*D27$	
29	Sep-16		Forecast: \$ 50,608	$=0.3*B28+0.7*D28$	

**FIGURE 5-7:** If you want to forecast farther ahead, consider a regression forecast.



**FIGURE 5-8:** The trendline in the chart is taken from the worksheet. You can also get one from the Chart menu.

Using regression (see Chapter 11 for the basics and Chapter 16 for some refinements), you're in a different position than with moving averages and exponential smoothing. As Figure 5-8 shows, you can create your forecasts using date itself as a predictor: Each forecast value there is based on the relationship in the baseline between date and revenue.

Because I know the value of the next two dates, August and September 2016, I can use the relationship between date and revenue in the baseline on the next two

dates to get a forecast. The forecast values appear in cells C28 and C29 and show up in the chart as the final two points in the Forecast series.

Now, the farther out into the future you forecast using regression, the thinner the ice gets (or, if you prefer the earlier metaphor, the more strain you're putting on the rubber band). The farther you get from the end of your baseline, the more opportunities there are for the actuals to change direction — for example, to turn down or to level off.

If you have a real need to forecast, say, 12 months into the future on a monthly basis, and if you think there's a dependable relationship between date and revenue amount, then regression may be your best choice. But keep in mind that things get flaky out there in the future.

Another method to push your forecast out beyond a one-step-ahead approach is seasonal smoothing. This approach, which depends on a seasonal component in your baseline, can support a forecast that goes that year into the future. It ain't necessarily so, but it's possible. (Chapter 18 has information on the seasonal approach.)

## Choosing your time periods

Suppose that you *do* need to forecast out a ways from the present — a year, for instance. Here's where judgment enters the picture, along with the nature of your requirements.

If your baseline consists of several years, with actuals broken out by months, one thing you may consider is to change the baseline's time period from months to years. Then you can forecast the entire next year — although your forecasts would not be month-by-month. You would get the one-step-ahead forecast, and that one step would be the entire year (see Figure 5-9).

Here's what's going on in Figure 5-9:

- » Column A contains the month during which revenue was recognized. It extends down past the bottom of the visible worksheet area to December 2016.
- » Column B contains the revenue for each month.
- » The range D3:E8 contains a pivot table. (Excel's pivot tables are huge assets for forecasting, and you can find out how to use them in Chapter 8.) This pivot table converts the monthly data in columns A and B to annual data — the sum of the revenue for each year.

Revenue		Actual Revenue						
Month	Amount		Row Labels		Sum of Actual Revenue Amount		Forecast	
Jan-12	\$ 46,018							
Feb-12	\$ 54,791		2012		\$478,094	Year	Value	
Mar-12	\$ 48,444		2013		\$505,502		#N/A	
Apr-12	\$ 39,854		2014		\$490,132	2014	\$491,798	
May-12	\$ 29,923		2015		\$795,116	2015	\$497,817	
Jun-12	\$ 47,476		2016		\$962,957	2016	\$642,624	
Jul-12	\$ 33,566					2017	\$879,037	
Aug-12	\$ 30,481							
Sep-12	\$ 41,270							
Oct-12	\$ 46,174							
Nov-12	\$ 32,493							
Dec-12	\$ 27,604							
Jan-13	\$ 53,818							
Feb-13	\$ 31,615							
Mar-13	\$ 34,672							
Apr-13	\$ 45,006							
May-13	\$ 30,419							
Jun-13	\$ 57,230							
Jul-13	\$ 29,094							
Aug-13	\$ 54,418							
Sep-13	\$ 31,597							

**FIGURE 5-9:** Pivot tables are useful for summarizing baseline data into longer time periods.

- » The forecast for each year, using moving averages, is in the range G6:H8. Your one-step-ahead forecast for 2017 is in cell H9. In this case, the forecasts are based on two-year moving averages, rather than the three-period averages that appear in Figure 5-6.

The approach in Figure 5-9 has a couple drawbacks:

- » **Your baseline goes from 60 observations (monthly revenue over 5 years, a good long baseline) to 5 observations (yearly revenue over 5 years, a short baseline indeed).** Reducing the length of your baseline so drastically often causes misleading results. But because the monthly revenues show the same gradual growth as the annual totals, you can have some confidence in the annuals.
- » **Whoever requested the forecast — an accountant, a bank, a sales manager, a sales VP — could want to see the forecast for 2017 on a monthly basis.** If so, you're probably going to have to back up to the monthly baseline and use regression (*probably*, because you may be able to find the seasonality in the baseline that would support seasonal smoothing).

## Spacing Time Periods Equally

Getting your baseline's time periods to line up properly is important. It's also important that, within reason, each time period in your baseline represents the same length of time. Here's a closer look at each of these two issues.

## Using periodic relationships

Over time, a baseline tends to display consistent behavior: Its level is increasing, decreasing, or remaining stationary (or it may be seasonal or cyclic). The relationships between time periods help measure this behavior: the relationship between one month and the next, or between one quarter and the next, or between one quarter and the same quarter in the prior year.

Your baseline might mix up the relationships between its time periods for various reasons, some good and some bad. A couple of examples:

- » Whoever assembled the baseline data (not you, certainly) overlooked the sales revenues for June 15 through June 30. This is a real problem, and it's really indefensible. "The dog ate my homework" doesn't cut it here.
- » The warehouse burned to the ground and nobody could sell anything until the factory could catch up with the loss of inventory. Again, a real problem, but it doesn't help your forecast even if the police do catch the arsonist.

The reason is this: If almost all your baseline consists of monthly revenues, and one time period represents just half a month, any forecast that depends on the entire baseline will be thrown off. Figure 5-10 shows an example of what can happen.

In Figure 5-10, cells A1:B27 contain a baseline with accurate revenues throughout. Exponential smoothing gives the forecast for August 2016 in cell C28.

Also in Figure 5-10, cells H1:I27 have the same baseline, except for cell I25. For some reason (careless accounting, that warehouse fire, or something else), the revenue for May 2016 has been underreported. The result is that the forecast for August 2016 is more than \$6,000 less than it is when the May 2016 revenues are the result of neither an error nor a one-time incident. Six thousand dollars may not sound like a lot, but in this context it's an 8 percent difference. And it's even worse right after the problem occurs: The difference in the two forecasts is 17 percent in June 2016.

If this happened to me, I'd send someone back to the files — whether hard- or soft-copy — to fill in the missing data for May 2016.

		Actual			Actual		
Revenue	Revenue	Exponential		Revenue	Revenue	Exponential	
Month	Amount	Smoothing		Month	Amount	Smoothing	
Jun-14	\$ 47,926	#N/A		Jun-14	\$ 47,926	#N/A	
Jul-14	\$ 45,875	\$ 47,926		Jul-14	\$ 45,875	\$ 47,926	
Aug-14	\$ 28,800	\$ 47,311		Aug-14	\$ 28,800	\$ 47,311	
Sep-14	\$ 60,000	\$ 41,757		Sep-14	\$ 60,000	\$ 41,757	
Oct-14	\$ 58,051	\$ 47,230		Oct-14	\$ 58,051	\$ 47,230	
Nov-14	\$ 53,438	\$ 50,476		Nov-14	\$ 53,438	\$ 50,476	
Dec-14	\$ 52,290	\$ 51,365		Dec-14	\$ 52,290	\$ 51,365	
Jan-15	\$ 66,247	\$ 51,642		Jan-15	\$ 66,247	\$ 51,642	
Feb-15	\$ 55,453	\$ 56,024		Feb-15	\$ 55,453	\$ 56,024	
Mar-15	\$ 72,936	\$ 55,853		Mar-15	\$ 72,936	\$ 55,853	
Apr-15	\$ 44,000	\$ 60,978		Apr-15	\$ 44,000	\$ 60,978	
May-15	\$ 85,900	\$ 55,884		May-15	\$ 85,900	\$ 55,884	
Jun-15	\$ 82,253	\$ 64,889		Jun-15	\$ 82,253	\$ 64,889	
Jul-15	\$ 61,300	\$ 70,098		Jul-15	\$ 61,300	\$ 70,098	
Aug-15	\$ 76,000	\$ 67,459		Aug-15	\$ 76,000	\$ 67,459	
Sep-15	\$ 57,100	\$ 70,021		Sep-15	\$ 57,100	\$ 70,021	
Oct-15	\$ 92,470	\$ 66,145		Oct-15	\$ 92,470	\$ 66,145	
Nov-15	\$ 60,000	\$ 74,042		Nov-15	\$ 60,000	\$ 74,042	
Dec-15	\$ 84,000	\$ 69,830		Dec-15	\$ 84,000	\$ 69,830	
Jan-16	\$ 62,900	\$ 74,081		Jan-16	\$ 62,900	\$ 74,081	
Feb-16	\$ 76,500	\$ 70,727		Feb-16	\$ 76,500	\$ 70,727	
Mar-16	\$ 67,100	\$ 72,459		Mar-16	\$ 67,100	\$ 72,459	
Apr-16	\$ 59,700	\$ 70,851		Apr-16	\$ 59,700	\$ 70,851	
May-16	\$ 82,041	\$ 67,506		May-16	\$ 41,020	\$ 67,506	
Jun-16	\$ 61,300	\$ 71,866		Jun-16	\$ 61,300	\$ 59,560	
Jul-16	\$ 80,700	\$ 68,696		Jul-16	\$ 80,700	\$ 60,082	
Aug-16 Forecast		\$ 72,297		Aug-16 Forecast		\$ 66,267	

**FIGURE 5-10:**  
Bad data from a recent time period can lead to a bad forecast.

If the missing data can't be located, due perhaps to an accounting error, or if no error was made but some really unusual incident interrupted the sales process during May 2016, I'd probably estimate the actuals for May. A couple of reasonable ways to do that:

- » Take the average of April and June and assign that average to May.
- » Use June 2014 through April 2016 as a baseline, and forecast May 2016. Then use that May 2016 forecast in your full baseline, January 2014 through July 2016.



TIP

Chapter 9 stresses the importance of charting your baseline. This situation is another good reason. Just looking at the baseline, you might not notice that May 2016 is an oddball. But it jumps right out at you if you chart the baseline — see Figure 5-11, particularly June through August 2016 in each chart.

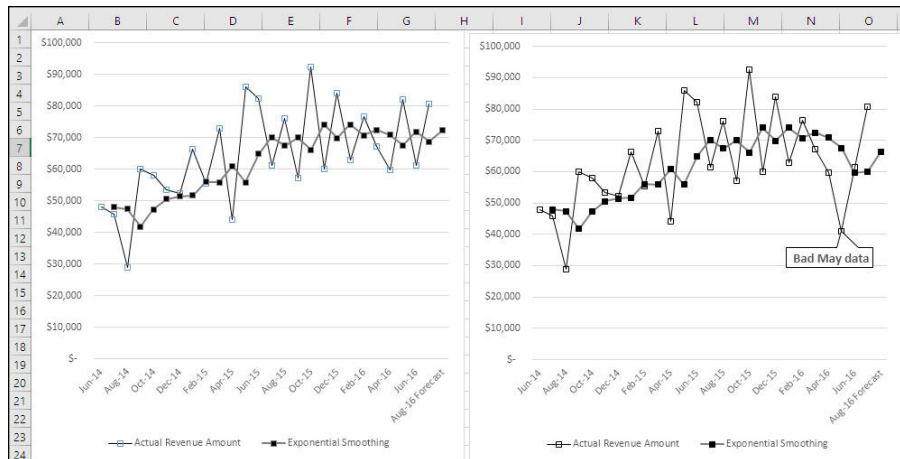


TIP

Don't worry about small differences in the length of the baseline's time periods. March has one more day in it than April does, but it's not worth worrying about. Two missing weeks is another matter.



**FIGURE 5-11:**  
Oddball data jumps out at you when you chart the baseline.



## When missing data causes unequal time periods

When you're working with forecasts that are based on moving averages and on exponential smoothing, you're working with forecasts that depend on a baseline with consecutive time periods. You might have a sequence of monthly sales revenues that make for a good, sound forecast into the next month, due to the relationship between the sales in consecutive months. But if some of those months are missing, you could be in trouble. Figure 5-12 shows how a full baseline works.

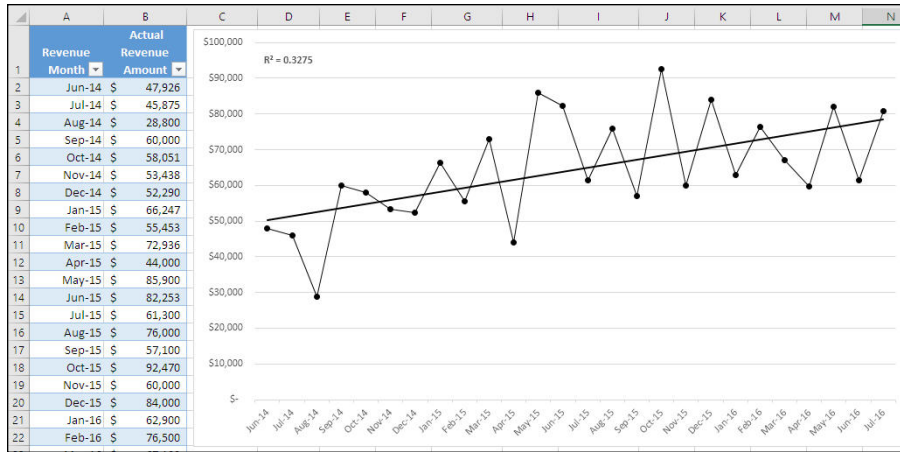
Notice the straight line in the chart in Figure 5-12. It's called a *trendline*. The trendline indicates how well the gradual growth in revenues tracks against the dates when the revenues were recognized. Just what you like to see. Generally, the greater the incline in the trendline, the stronger the relationship between the time period and the revenue, and this is a pretty decent result.



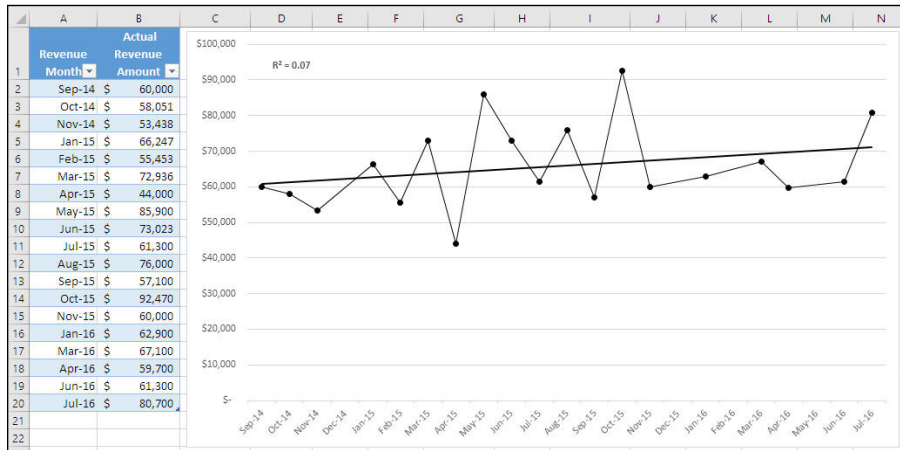
By the way, the data in Figure 5-12 is the same as in Figure 5-8. The difference is due to how the trendline is created. In Figure 5-8, the trend is calculated on the worksheet and that trend is charted explicitly. In Figure 5-12, the trendline is created by right-clicking the charted data series and choosing Add Trendline from the shortcut menu. Charting the trendline as a separate series gives you more control; using Add Trendline is faster.

Contrast the chart in Figure 5-12 with the one in Figure 5-13.

**FIGURE 5-12:**  
Excel can calculate the trendline and the equation for you.



**FIGURE 5-13:**  
The smaller the R-squared value, the less dependable the forecast. An R-squared of 0.07 is small.



The dates and revenues in Figure 5-13 are the same as in Figure 5-12, except that seven months are missing, some from the start of the baseline and a few from well into the baseline. Now, the relationship between the time period and its revenue has been disrupted by the missing periods, and the trendline in the chart has become nearly horizontal, indicating a much weaker relationship.

You wouldn't be able to put much trust in a forecast constructed from the baseline in Figure 5-13, even though the data that's in it is a subset of the same as the data in Figure 5-12. The lesson: Arrange for a baseline that's chock-full of consecutive time periods — or don't bother with a forecast.

Understanding how tables are structured

Moving from a list to a table

Looking at the records you're interested in

Getting your data into an Excel table

## Chapter 6

# Setting Up Tables in Excel

Excel is *not* a database program like Access or SQL Server or Oracle. Sure, you can store data in it, but it's really not intended to store large amounts of data or to manage relationships between different data sets.

Still, Excel has a rudimentary way, called a *list*, to store your data, whether it's sales data or something else. An Excel list is nothing more than data entered in adjacent columns and adjacent rows, such as the range C5:F20. The understanding is that different rows in the list contain different records, and that different columns contain different variables (or *fields*). The first row in the list often contains the names of the fields.

In Excel 2016, the list still exists in Excel as an informal way of arranging data. But Excel 2007 introduced a new structure called a *table*. A table is a formal object, as an Excel chart or an Excel worksheet is a formal object. A table has a formal name such as Table1, just as a chart has a formal name such as Chart1. A table knows how many records and fields it contains, just as a worksheet knows where its bottommost-used row and its rightmost-used column are located.

You can convert a simple list to a more sophisticated table with a couple of mouse clicks. Go to the Ribbon's Insert tab, select at least one cell in an existing list, and click the Table icon in the Tables group. You can also go the other direction and

convert an existing table back to a list: Select a cell in the table — a Design tab will appear on the Ribbon — and click the Convert to Range icon in the Tools group on that tab.

Tables are important as the basis for your forecasts: Sometimes you use them directly, and sometimes you use them indirectly as the basis for pivot tables. In this chapter, I show you how to set up tables and how to filter them so you can focus on specific sets of records.

You can enter new records to a table simply by typing the data into the first row below the table's current final row. You can establish a new column in the table in a similar way. This chapter shows you how to do that, and how to call for an automatic Totals row at the bottom of a table.

Because you sometimes want to forecast on the basis of a subset of the data in a table (for example, a sales forecast for the Northwest region only, or a sales forecast for dishwasher detergents only), you sometimes want to filter for those subsets. I show you how to do that here.

The process of importing data from another application such as a database can be time consuming and tedious unless you know the most effective way. This chapter shows you how to automate importing the data to a table or list on an Excel worksheet.

## Understanding Table Structures

An Excel table has a fairly simple structure. It has these characteristics:

- » **Different fields, also known as *variables*, are in different columns.** Notice in Figure 6-1 that three columns contain data, and each column contains a different field. Keeping the same sort of data — such as sales dates, product lines, sales regions, and so on — in one column is a good idea. Don't stick a markup percent in a column full of names of product lines.
- » **Different records are in different rows.** Each row in the list represents a different record. In this case, each record is a different sale, and the date, the sales rep, and the revenue are shown for each sales record.
- » **The columns have labels.** Each column in the table has a label in the first row. Here, the labels are Date, Sales Rep, and Revenue. This is not a requirement, but if you don't supply your own labels, Excel does so for you: Column1, Column2, Column3, and so on.

	A	B	C	D
1	Date	Sales Rep	Revenue	
2	12/9/16	Peters	\$ 9,237	
3	10/14/16	Williams	\$ 949	
4	11/5/16	Peters	\$ 6,829	
5	7/10/16	Johnson	\$ 9,946	
6	2/18/16	Simpson	\$ 5,777	
7	9/5/16	Williams	\$ 2,299	
8	11/19/16	Davis	\$ 6,530	
9	2/23/16	Johnson	\$ 7,437	
10	2/16/16	Simpson	\$ 4,428	
11	1/15/16	Edwards	\$ 9,164	
12	1/3/16	Edwards	\$ 9,858	
13	9/15/16	Davis	\$ 7,488	
14	11/26/16	Thompson	\$ 4,128	
15	10/18/16	Peters	\$ 9,151	
16	3/29/16	Thompson	\$ 9,451	
17	10/20/16	Simpson	\$ 8,578	
18	11/21/16	Davis	\$ 4,136	
19	6/29/16	Johnson	\$ 1,168	
20	2/2/16	Williams	\$ 4,453	

**FIGURE 6-1:**  
The range A1:C20  
contains a table.

Within the limits of an Excel worksheet's size (16,384 columns and 1,048,576 rows), there's no limit to the size of a list. It could have anywhere from 1 to 16,384 columns, and anywhere from 2 to 1,048,576 rows. Of course, you'd have to be nuts to put that much data in a worksheet — if you have that much data, use a true database instead.

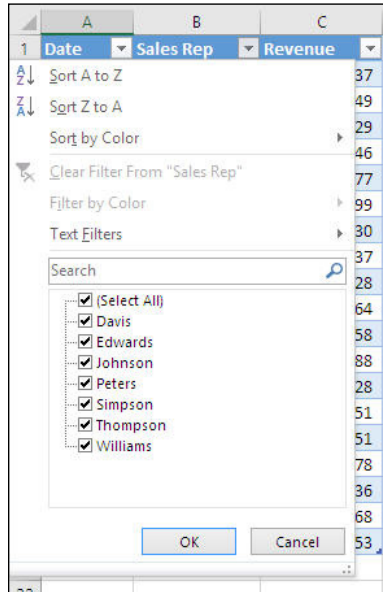
Speaking of databases, if you're familiar with how tables look in a database, you'll see the resemblance between the Excel table and the database table: fields in columns, records in rows, labels at the tops of the columns. In fact, an Excel table (or list, for that matter) is really just what, years ago, Excel called a database.

In Figure 6-1, the records aren't sorted by date, sales rep, or revenue. A table doesn't have to be sorted to be a table.

Why would you want to put a table like the one in Figure 6-1 into an Excel worksheet? Because a table is the basis for many kinds of analysis, including forecasting. For example:

» **If you do want to sort the records in the table, Excel gives you a couple of ways.** The fast way is to click a header cell's drop-down arrow and choose one of the sorting options. So in Figure 6-1, you could click the drop-down arrow in cell B1 to select a sort by Sales Rep name. But this way allows only one sort key. If you wanted to sort by Date and by Revenue within Date, select any cell in the table, click the Ribbon's Data tab, and click the Sort icon in the Sort & Filter group. Select a field as the primary sort key, and then click the Add button to include a secondary sort key.

» **Tables make filtering your data easy.** Just click a header cell in the table and choose one of the filtering options. For example, to show only the records for the sales rep named Davis, click the drop-down arrow in cell B1 shown in Figure 6-2. Clear the Select All check box to deselect all sales reps and then fill the check box for Davis. If you'd rather do without the drop-down arrows, turn them off by selecting any cell in the table, going to the Ribbon's Data tab, and clicking the Filter icon in the Sort & Filter group.



**FIGURE 6-2:** The drop-down lists offer you the values in their columns as well as a top-ten option.

» **Pivot tables make it easy to summarize data stored in a table or even in a list.** Pivot tables that group the records by date (week, month, quarter, or year) are the basis for many of the forecasting examples in this book. But if you're building a pivot table, using data in an Excel worksheet, the raw data has to be structured as a table or list. Figure 6-3 shows a pivot table based on the table you've been looking at.



REMEMBER

You have to follow the table structure carefully if you're preparing to create a pivot table. For example, if you forgot to put a label at the top of one of a table's columns, Excel would cooperate in creating the pivot table for you — but would replace the column header you omitted with something such as *Column1*. You probably don't want that used as a field name in a pivot table. A list isn't as cooperative as a table. If you omit a column header from a list, and then try to base a pivot table on that list, you'd get an error message telling you that one of your field names wasn't valid (because the cell where Excel expected to find a column header would be empty).

	A	B	C	D	E	F	G
1	Date	Sales Rep	Revenue				
2	12/9/16	Peters	\$ 9,237			Row Labels	Sum of Revenue
3	10/14/16	Williams	\$ 949			Davis	\$18,154
4	11/5/16	Peters	\$ 6,829			Edwards	\$19,022
5	7/10/16	Johnson	\$ 9,946			Johnson	\$18,551
6	2/18/16	Simpson	\$ 5,777			Peters	\$25,217
7	9/5/16	Williams	\$ 2,299			Simpson	\$18,783
8	11/19/16	Davis	\$ 6,530			Thompson	\$13,579
9	2/23/16	Johnson	\$ 7,437			Williams	\$7,701
10	2/16/16	Simpson	\$ 4,428			Grand Total	\$121,007
11	1/15/16	Edwards	\$ 9,164				
12	1/3/16	Edwards	\$ 9,858				
13	9/15/16	Davis	\$ 7,488				
14	11/26/16	Thompson	\$ 4,128				
15	10/18/16	Peters	\$ 9,151				
16	3/29/16	Thompson	\$ 9,451				
17	10/20/16	Simpson	\$ 8,578				
18	11/21/16	Davis	\$ 4,136				
19	6/29/16	Johnson	\$ 1,168				
20	2/2/16	Williams	\$ 4,453				

**FIGURE 6-3:**  
The pivot table automatically sorts itself by the data in the Row field — here, that’s the sales rep’s name.

If for some reason you want to omit a record — that is, have a blank row in the middle of the table — that’s okay (although the blanks may show up in the pivot table as blanks, confusing everyone). But you shouldn’t have blank columns, at least not if you’re going to create a pivot table using the table. Blank columns result in the same problem you get if you forget to label one of the table’s columns — you get an error message.

## Creating a Table

The most straightforward way to create a table in Excel is to start with a list: a rectangular range of cells with different fields in different columns, and column labels with field names in the first row. Subsequent rows have individual records.

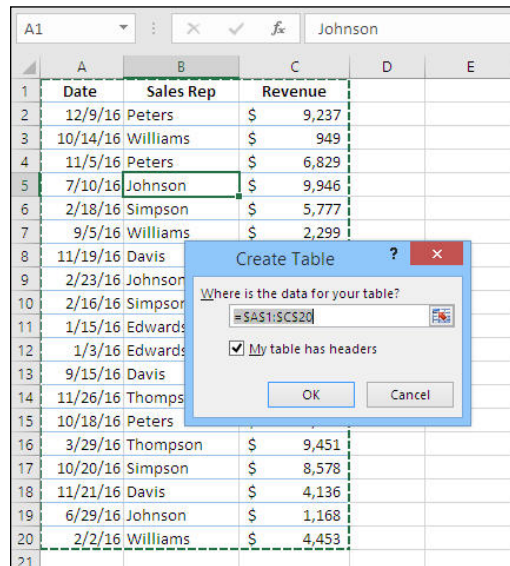
Then, select any cell in the list. Go to the Ribbon’s Insert tab and click the Table icon in the Tables group. Excel displays a dialog box where you can confirm the address of the existing list, and where you can indicate whether the list contains column headers. When you click OK, Excel changes the list to a table, probably one with banded rows and drop-down arrows in the Header row, as shown in Figure 6-4.



TECHNICAL  
STUFF

Notice that the dialog box in Figure 6-4 makes a liar out of me. I’ve been telling you that you need to put labels at the top of each column in your list. But the Create Table dialog box has a My Table Has Headers check box, which allows you specify whether your list has *headers* (one way that Excel refers to those column labels). The truth is that you *can* start the process with a list that doesn’t have column labels. (In fact, you can sort a list that’s missing one or more column

labels, and you can filter a list with missing labels.) But use column headers (or labels, if you prefer that term) anyway. Why? Because you can't build a pivot table if a label is missing. And if all the labels are missing, Excel will assume that the first row of data contains the field names for the pivot table, and that leads to ridiculous results.



**FIGURE 6-4:** If you start by clicking in the list, Excel automatically proposes the full list range for the table.



**WARNING**

If the My Table Has Headers check box is cleared, Excel puts headers in on your behalf. It shifts the list down a row and inserts the labels Column1, Column2, and so on as column headers. So be careful: If for some other reason you need the list to stay where it was, be sure that you've provided your own column labels. If you have data below the list, Excel shifts it down too when it converts the list to a table, to prevent an over-write.

After you're sure that Excel has figured out properly where your existing list is, and you've select the My Table Has Headers check box (if appropriate), click OK. Excel converts the list to a table, and you'll see something similar to Figure 6-5.

It's easy to add data to an existing table. Just select a cell in the row immediately below the bottom of the table, and in a column that the table occupies. Type a value, and when you press Enter Excel extends the table to capture the new record. You can also start in the cell that is a table's lower-right corner and press the Tab key to establish a new record.



	A	B	C	D
1	Date	Sales Rep	Revenue	
2	12/9/16	Peters	\$ 9,237	
3	10/14/16	Williams	\$ 949	
4	11/5/16	Peters	\$ 6,829	
5	7/10/16	Johnson	\$ 9,946	
6	2/18/16	Simpson	\$ 5,777	
7	9/5/16	Williams	\$ 2,299	
8	11/19/16	Davis	\$ 6,530	
9	2/23/16	Johnson	\$ 7,437	
10	2/16/16	Simpson	\$ 4,428	
11	1/15/16	Edwards	\$ 9,164	
12	1/3/16	Edwards	\$ 9,858	
13	9/15/16	Davis	\$ 7,488	
14	11/26/16	Thompson	\$ 4,128	
15	10/18/16	Peters	\$ 9,151	
16	3/29/16	Thompson	\$ 9,451	
17	10/20/16	Simpson	\$ 8,578	
18	11/21/16	Davis	\$ 4,136	
19	6/29/16	Johnson	\$ 1,168	
20	2/2/16	Williams	\$ 4,453	
21				

**FIGURE 6-5:**  
A new table has banded rows by default.

Excel lets you do some special things with tables that you've set up. I go into these special features in the following sections.

## Using the Total row

Click in the table to display the Design tab for tables. Go to the Design tab and fill the Total Row check box in the Table Style Options group. You'll get a new row below the table. The new row shows you summaries for your data (see Figure 6-6).

The Total row automatically gives you a summary of the rightmost column in the list. If the column contains numeric data only, you'll get a sum. If it contains text data, you'll get a count of the values.

But each cell in the Total row displays a drop-down list if you click its drop-down arrow. If you click that, you'll see a selection of summary statistics that you can choose from, similar to the summaries you can choose in a pivot table:

- » **None:** The user wants a Total row, but doesn't want the selected cell to display a summary.
- » **Average:** The sum of the numeric values divided by the number of numeric values.
- » **Count:** How many values there are, whether numeric or text.

	A	B	C	D	E
1	Date	Sales Rep	Revenue		
2	12/9/16	Peters	\$ 9,237		
3	10/14/16	Williams	\$ 949		
4	11/5/16	Peters	\$ 6,829		
5	7/10/16	Johnson	\$ 9,946		
6	2/18/16	Simpson	\$ 5,777		
7	9/5/16	Williams	\$ 2,299		
8	11/19/16	Davis	\$ 6,530		
9	2/23/16	Johnson	\$ 7,437		
10	2/16/16	Simpson	\$ 4,428		
11	1/15/16	Edwards	\$ 9,164		
12	1/3/16	Edwards	\$ 9,858		
13	9/15/16	Davis	\$ 7,488		
14	11/26/16	Thompson	\$ 4,128		
15	10/18/16	Peters	\$ 9,151		
16	3/29/16	Thompson	\$ 9,451		
17	10/20/16	Simpson	\$ 8,578		
18	11/21/16	Davis	\$ 4,136		
19	6/29/16	Johnson	\$ 1,168		
20	2/2/16	Williams	\$ 4,453		
21	Total		\$ 121,007		
22					

**FIGURE 6-6:**  
The Total row remains in place after you select some cell outside the list.

- » **Count Numbers:** How many numeric values there are.
- » **Max:** The largest numeric value.
- » **Min:** The smallest numeric value.
- » **Sum:** The total of the numeric values.
- » **StdDev:** A quantity that describes how much variability, or spread, there is around the field's average value.
- » **Var:** The square of the standard deviation.
- » **More functions. . .** Clicking this item opens the Insert Function dialog box, so you can select any summary value from the median to the imaginary coefficient of a complex number.



TIP

You can use Count Numbers if a column in the table mixes numeric values with text values. Having a column that mixes numeric and text values usually doesn't make much sense, though — you normally want to store the same kind of information in each column, and if you've mixed text and numeric values, then you probably haven't done that.

You can arrange for a different summary statistic in each cell of the Total row. In the table we've been looking at, you could get the earliest sales date, a count of the records, and the sum of the revenue (see Figure 6-7).

	A	B	C	D	E	F
1	Date	Sales Rep	Revenue			
2	12/9/16	Peters	\$ 9,237			
3	10/14/16	Williams	\$ 949			
4	11/5/16	Peters	\$ 6,829			
5	7/10/16	Johnson	\$ 9,946			
6	2/18/16	Simpson	\$ 5,777			
7	9/5/16	Williams	\$ 2,299			
8	11/19/16	Davis	\$ 6,530			
9	2/23/16	Johnson	\$ 7,437			
10	2/16/16	Simpson	\$ 4,428			
11	1/15/16	Edwards	\$ 9,164			
12	1/3/16	Edwards	\$ 9,858			
13	9/15/16	Davis	\$ 7,488			
14	11/26/16	Thompson	\$ 4,128			
15	10/18/16	Peters	\$ 9,151			
16	3/29/16	Thompson	\$ 9,451			
17	10/20/16	Simpson	\$ 8,578			
18	11/21/16	Davis	\$ 4,136			
19	6/29/16	Johnson	\$ 1,168			
20	2/2/16	Williams	\$ 4,453			
21	1/3/16	19	121,007			

**FIGURE 6-7:** You can get the earliest date by choosing the Min summary for the Date column.



TIP

A quicker way to put a Total row at the bottom of a table is to right-click any cell in the table. You'll see a shortcut menu with Table as one of its items. Click the Table item and then click Totals Row. It's a toggle, so you can also use it to remove an existing Total row.

Suppose you begin to build a pivot table based on a table with a Total row. You select a single cell inside the table, go to the Ribbon's Insert tab, and click the pivot table icon in the Tables group. (Or, with a table cell selected, go to the Ribbon's Design tab and click Summarize With PivotTable.) When the Create PivotTable dialog box appears, you'll find the name of the selected table in the Table/Range edit box. Notice that a dashed border surrounds the part of the table that contains the actual data, omitting both the Header row and the Total row.

This process — starting with a cell inside the table, starting by selecting one cell only, and then inserting a pivot table — is usually the way to begin building a pivot table from a table. If you start by selecting a cell outside the table, you'll need to type or drag through the table to show Excel where your data is. You'll also need to be careful not to include the Total row in the range of cells you select — you don't want to duplicate totals in your pivot table.



TIP

## Using other table features

You might find that some of the following actions with tables are handy:

- » **Convert a table back to just a normal range.** Click inside a table that you've created, go to the Ribbon's Design tab, and choose Convert to Range in the Tools group. Your data will assume a list structure, and the special features of a table — in particular, the Total row and the Header row — will be gone. If you included a Total row, its summaries stay in place.
- » **Use a table's name in formulas.** A table has a defined name, like any other range of cells that you might name using the Define Name icon on the Formulas tab. The table's fields are also named. So you can use those names (which are termed *structured references*) in formulas that are outside the table. For example, in Figure 6-7, suppose that Excel has named the table as Table575. You could enter this formula to calculate the sum of the Revenue values in that table:

```
=SUM(Table575[Revenue])
```

- » **Resize a table.** I pointed out earlier that if you want to add a row or a column to a table, one way is to just start typing in a cell immediately below or to the right of an existing table. If you want to add rows or columns wholesale, click somewhere in the table and go to the Ribbon's Design tab. Click the Resize Table icon in the Properties group; a Resize Table dialog box appears. Use its range edit box to drag through the range that you want the table to occupy. The resized table's Header row must be in the same row as that of the existing table, and the resized table must overlap the existing table.

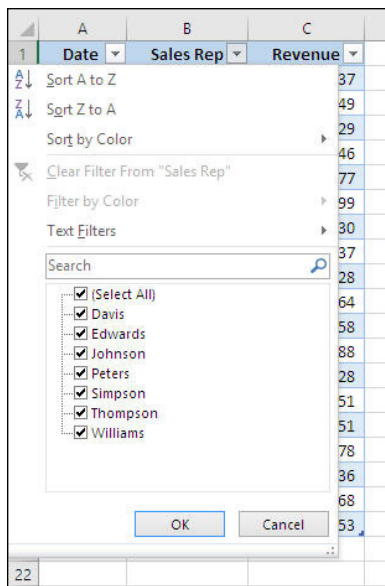
## Filtering Lists

You frequently want to filter a list so as to look more closely at specific records. In the context of forecasting sales, for example, you might want to focus on sales results for a particular sales rep, territory, or product, or for a particular date range. Excel offers you a couple of ways to do this. One is quick and easy, and the other takes just a little more time, but you get more ways to filter.

## Using Excel's table filters

As earlier sections of this chapter mention, when you create a table, Excel by default includes drop-downs that sort and filter in the table's Header row.

To filter a table, click one of the drop-down arrows in the table's Header row. (You do not need to have already selected a cell in the table.) You'll see a list box containing first some sorting options, and then all the unique values in that column, plus a few special options (see Figure 6-8). In the figure, to focus in on just one sales rep, you would clear the Select All check box to clear all the check boxes, and then click the check box by the name of the rep you're interested in.



**FIGURE 6-8:** Excel filters the records by hiding rows that don't conform to the filter that you set.

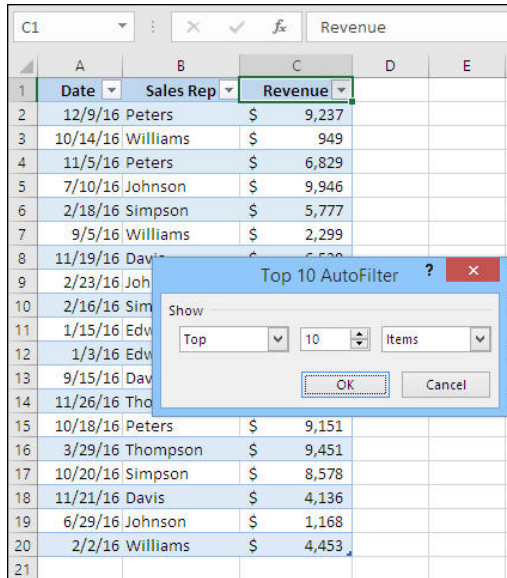
The sorting capability of the table's drop-downs isn't as sophisticated as you get by going to the Ribbon's Data tab and choosing Sort in the Sort & Filter group, but it's quick. You can sort by only one column at a time, whereas using the Sort icon in the Sort & Filter group does not limit the number of sort keys you can use.



TIP

If you've already used a filter on your table and you want to return all the records, click the drop-down again and place a check mark in the Select All check box, or click the Clear Filter From item. You can tell which drop-downs are being used to filter your data, because the arrow on the drop-down turns into the outline of a funnel when it's in use.

If you're filtering on a numeric field, such as Revenues, you may want to focus on the largest or the smallest value. Select (Top 10 . . .) from the drop-down list. You'll get the Top 10 AutoFilter dialog box, as shown in Figure 6-9.



**FIGURE 6-9:** The Top 10 AutoFilter dialog box allows you to show either a number of items or a percent of items.

To use the Top 10 AutoFilter, follow these steps:

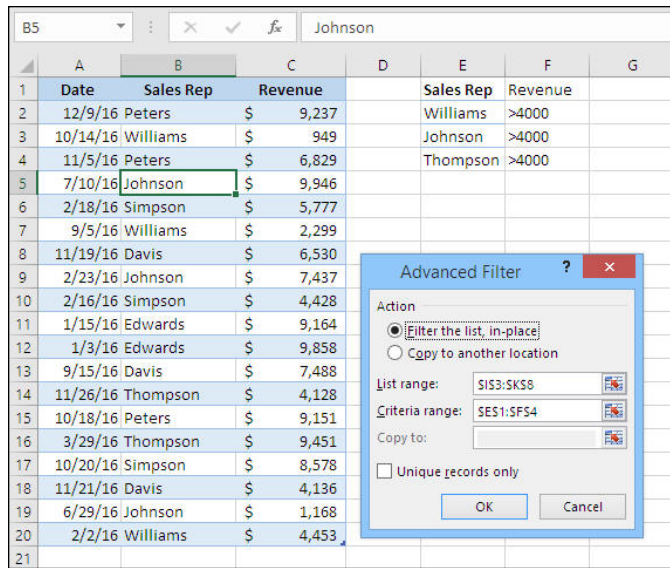
1. Click the drop-down in the Header row of a numeric field, such as Revenue in Figure 6-9.
2. Click Number Filters in the list box, and then click Top 10 in the cascading menu.
3. The Top 10 AutoFilter dialog box appears. In the leftmost drop-down list, select either Top (the largest values) or Bottom (the smallest values).
4. In the center drop-down list, select however many values you want to see.
5. In the rightmost drop-down list, select either Items or Percent.

For example, in Figure 6-9, if you chose Top, 10, and Items, respectively, you'd see the largest ten Revenue values. If you chose Top, 20, and Percent, respectively, you'd see the top 20 percent of Revenue values.

## Using the Advanced Filter

If you want to do a little more in the way of filtering a list, go to the Ribbon's Data tab and choose the Advanced icon from the Sort & Filter group. The Advanced Filter dialog box, shown in Figure 6-10, appears.

**FIGURE 6-10:**  
The Advanced Filter dialog box allows you greater control over the filtering.



The Advanced Filter dialog box allows you to apply a criteria list to select records with specific values (which you could do with the check boxes in a table's drop-down). You can also copy unique values only to a separate location, or copy a filtered list to a separate location. Suppose you wanted to see just the sales belonging to Williams, Johnson, and Thompson. You set up a criteria range in a different part of the worksheet — E1:E4 in Figure 6-10. Put the column label for the column you want to filter on in the first row of the criteria range, and underneath it put the values you want to select.

Then click one of the table's cells and choose Advanced from the Sort & Filter group on the Ribbon's Data tab. The Advanced Filter dialog box will automatically fill the List Range box with the address. Click in the Criteria Range box and drag through the cells where you put the criteria range. Click OK, and you'll see the records that meet your criteria.



TIP

To get all the records back, just click the Filter icon in the Sort & Filter group.

You can use more than one set of criteria. If in addition to the sales rep criteria you wanted to see only records with revenues greater than \$4,000, you could enter **Revenue** in cell F1 and **>4000** in each cell in the range F2:F4. Adjust the criteria range address in the Advanced Filter dialog box to E1:F4, and click OK.



TIP

If you want to create a new list from the filtered data, select the Copy to Another Location radio button. The Copy To box becomes enabled, so you can click in it and then click in a worksheet cell where you want your new list to start. Note that Excel creates a new *list*, not a new *table*, when you use the Advanced Filter in this way.

Suppose you wanted a list of the sales reps' names, but you didn't want them repeated as they are in your original list. Follow these steps:

1. **Click Advanced in the Sort & Filter group on the Data tab.**
2. **Click in the List Range box (yes, even if your data is in a table) and — as the data is laid out in Figure 6-10 — drag through B1:B20 to capture the column label and the sales reps' names.**
3. **Choose to copy the data to another location and indicate that location.**
4. **Select the Unique Records Only check box and click OK.**

You now have a new list that contains the sales reps' names, one apiece.

## Importing Data from a Database to an Excel Table

If you're setting up a table so you can use it to help forecast sales, there's a good chance that the data is stored in your company's accounting system. Many companies use a database management system for accounting purposes and Excel for financial analysis. If you're in that situation, or a similar one, there's an easy way to get the data out of the database and into an Excel table. Take the following steps. (They assume you have Microsoft Access readily available to you. I can't assume that you have Sage 100 or Oracle Financials on your laptop.)

1. **Go to the Ribbon's Data tab, click the Get External Data icon, and then click From Access.**

The Select Data Source dialog box, shown in Figure 6-11, appears.

2. **When you've located your data source, select it and click Open.**

If more than one table is in the data source, you'll see the Select Table dialog box, shown in Figure 6-12.

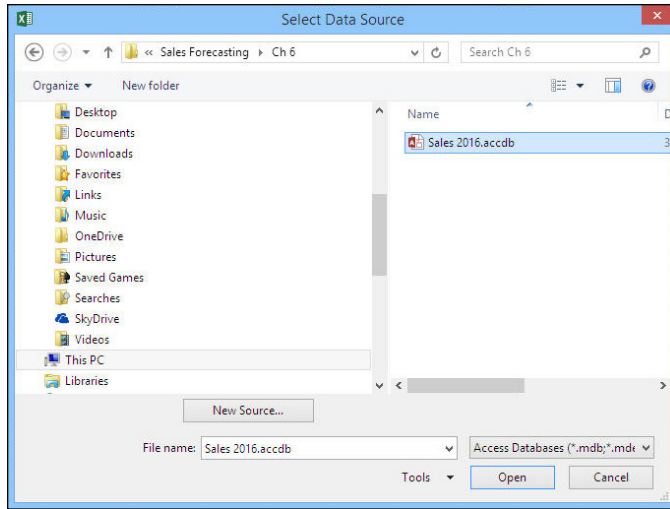
3. **Select the table you want and click OK.**

The Import Data dialog box, shown in Figure 6-13, appears.

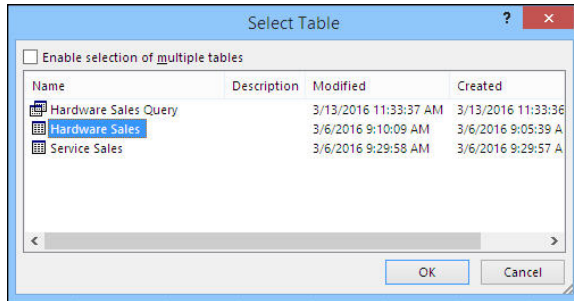
4. **Select where you want to put the data and click OK.** Figure 6-14 shows what the data might look like (including each record's unique ID number, automatically supplied by Access).



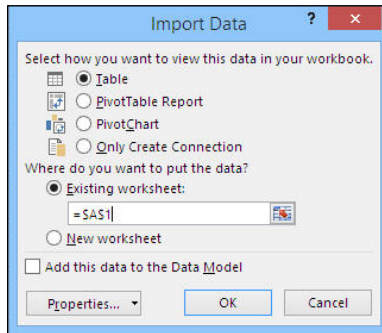
**FIGURE 6-11:** Use the Select Data Source dialog box to browse to the location where your database is stored.



**FIGURE 6-12:** In the Select Table dialog box, pick the table or query that you want to import from.



**FIGURE 6-13:** Use the Properties button to control various aspects of the table, which will occupy an external data range.



	A	B	C	D
1	ID	Sales Date	Sales Rep	Revenue
2	1	12/9/2016	Peters	9237
3	2	10/14/2016	Williams	949
4	3	11/5/2016	Peters	6829
5	4	7/10/2016	Johnson	9946
6	5	2/18/2016	Simpson	5777
7	6	9/5/2016	Williams	2299
8	7	11/19/2016	Davis	6530
9	8	2/23/2016	Johnson	7437
10	9	2/16/2016	Simpson	4428
11	10	1/15/2016	Edwards	9164
12	11	1/3/2016	Edwards	9858
13	12	9/15/2016	Davis	7488
14	13	11/26/2016	Thompson	4128
15	14	10/18/2016	Peters	9151
16	15	3/29/2016	Thompson	9451
17	16	10/20/2016	Simpson	8578
18	17	11/21/2016	Davis	4136
19	18	6/29/2016	Johnson	1168
20	19	2/2/2016	Williams	4453
21				

**FIGURE 6-14:** Consider using the database queries to manage what data returns to Excel, such as the record ID and tables that are linked by shared keys.



TIP

You could, of course, open the database directly, open the table, and do a copy and paste — and that probably seems easier. But the nice thing about doing it the longer way is that your table will save not only the data, but also information about the database’s location. The next time you want to bring the data into your worksheet, just right-click a cell in the worksheet’s table, choose Refresh Data from the context menu, and the most current information will flow from the database into the table.

## Chapter 7

# Working with Tables in Excel

In Chapter 6, I show you how to set up a table — what a table's structure is like, how to get Excel to help you manage it, and how to focus on specific items in the table. In other words, the dull routine of data management. Not fun, not always interesting, but necessary.

In this chapter, I get into more interesting things — certainly more colorful things. Excel's charts are great ways to visualize what's going on with the data in your table. They put you in a position to see patterns that are buried in the numbers. And you can use charts in some subtle ways to get more deeply into the business of sales forecasting.

I also show you how to use Excel's Data Analysis add-in, how it interacts with tables, and how it helps you build your forecasts (as well as some annoying traps to avoid).

## Turning Tables into Charts

Excel has a laundry list of charts for you to choose from — check out the Charts group on the Ribbon's Insert tab to see some of your options. Looking at them, you may think that making the choice is just a matter of which type looks prettiest — and, truth to tell, the decision is partly a matter of what you like to look at.

Or what you *can* look at. I'll never forget the client who asked me very nicely to stop using the default gray background in Excel 1997 charts. Her eyes, which were beginning to age, had trouble distinguishing the black data series in the charts from the gray background of their plot areas.

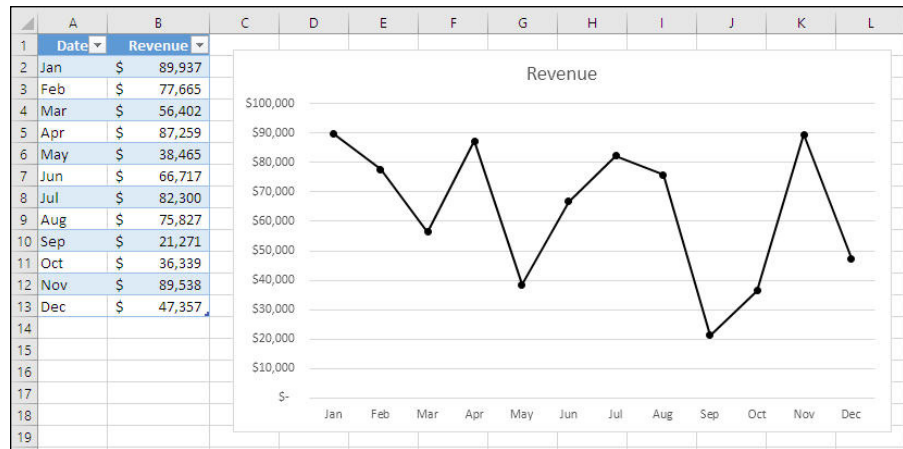
But you need to know about some issues beyond appearance, especially if you're going to use charts to help you forecast sales. And using charts to help forecast sales is a good idea. I hope to convince you of that before this chapter's over.

## Understanding chart types

Most Excel charts have at least two axes. The axes are represented by the chart's left, vertical border and its bottom, horizontal border. Figure 7-1 shows an example.

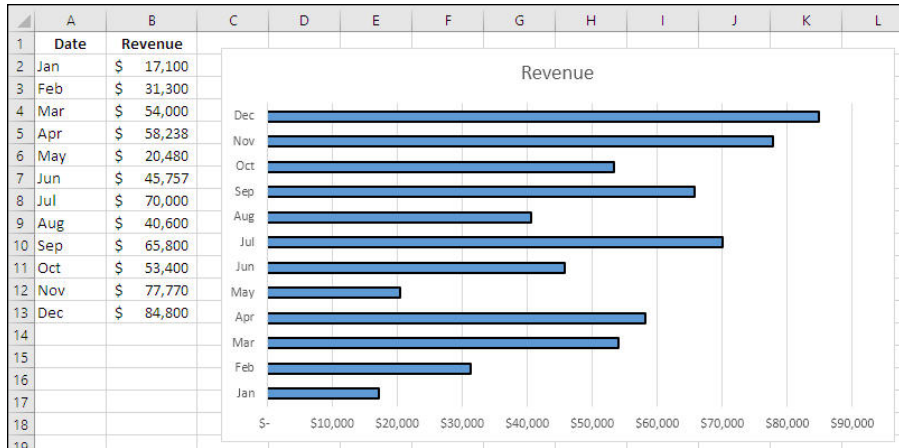
The chart in Figure 7-1 is a *Line chart*. Excel has names for the two different types of axes: the category axis and the value axis. In Figure 7-1, the horizontal axis with the months is called a *category axis*, and the vertical axis with the revenues is called a *value axis*.

**FIGURE 7-1:**  
You can tell the amount of revenue for each month by checking the values on the two axes.



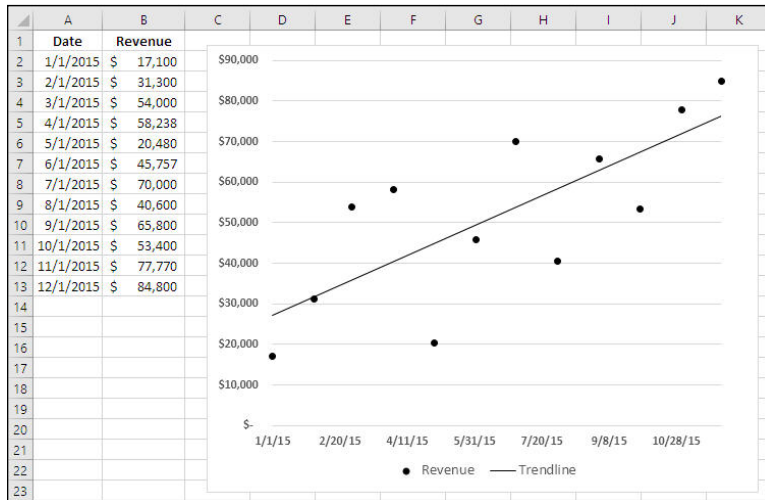
The idea is that different *categories* (here, months) have different numeric *values* (here, revenues). You show the category values on one axis and the numeric values on the other axis.

But the categories aren't always shown on the horizontal border and the values aren't always shown on the vertical. For example, Figure 7-2 shows a *bar chart*. With a bar chart, the categories are on the vertical axis and the values are on the horizontal axis.



**FIGURE 7-2:**  
The categories show up on the vertical axis in a bar chart.

Each of the charts shown in Figure 7-1 and Figure 7-2 has one category axis (date) and one value axis (revenue). But another type of Excel chart has not just one but two value axes: the XY (Scatter) chart. The XY (Scatter) chart (see Figure 7-3) is an important type of chart because it helps you understand what's going on when you forecast by using regression.



**FIGURE 7-3:**  
The XY (Scatter) chart resembles a Line chart, but both the horizontal and the vertical axes are value axes.

When you're forecasting sales, you often pair up dates — usually months or years — with a measure of sales revenue. Revenues are of course numbers, so they belong on a value axis. Dates are also numbers (Excel measures dates by counting the number of days since January 1, 1900). When both fields are numbers, putting them both on value axes is often correct — so, you end up with an XY (Scatter) chart.

Here's one reason that you often want to have two value axes on a chart: You can then use meaningful *trendlines*. A trendline shows the relationship between the two numeric variables. In Figure 7-3, that would be the relationship between the date on the horizontal axis and the revenues on the vertical axis. To get a trendline, follow these steps:

**1. Select the chart.**

If the chart is on its own sheet, select that sheet. If the chart is embedded in a worksheet, click on the embedded chart. A new Design tab for charts appears on the Ribbon.

**2. Go to the Design tab. Click the Add Chart Element icon and choose Trendline from the list box.**

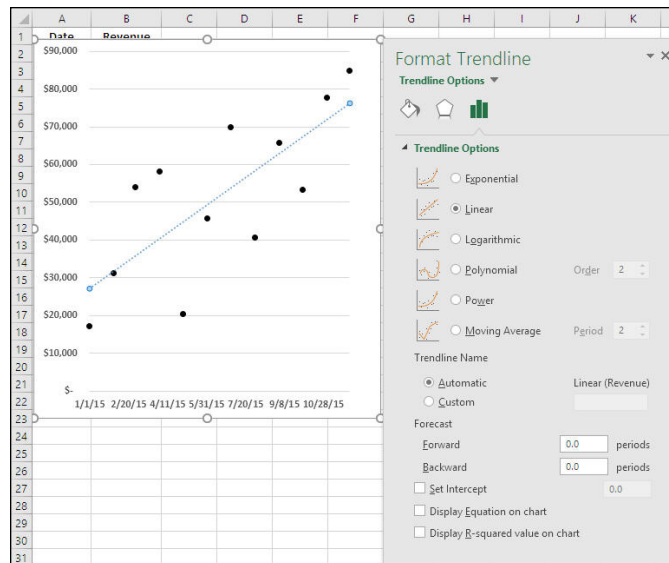
**3. In this case, click Linear in the cascading menu.**

A trendline appears on the chart.



TIP

If you want to add more information along with the trendline, you can choose More Options in the cascading menu that appears in Step 3. But it's quicker to right-click a charted data series and choose Add Trendline from the cascading menu. A Format Trendline pane appears on your screen, as in Figure 7-4.

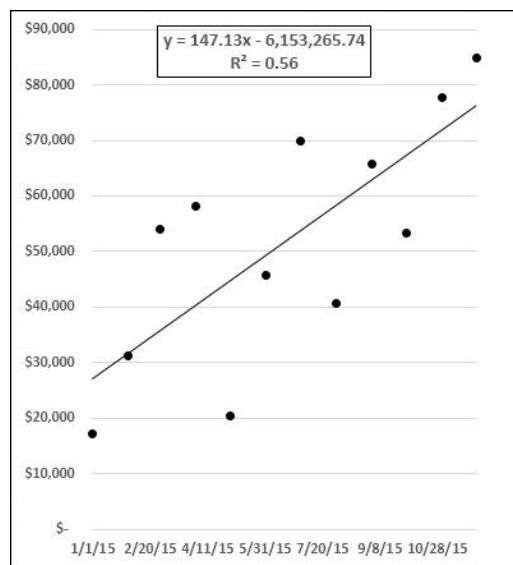


**FIGURE 7-4:** This pane enables you to set all available trendline options.

Then take these steps:

1. Click the Linear option button.
2. Scroll down the Format Trendline pane if necessary, and fill at least two check boxes: Display Equation on Chart and Display R-Squared Value on Chart.
3. Click OK.

Figure 7-5 shows the result, based on the original chart from Figure 7-3.



**FIGURE 7-5:** By showing the equation on the chart, you can preview what the LINEST function would show on the worksheet.

Three new sources of information in Figure 7-5 come from adding the trendline:

- » **The R-squared value:** This tells you how closely the date is related to the revenues. The farther that the R-squared value is from zero, the stronger the relationship. Zero means no relationship — and you may want to try a forecast with one of the other approaches that this book discusses: moving averages or exponential smoothing. Plus 1.0 is a perfect relationship, and if you get it, you can name your own ticket. Sorry, but that never happens and it's not about to.
- » **The equation:** This tells you how to forecast the next revenue value. Plug the next date value into the equation in place of the equation's x value, and the result is your forecast. The better the R-squared value — the farther it is away

from zero and therefore the closer it is to 1.0 — the more confidence you can have in that forecast value.

» **The trendline itself:** This shows you what the equation forecasts in revenues for any given date value. And, of course, it shows you whether your revenues are increasing, decreasing, or flat over time. As the figure shows, I'm assuming your revenues are increasing. That trendline's heading up: You've got a good product, good sales reps, or a good ad firm.



TIP

The chart trendline is a handy way to find the relationship between sales and another variable such as date, as in Figure 7-5, or some other predictor such as advertising budget. But if you're going to use the equation to make a forecast, you should also use the LINEST function on the worksheet. That will give you the equation in worksheet cells, which makes it easier to create the forecast. LINEST also provides you with useful information about how strongly you can rely on the equation as more and more data comes in. Chapter 16 discusses LINEST and regression in more depth.



TIP

You can get a trendline on a chart that's not an XY (Scatter) chart, such as a Line chart. It might well look like the trendline on an XY (Scatter) chart. But if you look at the equation, it will probably be very different. The reason is that in a Line chart, Excel calculates the equation not by relating the revenues to the actual date values, but to the category number (1, 2, 3, and so on). That's why I recommend that you use an XY (Scatter) chart if you're going to use it to have a look at the trendline, equation, and R-squared to evaluate the relationship. Later, if you want to put the equation on the worksheet by using LINEST, it will match the one in your chart.

## Creating the chart from your table

Excel makes it easy to create a chart from a table. Here's how you can create the chart in Figure 7-5 from the two-column table in cells A1:B13:

1. **Click any cell in the table — A6, for example.**
2. **Go to the Ribbon's Insert tab and click the icon for the XY (Scatter) chart.**
3. **Click the default subtype, which has no lines connecting the dots.**

An XY (Scatter) chart appears, embedded in the active worksheet.

So it's just three steps to get from a table to an embedded chart. Because it's so easy to produce an XY (Scatter) chart, there's no excuse not to — particularly



when a chart is such a great way to identify problems that might lead you astray, such as a couple of extreme outliers or perhaps a U-shaped rather than a straight-line relationship.



TIP

It's also easy to set other chart options when you've created the basic chart — options such as its location, whether the plot area has gridlines, whether it has a legend, and so on. With the chart selected, go to the Ribbon's Design tab and then click the Add Chart Element icon in the Chart Layouts group. From there, it's easy to call for more elements such as axis titles, error bars, and trendlines. Or, select an element in the chart such as its legend. Then go to the Ribbon's Format tab and click Format Selection in the Current Selection group to set attributes such as color, size, pattern, and so on.



TIP

No matter what type of chart — Line, XY (Scatter), Column, or something else — you select, if one of your table's fields is a date field, you'll get each individual date from that table on the chart. In Figures 7-1 through 7-5, I've been careful to show each date as a separate month. If your data source provides sales dates on a daily basis, charting that table might result in a chart axis that shows daily sales, and the dates are so jammed together that they're almost impossible to read. One way to manage that is to format the date axis with a larger major unit (7, perhaps, rather than 1). Chapter 8 shows you another way to group your records to get summaries by week, month, quarter, or year. After they're grouped, you'll be able to create a chart that's easier to interpret.

## Refining charts

I go on and on about charts here, and I write even more in other chapters in this book, because I've always found representing my baseline data visually very important. Looking at how the baseline falls out over time is important to how you analyze it. The chart may very well convince you to use a completely different approach to creating your forecast — such as smoothing or moving averages rather than regression.

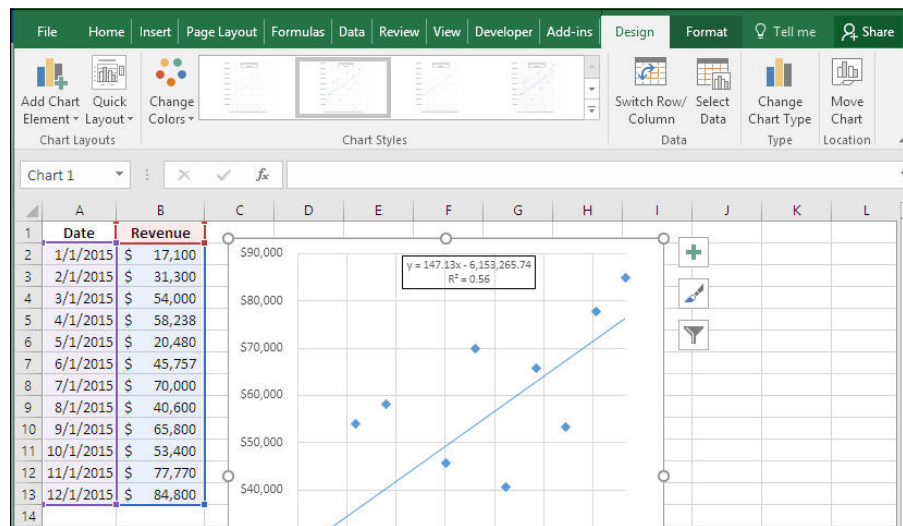
After you've created a chart, you can revise it in different ways to make it more informative. Here are some of them.

## Using the Chart menu

Start off in one of two ways. If you left the chart embedded in a worksheet, as shown in Figure 7-5, click on the chart to activate it. If you put the chart in its own sheet, all by itself, select that sheet. You'll get two Ribbon tabs that you don't see

unless a chart is active: **Design and Format**. You can use the **Design** menu (see Figure 7-6) to control these chart elements:

- » **Add Chart Element:** Include elements such as titles, data labels, legends, and gridlines.
- » **Chart Styles:** Choose from among several different predefined styles (such as colors, fonts, and so on) that alter the look of your chart.
- » **Switch Row/Column:** I doubt that you'll have much use for this tool. One fundamental aspect of Excel tables is that different records occupy different rows, and different variables occupy different columns. If you switch those roles with this tool, you're telling the chart to behave as though the table has different records in different *columns* and each row has a different variable.
- » Click the **Select Data** icon to change the worksheet location of the charted data. You can add a data series to the chart or remove one. And you can extend or reduce the length of a data series that's already in the chart.
- » **Change Chart Type:** Switch from, say, an XY (Scatter) chart to a Line chart, from a Column chart to a Bar chart, and so on.
- » **Move Chart:** Store the chart in its own chart sheet, or as an embedded object in a different worksheet.

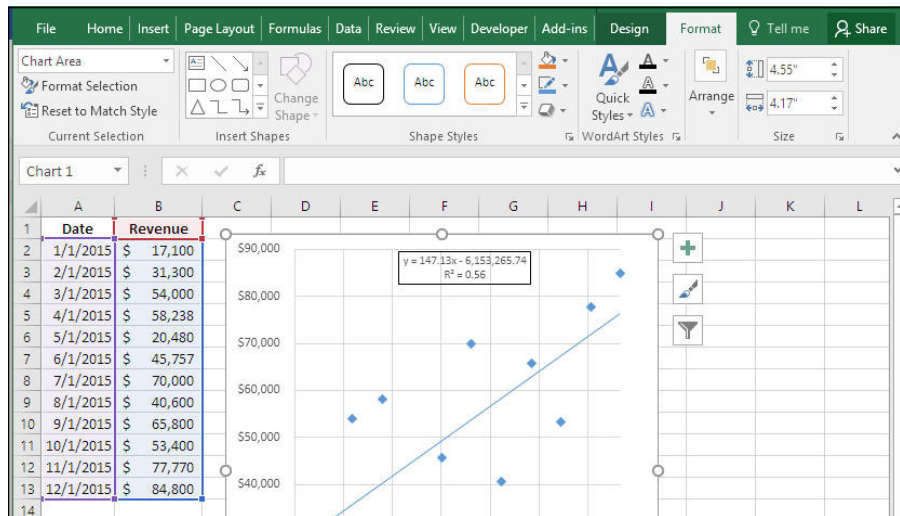


**FIGURE 7-6:** The Chart Design tab puts you in a position to change much of what your chart shows.

As you'd expect, the chart's Format tab, shown in Figure 7-7, gives you tools to control *how* to display what the Design tab enables you to show. Use the tools in the Format tab to

- Format the currently selected chart element. Use this tool to display a pane that contains the formatting options for the selected chart element, such as a chart title or a trendline. Using this tool is not quite as fast as right-clicking the chart element and choosing, say, Format Chart Title or Format Trendline from the shortcut menu. But I've found that it's often easier to select a chart element from the drop-down box at the top of the Current Selection group than to try right-clicking it on the chart. (This is especially true when you're trying to format an element in a particularly complex chart.)
- Insert shapes such as callouts and text boxes into the chart.
- Apply predefined styles to existing shapes, including the colors of the shape's fill and outline.
- Use Word Art Styles to apply predefined styles to selected text objects.
- Align objects in the chart to the chart area, group them together (or ungroup them) to make it easier to move them as one, or rotate them around a focal point.

**FIGURE 7-7:**  
Use the tools in the chart's Format tab to adjust the appearance of individual elements.



## Formatting the axes

You can control more about your chart's appearance by focusing on its axes.

### 1. Right-click one of the axes and choose Format Axis from the shortcut menu.

The Format Axis pane appears. From that pane you can set a large number of options pertaining to the axis and to the axis's text, including the color and shape of its fill and lines, effects such as shadows and soft edges, the size and properties of different parts of the axis, and various other options such as the minimum and maximum value shown on the axis. These options are nice to have and it won't hurt to know that they exist, but few of them have anything to do with forecasting in general or sales forecasting in particular. One exception is the category axis of a Line chart.

### 2. Set up a Line chart with date values showing on its horizontal, category axis.

Click that axis and choose Format Axis from the shortcut menu.

### 3. Click the Axis Options icon near the top of the Format Axis pane and expand its Axis Options item.

Notice that you have three choices as to how the individual values on the axis are treated:

- Automatically select based on data
- Text axis
- Date axis

In most forecasting applications you'll want to choose to treat the axis as a Date axis. If you treat a Date (or any numeric) axis as a Text axis, or if you tell Excel to choose and it chooses poorly, any trendline equation that the chart calculates is apt to be wrong, and the distance between labels on the axis won't necessarily reflect the elapsed time between them.

This is one reason that I generally prefer XY (Scatter) charts to Line charts in forecasting situations. More often than not, both the predictor variable and the predicted variable are numeric measurements (Excel treats dates as numbers), and XY charts automatically have two numeric axes. So it's more difficult to go wrong.

## Using the Data Analysis Add-in with Tables

The Data Analysis add-in (known in earlier versions of Excel as the Analysis Tool-Pak or ATP) helps you do statistical analyses of all sorts — and sales forecasting is definitely one sort of statistical analysis.

An add-in contains Visual Basic code: a program, often written in a version of BASIC, that Excel can run. It's password protected, locked up so that you don't get to look at the code itself. That's okay — you probably wouldn't want to see it any more than you'd want to watch legislators making sausage.



REMEMBER

You need to install the Data Analysis add-in on your computer from your Office installation CD or your download source. You'll find it among the Add-Ins under Excel if you do a custom installation; normally the add-in will be installed automatically if you do a complete installation.

But installing the Data Analysis add-in on your computer doesn't mean it's installed in Excel. If you don't see the words Data Analysis in the Analyze group of the Ribbon's Data tab, then so far you might have just installed it on your hard drive. As with all add-ins, you need to bring it to Excel's attention. To do so, follow these steps:

- 1. In Excel, click the File tab.**
- 2. Choose Options from the navbar at the left of the Excel window.**
- 3. Choose Add-Ins from the navbar at the left of the Excel Options window. Click OK.**
- 4. Make sure that the Manage drop-down near the bottom of the Excel Options window contains *Excel Add-ins*. Click Go.**
- 5. The Add-ins dialog box appears. Make sure that the check box next to Analysis ToolPak (*sic*) is checked, and click OK.**

Now you'll find a new item, Data Analysis, in the Ribbon's Data tab, in the Analyze group. Click that item to get at the add-in's tools.

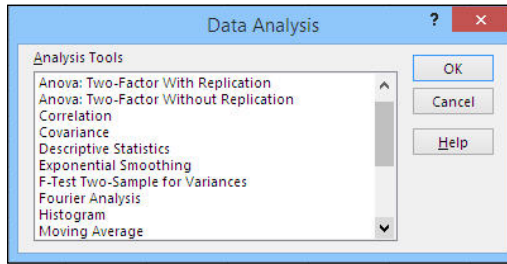
With the Data Analysis add-in installed both on your computer and in Excel, you'll find 19 analysis tools. Suppose you want to wield the Moving Average tool on your table. Do this:

- 1. Click the Ribbon's Data tab and click Data Analysis in the Analyze group.**  
The Data Analysis dialog box, shown in Figure 7-8, appears.
- 2. Scroll down the Data Analysis list box and click Moving Average, and then click OK.**

The Moving Average dialog box, shown in Figure 7-9, appears.

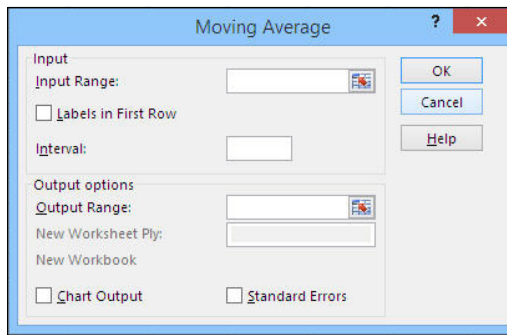
**FIGURE 7-8:**

Not all the analysis tools are useful for doing forecasting. The three best are Exponential Smoothing, Moving Average, and Regression, but you might want to use others for other purposes.



**FIGURE 7-9:**

Using a Data Analysis tool is easiest if the worksheet that contains the table is active — but this isn't required.



3. Click in the **Input Range** field and, using your mouse pointer, drag through the Revenues part of your table.
4. I recommend that you include the column label in the Input Range. If you do so, select the **Labels in First Row** check box.
5. Click in the **Interval** field and enter the number of months (or whatever date period your table uses) you want to base your moving average on.

For example, to base your moving average on a three-month interval, enter 3. If your table measures dates in weeks and you want to base the analysis on a two-week interval, enter 2.

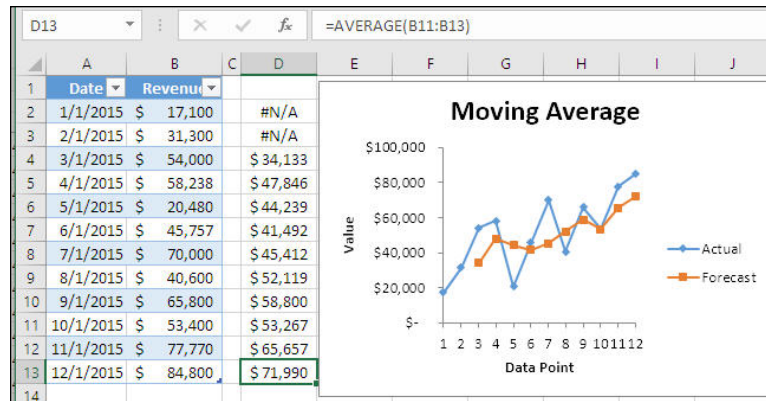
6. Click in the **Output Range** field, and then click the worksheet cell where you want the results to start showing up.
7. If you want to see a chart of the moving average, select the **Chart Output** check box.

Don't make me beg here. Select the check box.

8. Click **OK**.

You'll see the results shown in Figure 7-10.

**FIGURE 7-10:**  
This moving average is based on three intervals — that is, each average consists of three months.



Notice how the moving average smooths out the individual observations. This tends to suppress the *noise* (the random variation in each of your table’s records) and to emphasize the *signal* (the main direction of the baseline).

Also notice how much easier it is to see what’s going on when you look at the chart than when you just look at the table. The lesson: Chart your baselines.

There’s more, much more, to moving averages, and you can find it in Chapters 13 and 14. But now you know the basics of getting and charting a moving average from a table.

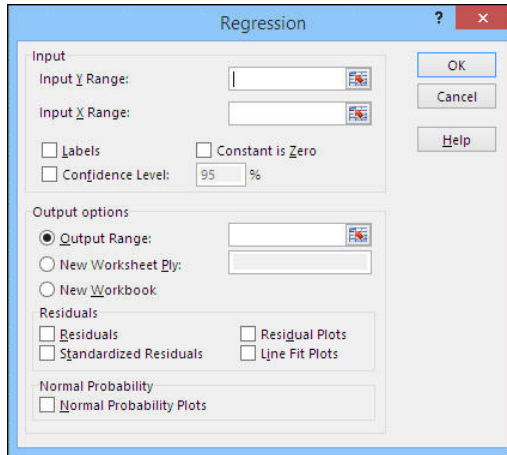
## Avoiding the Data Analysis Add-in’s Traps

For years, some of the Data Analysis add-in’s tools (for example, the Regression tool) have confused the input range with the output range. As you find in Chapter 16, if you’re going to forecast by means of regression you need at least two variables: a predictor variable (such as date or advertising dollars) and a predicted variable (in this context, something such as sales revenues or unit sales).

The Regression tool refers to the predictor variable’s values as the Input X Range, and the predicted variable’s values as the Input Y Range. See Figure 7-11.

Now, suppose you do this:

1. Go to the Ribbon’s Data tab and click Data Analysis in the Analyze group.
2. Locate and click on the Regression tool in the list box and then click OK.



**FIGURE 7-11:**  
Be aware of which reference edit box has the focus.

3. Click in the Input Y Range field, and then drag through something such as your sales revenue values on the worksheet.
4. Click in the Input X Range box, and then drag through something such as the date values on the worksheet.

Notice that the default option for the Output Options is New Worksheet Ply.

If you now override the default option and select the Output Range option button (which lets you put the Regression output on the same sheet with your table), the focus snaps back to Input Y Range. If you then click in some worksheet cell to select it as the output location, that cell becomes the Input Y Range. Because you normally want to use an empty range for the output, you certainly won't select a cell with input values in it. So you choose an empty cell, and because of the change in focus, that cell becomes the Input Y Range.

In other words, the Regression tool is trying to get you to choose a range, or cell, without any data in it to supply your Input Y Range — that is, the values of the variable that's to be predicted.

If you're not aware of what's going on, this can cost you time and unnecessary skull sweat. Unfortunately, there's no good solution — remember, you can't open the code that drives the Data Analysis add-in's tools — other than to be aware that it happens, and to know you have to select the Output Range option button and then its associated edit box again to reset the focus where you want it.





WARNING

Several tools in the Data Analysis add-in have this problem. Take care when you're identifying an output range for one of these tools. If the problem occurs, usually no great harm is done. But it's really annoying after the third or fourth instance.

The other main problem with the Data Analysis add-in is that its output is often static. The Regression tool, for example, puts calculated values in cells rather than formulas that can recalculate when the inputs change. If you get new or changed input values, you'll have to rerun the tool to get the revised results.

Other tools, such as Moving Average and Exponential Smoothing, report their results as formulas, so they'll recalculate if you change the inputs. If you have new values for these tools to use (for example, your input range changes from A1:A20 to A1:A25), you'll need to reset the input range address; but if you're just revising an earlier value, the formulas will recalculate and the charts will redraw without any extra effort on your part.





# **Making a Basic Forecast**

## IN THIS PART . . .

Excel has some powerful tools to help you make your forecasts. One of them is the PivotTable, which organizes your sales history for you and makes creating your baseline a snap. Another is the PivotChart — it's hard to overemphasize how important a chart is to understand what's going on with your sales. If you can visualize it, you're a lot closer to understanding it. Part 3 also introduces you to the Data Analysis add-in (which has tools that help you make basic forecasts from one variable) and to regression analysis (a way to forecast sales using some other variable).

## IN THIS CHAPTER

Getting used to the idea of pivot tables

Putting a pivot table on your worksheet

Grouping similar records together

Staying out of pivot table trouble

## Chapter 8

# Summarizing Sales Data with Pivot Tables

If you're going to create good forecasts, you frequently need to match up some amount of revenue with the period that the revenue came in. Or, if you need to deal with the number of units sold rather than revenue, you need to pair a count of units with the period when they were sold. Excel's pivot tables give you an excellent method of doing that, if you set your data up right.

When you have your pivot table, it often makes sense to look at your revenues from different viewpoints — usually in terms of months or quarters or years, but often by means of other variables such as ranges of commission percentages. Pivot tables refer to this as *grouping*. In this chapter, I cover building pivot tables and grouping dates.

# Understanding Pivot Tables

A pivot table summarizes one variable — typically, one that can be counted or summed, such as units sold or revenue — in terms of another variable, typically one that comes in categories. Pivot tables are the most powerful method in Excel of summarizing any kind of data, including sales information. That makes pivot tables a useful way to prepare a baseline for your forecast.

A pivot table is often based on an Excel table. (If you're not yet a table maven, check out Chapter 6.) Excel tables put different variables, or *fields*, in different columns and different records in different rows. So if you have daily sales revenues in an Excel table, one column would have the date field (like March 10, 2016) and another column, probably an adjacent one, would have the revenue field (like \$5,841). One row would show the revenue for March 10, another would show the revenue for April 4, and so on. Different fields are in different columns, and different records are in different rows.



REMEMBER

*Tables and pivot tables* are two different things. In Excel, a table is just a range of cells that contains data, tricked out with a few special characteristics like a Totals row and sorting and filtering tools. A pivot table is a much more powerful device that can summarize raw data from a table into, for example, a breakdown of sales results according to product line and month sold. Pivot tables frequently use tables as the source of their data.

You can create a pivot table that helps you summarize the information in a table. Pivot tables always have a *value* field. If you have a table that contains revenues by month, you would probably treat the table's revenue field as the pivot table's value field. With revenue in the value field, you're prepared to total up (or otherwise summarize) revenue by date, territory, product line, or any other grouping variable in the original table. (The term *value field* is somewhat misleading. Any field in a pivot table contains values. But Excel reserves *value field* for a field, such as revenue, that's analyzed by some other, often categorical variable such as product line or sales region.)

Pivot tables also usually have *row* fields. If you wanted to analyze revenue by date, you'd probably decide to use the table's date field as the pivot table's row field. That means that different dates (such as different months) show up in different rows of the pivot table, right next to the revenue for each different date.

With that setup, you can *group* the date field so that each row in the pivot table has a broader span than just one day — say, a month, a quarter, or a year. Figure 8-1 gives an example of totaling revenue by month.

	A	B	C	D	E
1	Sales Date	Sales Revenue			
2	1/1/2016	\$ 5,860			
3	1/2/2016	\$ 5,041		Sum of Sales Revenue	
4	1/3/2016	\$ 4,036		Months	Total
5	1/4/2016	\$ 5,543		Jan	\$150,045
6	1/5/2016	\$ 5,429		Feb	\$123,507
7	1/6/2016	\$ 4,362		Mar	\$139,794
8	1/7/2016	\$ 4,993		Apr	\$135,436
9	1/8/2016	\$ 3,767		May	\$138,606
10	1/9/2016	\$ 5,795		Jun	\$139,103
11	1/10/2016	\$ 4,947		Jul	\$140,048
12	1/11/2016	\$ 4,461		Aug	\$140,032
13	1/12/2016	\$ 3,637		Sep	\$131,677
14	1/13/2016	\$ 4,812		Oct	\$141,306
15	1/14/2016	\$ 3,902		Nov	\$138,164
16	1/15/2016	\$ 5,259		Dec	\$137,530
17	1/16/2016	\$ 5,461		Grand Total	\$1,655,248
18	1/17/2016	\$ 5,325			
19	1/18/2016	\$ 4,221			
20	1/19/2016	\$ 5,807			
21	1/20/2016	\$ 5,307			

**FIGURE 8-1:**  
The table in columns A and B gives daily sales revenues. The pivot table starting in column D summarizes the revenues by month.

Pivot tables are handy in various ways. If you want to chart your sales by week, month, quarter, or year, you can start by using a pivot table to summarize them. If you just want a tabulation, the pivot table gives it to you. And if you want to use Excel’s Data Analysis add-in to create a forecast, the pivot table can give you just the layout you’ll need.

In the following sections, you see how to use pivot tables to support your forecasts, by turning basic sales data into a baseline for your forecast, and getting summaries of revenue data and counts of units sold.

## Making baselines out of sales data

Most companies that are in the business of selling products and services record their sales on a daily basis, whether they’re recording revenues or the number of units that were sold. You can usually tell from their corporate accounting system how many dollars they brought in on May 4 and on October 12, or how many widgets they sold on February 8 and on August 25.

The accounting system usually breaks out individual sales. So, if the company made ten sales on June 3, you’re going to see a different record for each of the ten sales. Seeing those sales one by one is great if you’re an accountant, or if you have some other reason to need information about individual sales, or if you’re having trouble sleeping. But if you’re forecasting, individual records are a nuisance.

You need a way of summarizing all those individual records into a baseline for forecasting. Consider the following ideas while deciding the best way to summarize your sales data:

- » **You don't need individual sales records.** If your company made three sales on January 5 — one for \$2,500, another for \$8,650, and another for \$4,765 — an important fact you want to know is that, on January 5, you brought in \$15,915.
- » **You don't forecast sales on a daily basis.** If your company is like most, you need a bigger picture. To plan your inventory levels, decide how many salespeople your company needs, and figure out what you can expect in revenue and what your company's tax liability will be, you need a longer time period such as a month or a quarter for your forecast.
- » **You do need to match the length of your time period with your reasons for forecasting.** Typical time periods are a month, a quarter, or a year, depending on why you're forecasting. For purchasing materials, you may want to forecast your sales for next month. For estimating earnings, you may want to forecast your sales for the next quarter. For hiring decisions, you may want to forecast your sales for the next year.

The point — and I do have one — is that if you're going to forecast sales for next month you need to organize your baseline in months: how much you sold in January, in February, in March, and so on. If you're going to forecast sales for next quarter, then that's how you need to organize your baseline: how much you sold in Q1, in Q2, in Q3, and so on.



REMEMBER

You need a much longer baseline than just three periods to make a forecast that won't embarrass you.

Excel's pivot tables are ideal for helping you total up your sales data to establish a baseline for forecasting. You feed your raw sales data into Excel, where you can build pivot tables in two primary ways:

- » **From an Excel table:** Suppose your Accounting or IT department can send you sales data in a soft copy format, like a .csv (comma-separated values) file. You can paste that data into an Excel workbook as a list, convert that list to a table, and base a pivot table on it.
- » **From (what Excel calls) external data:** In other words, the underlying data, the individual sales figures, aren't stored in an Excel worksheet. They're kept in a separate database or a text file or even another Excel workbook.





TIP

Building your pivot tables on external data can be handy because the sales data are usually updated routinely in the external data source (in practice, this is often a true relational database, as distinct from a flat file like a standard Excel list or table). Then when you want to update your forecast, you don't have to get and paste new data into your workbook. The pivot table can update itself automatically from the external data source.

In the following sections, I show you some of the uses for row fields, column fields, and filter fields in pivot tables.

## Using row fields

Row fields are important in pivot tables, because you can use them to organize your data as a summarized table. In forecasting situations, this means that each row in the pivot table represents each time period in your baseline.

For example, suppose you want to forecast sales for the next quarter. You set up your pivot table so that each of the quarterly sales totals for the previous, say, ten years shows up in a different row — even though each record in your source data represents an individual sale on a particular day.

To total according to quarters, months, or some other time span, you need to *group* the date field. Turn to the “Grouping Records” section, later in this chapter, for more information.

With your pivot table set up that way, getting the Data Analysis add-in, or an XY (Scatter) chart, to create a forecast is easy. And the time period will be right. Summed up into quarterly totals, you can get the next quarter's estimate. Summed up into monthly totals, you can get the best estimate for the next month. You can't do that if the data is still organized by individual sale or by specific day.

The way to start is to put the dates that the sales were made into a pivot table's row field. If the company's accounting system, or sales database, already summarizes sales into the time period you want to use, so much the better. Then you don't need to do any grouping of individual dates into months, quarters, or years. If the company's accounting system, or sales database, doesn't already summarize sales into the time period you want to use, you can do that in a snap (see “Grouping Records,” later in this chapter).

## Using column fields

In setting up your sales data to do forecasting, you're going to have a lot more use for row fields than for column fields, because Excel handles tables so well.

A column field puts different records into different columns — but an Excel table asks you to put different records into different rows. Suppose you have a table that shows the date and the revenue for each date. You want your pivot table to summarize your sales data by month. If you put the date field into the Column area, you'll wind up with a different month in each column.

That's inconvenient, especially if you want to use the Data Analysis add-in to create a forecast. (It can also be mildly inconvenient if you want to chart your revenues over time.)

In sales forecasting, column fields really come into their own when you have several product lines or territories that you want to analyze along with sales dates. Pivot tables can have row fields *and* column fields (and, by the way, filter fields, which in some prior versions of Excel were called *page fields*). The combination of a row field with a column field is ideal for forecasting sales of different product lines.

Your table could have three fields: date of sale, product line, and sales revenue. Then your pivot table could use the fields like this (see Figure 8-2):

- » **Date of sale as a row field:** Each row in the pivot table corresponds to a different total of sales during a particular month.
- » **Product line as a column field:** Each column in the pivot table corresponds to a different product line.
- » **Revenue as a value field:** Each cell in the pivot table sums the revenue for a particular product line on a particular date.

Now the pivot table resembles a standard Excel table. Its structure makes it easy to get a forecast for each product line, whether you're using the Data Analysis add-in to get the forecast or using a trendline in a chart.

## Getting subsets with filter fields

A filter field is fundamentally different from a column, row, or value field. A value field, as I mention in "Understanding Pivot Tables," is the field that you're summing, or averaging, or getting mins and maxs from. Column and row fields give you a way to see something such as the total revenues for Fords versus Chevys, or the units sold for Mixmasters versus Toastmasters.

You use a filter field, instead, to get a subset of data. Suppose your sales data looks like that in Figure 8-3.

	A	B	C	D	E	F	G	H	I	J	K
1	Sales	Product	Sales								
2	Date	Revenue									
3	1/1/2016	24172	\$ 5,860								
4	1/1/2016	B000655-02	\$ 5,041	Sum of Sales							
5	1/1/2016	B000655-03	\$ 4,036	Revenue	Product						
6	1/1/2016	BCD 1554	\$ 5,543	Months	24172	B000655-02	B000655-03	BCD 1554	GNPD-032	Grand Total	
7	1/1/2016	GNPD-032	\$ 5,429	Jan	\$140,259	\$137,765	\$135,345	\$149,008	\$138,580	\$700,957	
8	1/2/2016	24172	\$ 4,362	Feb	\$133,584	\$127,664	\$125,762	\$129,135	\$141,320	\$657,465	
9	1/2/2016	GNPD-032	\$ 4,993	Mar	\$139,281	\$134,659	\$145,851	\$149,119	\$147,298	\$716,208	
10	1/2/2016	BCD 1554	\$ 3,767	Apr	\$137,487	\$137,480	\$136,997	\$134,553	\$125,746	\$672,263	
11	1/2/2016	B000655-03	\$ 5,795	May	\$138,050	\$140,215	\$140,026	\$151,963	\$147,918	\$718,172	
12	1/2/2016	B000655-02	\$ 4,947	Jun	\$133,909	\$136,207	\$133,139	\$134,685	\$128,962	\$666,902	
13	1/3/2016	24172	\$ 4,461	Jul	\$132,710	\$143,616	\$138,890	\$142,125	\$139,203	\$696,544	
14	1/3/2016	B000655-02	\$ 3,637	Aug	\$134,680	\$146,564	\$138,539	\$141,324	\$136,972	\$698,079	
15	1/3/2016	B000655-03	\$ 4,812	Sep	\$137,891	\$128,977	\$133,911	\$132,556	\$142,020	\$675,355	
16	1/3/2016	BCD 1554	\$ 3,902	Oct	\$135,299	\$135,683	\$144,694	\$144,794	\$145,938	\$706,408	
17	1/3/2016	GNPD-032	\$ 5,259	Nov	\$125,402	\$133,347	\$138,427	\$134,535	\$139,102	\$670,813	
18	1/4/2016	24172	\$ 5,461	Dec	\$135,778	\$132,743	\$136,479	\$141,123	\$137,783	\$683,906	
19	1/4/2016	B000655-02	\$ 5,325	Grand Total	\$1,624,330	\$1,634,920	\$1,648,060	\$1,684,920	\$1,670,842	\$8,263,072	
20	1/4/2016	B000655-03	\$ 4,221								
21	1/4/2016	BCD 1554	\$ 5,807								
22	1/4/2016	GNPD-032	\$ 5,307								
23	1/5/2016	B000655-02	\$ 4,254								

**FIGURE 8-2:** This pivot table resembles a standard table, and you can use the Data Analysis add-in to create forecasts from it.

	A	B	C	D	E	F	G	H	I	J	K
1	Sales	Make	Sales								
2	Date	Revenue		Make (All)							
3	1/1/2016	Ford	\$ 20,355	Sum of Sales							
4	1/1/2016	General Mo	\$ 24,341	Revenue	Months	Total					
5	1/1/2016	Toyota	\$ 21,880	Jan	\$2,153,166						
6	1/2/2016	Ford	\$ 20,765	Feb	\$2,042,694						
7	1/2/2016	General Mo	\$ 25,727	Mar	\$2,177,161						
8	1/3/2016	Toyota	\$ 23,588	Apr	\$2,104,159						
9	1/3/2016	Ford	\$ 21,161	May	\$2,173,506						
10	1/3/2016	General Mo	\$ 25,078	Jun	\$2,073,836						
11	1/3/2016	Toyota	\$ 21,258	Jul	\$2,163,797						
12	1/4/2016	Ford	\$ 21,876	Aug	\$2,180,351						
13	1/4/2016	General Mo	\$ 23,536	Sep	\$2,093,834						
14	1/4/2016	Toyota	\$ 20,925	Oct	\$2,171,971						
15	1/5/2016	Ford	\$ 22,877	Nov	\$2,101,149						
16	1/5/2016	General Mo	\$ 24,053	Dec	\$2,157,767						
17	1/5/2016	Toyota	\$ 23,967	Grand Total	\$25,593,391						
18	1/6/2016	Ford	\$ 22,263								
19	1/6/2016	General Mo	\$ 24,035								
20	1/6/2016	Toyota	\$ 21,144								
21	1/7/2016	Ford	\$ 22,549								
22	1/7/2016	General Mo	\$ 24,870								
23	1/7/2016	Toyota	\$ 20,438								
24	1/8/2016	Ford	\$ 22,388								

**PivotTable Fields**

Choose fields to add to report:

Search

Sales Date  
 Make  
 Sales Revenue  
 Months

Drag fields between areas below:

**Filters:** Make

**Columns:**

**Rows:** Months

**Values:** Sum of Sales Revenue

Defer Layout Update Update

**FIGURE 8-3:** The row field breaks revenues down by month. Here, you can use the filter field to focus on sales of Fords, or GM, or Toyota, or All makes.

Click the drop-down list in the filter field. In Figure 8-3, that's the one at the top of the pivot table labeled Make. You now can choose Ford to see the monthly sales revenue for Fords, or Toyota to see the monthly sales revenue for Toyota, or General Motors for GM's monthly revenue.

The filter field is much like an AutoFilter. Use it when you want to zoom in on just one value — Ford, Toyota, whatever — in a field from the pivot table’s underlying data source. When you want to combine all the values in the filter field, just choose All.

## Totaling up the data

In a pivot table, you always have to have a value field. You can manage with no row field or column field or filter field — it doesn’t make a lot of sense to do so, but technically you can. What you can’t do without is a value field. You always have to have a value field that the pivot table can sum, count, or summarize in some other way.

## Summing revenues

In sales forecasting, you’re in the business of summing revenues and counting units sold. So generally you’re going to want to have your value field sum up revenues if it’s dollars, or count units if it’s the number of items you’ve sold.

Figure 8–4 shows what your worksheet looks like when you’re ready to put a field into the  $\Sigma$  Values area of the Field List.

	A	B	C	D	E	F
2	1/1/2016	Ford	\$ 20,355		Drop Report Filter Fields Here	
3	1/1/2016	General Mo	\$ 24,341			
4	1/1/2016	Toyota	\$ 21,880		Sum of Sales Revenue	Total
5	1/2/2016	Ford	\$ 20,765		Sales Date	
6	1/2/2016	General Mo	\$ 25,727		1-Jan	66576
7	1/2/2016	Toyota	\$ 23,588		2-Jan	70080
8	1/3/2016	Ford	\$ 21,161		3-Jan	67497
9	1/3/2016	General Mo	\$ 25,078		4-Jan	66337
10	1/3/2016	Toyota	\$ 21,258		5-Jan	70897
11	1/4/2016	Ford	\$ 21,876		6-Jan	67442
12	1/4/2016	General Mo	\$ 23,536		7-Jan	67857
13	1/4/2016	Toyota	\$ 20,925		8-Jan	70784
14	1/5/2016	Ford	\$ 22,877		9-Jan	66066
15	1/5/2016	General Mo	\$ 24,053		10-Jan	72821
16	1/5/2016	Toyota	\$ 23,967		11-Jan	68865
17	1/6/2016	Ford	\$ 22,263		12-Jan	67604
18	1/6/2016	General Mo	\$ 24,035		13-Jan	69544
19	1/6/2016	Toyota	\$ 21,144		14-Jan	70685
20	1/7/2016	Ford	\$ 22,549		15-Jan	69198
21	1/7/2016	General Mo	\$ 24,870		16-Jan	67941
22	1/7/2016	Toyota	\$ 20,438		17-Jan	71675
23	1/8/2016	Ford	\$ 23,399		18-Jan	67879
24	1/8/2016	General Mo	\$ 23,914		19-Jan	71417
25	1/8/2016	Toyota	\$ 23,471		20-Jan	70014
26	1/9/2016	Ford	\$ 21,145		21-Jan	67784
27	1/9/2016	General Mo	\$ 24,479		22-Jan	69840
28	1/9/2016	Toyota	\$ 20,442		23-Jan	72228

**FIGURE 8-4:** You’ve completed the pivot table as soon as you put a field in the  $\Sigma$  Values area.

Excel stalks you. It's always watching what you're doing. Suppose your value field is revenues. That's a *numeric* field: It's numbers. When you put a numeric field into the  $\Sigma$  Values area, Excel notices that and it automatically gives you a sum. For each value of the row or column field, you'll get a sum of, for example, your sales revenues.

You'll almost always have a row, column, or filter field in your pivot table. In sales forecasting, you'll usually find yourself using dates as a row field, so January might be in one row, February in the next, and so on. In Figure 8-4, the pivot table's Sales Date field shows the date of each individual sale, so it will have to be grouped to summarize on month.

For each item in the row field, the pivot table's value field will give you the total of all the revenue that goes with that item. Maybe your list shows 50 sales records for February. The pivot table can total the revenue for those 50 records and show the sum of that revenue in one row — which would then be labeled “February.”

## Counting units

But what if you want to summarize units sold rather than revenues? Say you sell cars. Then your table might have a field named Make. You'd put that field into the Row area and into the  $\Sigma$  Values area. Because Make is a text field — with values like Ford, Toyota, and General Motors — the pivot table's  $\Sigma$  Values area defaults not to a sum but to a count. After all, it can't add a Ford to a Chevy. But it can count the number of Fords, Chevys, and GMs that you've sold.

Putting the same field into both the Row area and the  $\Sigma$  Values area may seem a little strange, but that's how it's done. You'll get a different row for each distinct value. And — if you're sure it's a text field — you'll get a count of records for each row. Figure 8-5 shows an example.

	A	B	C	D	E	F
	Sales	Product	Sales			
1	Date		Revenue			
2	1/1/2016	General Motors	\$ 20,355			
3	1/1/2016	Ford	\$ 24,341		Row Labels	Count of Product
4	1/1/2016	Ford	\$ 21,880		Ford	449
5	1/2/2016	General Motors	\$ 20,765		General Motors	276
6	1/2/2016	Toyota	\$ 25,727		Toyota	373
7	1/2/2016	General Motors	\$ 23,588		Grand Total	1098
8	1/3/2016	General Motors	\$ 21,161			
9	1/3/2016	Ford	\$ 25,078			

**FIGURE 8-5:** Pivot tables use Count as the default when the value field has text values.

Sometimes Excel defaults to the wrong summary. This can happen, for example, when you accidentally wind up with a text value in what you expected to be an

entirely numeric field. If Excel finds even one text value in the pivot table's value field, it will default to Count as the summary.

If this happens, you need to do two things:

- » Find the text value in the data source and either change it to a number, or delete it.
- » Right-click a cell in the pivot table's value field and choose Value Field Settings from the shortcut menu. Then click on Sum in the Summarize Value Field By list box.

## Other summaries

When you're forecasting sales, you're generally most interested in total revenues and counts of units sold. But sometimes you want other information, such as the largest single sale made each month, or the smallest number of units sold each quarter.

The pivot table's value field offers you several summaries besides Sum and Count. You can look at a Maximum, a Minimum, and an Average. (There are several other summaries that you're unlikely to have use for in sales forecasting. They have to do mostly with how the individual records in the underlying data set vary around their average value.)

# Building the Pivot Table

Okay, it's time to actually create a pivot table and group the date field. You've got your baseline of data, laid out as a table. What now?

Suppose your table looks like Figure 8-6.

The problem is that you don't need to forecast sales by sales rep. And you don't need to forecast sales by day. In this case, the whole idea is to forecast sales by month. Corporate couldn't care less who made the sale, or on which day. They don't want to know how much George Smith sold on February 24. They want you to tell them how much your sales team will sell in January 2017. For that, you don't need daily sales data and you don't need salesperson data.

You do need to know monthly totals. With that information, you can generate a forecast that's rational and credible. You'll get the monthly totals, but first you need to build the pivot table.

	A	B	C	D
1	Sales	Sales Rep	Sales	
2	Date		Revenue	
2	1/1/2016	Jones	\$ 20,355	
3	1/1/2016	Smith	\$ 24,341	
4	1/1/2016	Williams	\$ 21,880	
5	1/2/2016	Jones	\$ 20,765	
6	1/2/2016	Smith	\$ 25,727	
7	1/2/2016	Williams	\$ 23,588	
8	1/3/2016	Jones	\$ 21,161	
9	1/3/2016	Smith	\$ 25,078	
10	1/3/2016	Williams	\$ 21,258	
11	1/4/2016	Jones	\$ 21,876	
12	1/4/2016	Smith	\$ 23,536	
13	1/4/2016	Williams	\$ 20,925	
14	1/5/2016	Jones	\$ 22,877	
15	1/5/2016	Smith	\$ 24,053	
16	1/5/2016	Williams	\$ 23,967	
17	1/6/2016	Jones	\$ 22,263	
18	1/6/2016	Smith	\$ 24,035	
19	1/6/2016	Williams	\$ 21,144	
20	1/7/2016	Jones	\$ 22,549	
21	1/7/2016	Smith	\$ 24,870	
22	1/7/2016	Williams	\$ 20,438	
23	1/8/2016	Jones	\$ 23,399	

**FIGURE 8-6:**  
The table shows sales by date and by sales rep.



TIP

**1. Click one of the cells in your table.**

When Excel creates the pivot table, it needs to know where to find the table that contains the raw data. If you start by selecting a cell in that table, Excel locates its boundaries for you.

**2. Go to the Ribbon's Insert tab and click the Pivot Table icon in the Tables group.**

The dialog box shown in Figure 8-7 appears.

**3. Select the Existing Worksheet option button if you want the pivot table to appear on the same worksheet as the table.**

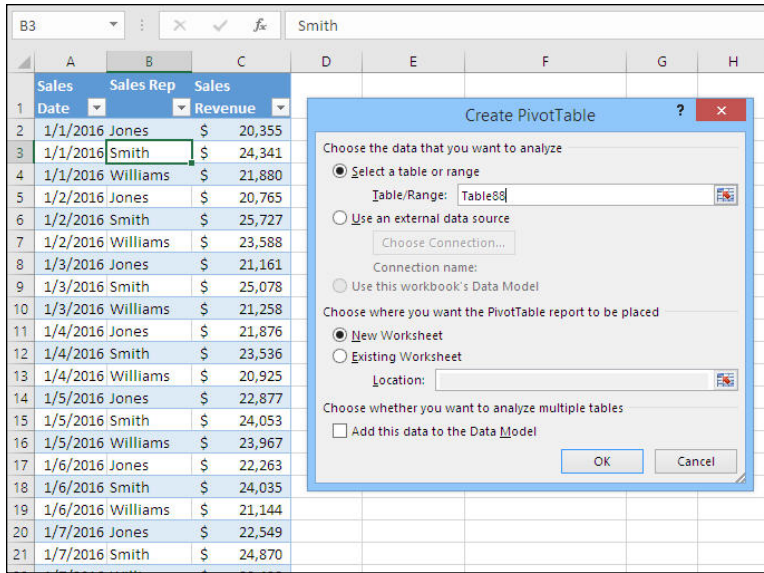
Especially during initial development, I find it's useful to keep the raw data and the pivot table on the same worksheet.

**4. Click in the Location reference edit box and then click a worksheet cell to the right of or below the table. Click OK. Your worksheet will resemble the one shown in Figure 8-8.**

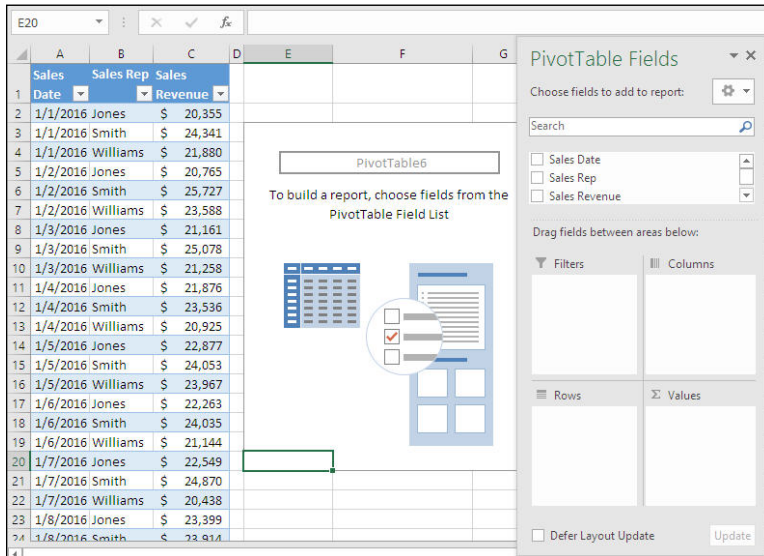
When Excel creates the pivot table, it builds it to the right and down from the cell you select. If you start the pivot table to the left of or above your list, Excel will ask you if you really want to overwrite those cells.



**FIGURE 8-7:** This is where you tell Excel how to find the basic data. In this case, your data is already on the worksheet, in the form of an Excel table.



**FIGURE 8-8:** Finish defining the pivot table with the PivotTable Fields pane.



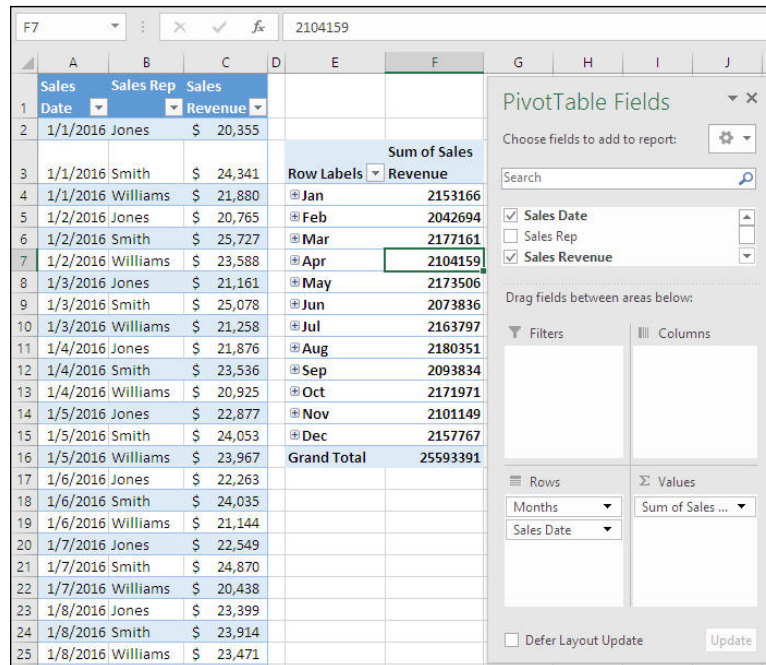
**5. Click and drag the Sales Date field from the Fields section to the Rows area.**

Depending on the version of Excel you're running, Excel might add a Months field along with Sales Date to the Rows area. If so, you'll see the sales month on the worksheet as the new pivot table's row field.



**6. Click and drag the Sales Revenue field from the Fields section to the  $\Sigma$  Values area.**

You now have a pivot table that shows a baseline of sales revenues by sales date. It should look something like the pivot table in Figure 8-9.



**FIGURE 8-9:** You can display individual days by clicking the expand box next to any month name.



**TIP**

Excel might automatically put Month into the pivot table's rows. If you don't want it there, just click Month in the PivotTable Fields pane's Rows area and drag Month back up to the Fields section.

See the next section, "Grouping Records," for information about how to show the sales revenues by month (or by year, or some other date grouping) if you need the pivot table's row field to aggregate the records by a different time period.

In the meantime, it's usually a good idea to format the Values field properly. Here, that's using the Currency format. Here are the steps:

- 1. With the PivotTable Fields pane visible, click on the drop-down arrow next to Sum of Sales Revenue in the  $\Sigma$  Values area.**
- 2. Choose Value Field Settings from the shortcut menu.**

3. Click the **Number Format** button.
4. Choose the **Currency** format.
5. Adjust the number of decimal places.
6. Click **OK** to return to the **Value Field Settings** dialog box, and click **OK** again to return to the worksheet.

If you're using Excel 2016, which provides automatic time grouping, the pivot table should now appear as in Figure 8-10. Otherwise the rows will reflect the periods used in the pivot table's data source. If you still see individual dates, for example, instead of months, refer to the next section for information on arranging individual dates into longer periods of time.

	A	B	C	D	E	F
	Sales	Sales Rep	Sales			
1	Date		Revenue			
2	1/1/2016	Jones	\$ 20,355			
3	1/1/2016	Smith	\$ 24,341			Sum of Sales
4	1/1/2016	Williams	\$ 21,880	⊕ Jan		\$2,153,166
5	1/2/2016	Jones	\$ 20,765	⊕ Feb		\$2,042,694
6	1/2/2016	Smith	\$ 25,727	⊕ Mar		\$2,177,161
7	1/2/2016	Williams	\$ 23,588	⊕ Apr		\$2,104,159
8	1/3/2016	Jones	\$ 21,161	⊕ May		\$2,173,506
9	1/3/2016	Smith	\$ 25,078	⊕ Jun		\$2,073,836
10	1/3/2016	Williams	\$ 21,258	⊕ Jul		\$2,163,797
11	1/4/2016	Jones	\$ 21,876	⊕ Aug		\$2,180,351
12	1/4/2016	Smith	\$ 23,536	⊕ Sep		\$2,093,834
13	1/4/2016	Williams	\$ 20,925	⊕ Oct		\$2,171,971
14	1/5/2016	Jones	\$ 22,877	⊕ Nov		\$2,101,149
15	1/5/2016	Smith	\$ 24,053	⊕ Dec		\$2,157,767
16	1/5/2016	Williams	\$ 23,967	Grand Total		\$25,593,391
17	1/6/2016	Jones	\$ 22,263			
18	1/6/2016	Smith	\$ 24,035			
19	1/6/2016	Williams	\$ 23,144			

**FIGURE 8-10:**  
The Sales Revenue field is now formatted as Currency.

## Grouping Records

You can use pivot tables to summarize data. One of the ways that you fine-tune a summary in a pivot table is to *group* a row field or a column field.

When you group a row field or column field, you combine the values in the field. For example, you might have a field that shows the dates when sales were made. Excel's pivot tables make it possible for you to group individual dates like September 4 and September 6 into weeks, like September 1 through September 7.

Or you can group into months, or quarters, or years. It all depends upon how you want to forecast — by week, by month, by quarter, by year. Keep your time period in mind.

## Knowing when to group records

If you followed the steps in the preceding section, you have a pivot table that summarizes revenue, but it might still do so day by day. The reason is that the table you've based the pivot table on shows revenue by day. Summarizing revenue by month (or week, or quarter, or year) would be much better. Here's how to do that.

Notice in Figure 8-10 that the underlying table has several identical dates. The pivot table combines any identical dates into just one row, adding up their revenue along the way. One of the things that pivot tables do is to combine identical values in a table into the same item in a row, column, or filter field.

But you don't want to forecast by individual dates. That's not the right time period. In this kind of situation, you need to group individual dates that belong to the same month into one row of the pivot table. Then all the revenues for the grouped dates are also totaled.

Of course, you're not restricted to grouping on months. You can use seven-day periods, quarters, years, and others (such as three-day or ten-day periods). If you're concerned about time of day, as you might be with online sales, you can use the hour of the day as a period. Or if your raw data shows the time of day when a sale was made, you can group the individual hours into 8-hour shifts.

## Creating the groups

We'll start with the day as the row field as shown in Figure 8-11.

After you have a pivot table set up, as in Figure 8-11, grouping the individual dates in the row field is simple. Follow these steps:

- 1. Right-click any of the cells in the column of the pivot table that has the dates — for example, cell E7 in Figure 8-11.**
- 2. Choose Group from the shortcut menu.**
- 3. Choose Months from the list box in the Grouping dialog box. See Figure 8-12.**
- 4. Click OK.**

	A	B	C	D	E	F
1	Sales	Product	Sales			
2	Date		Revenue	Row Labels	Sum of Sales Revenue	
3	1/1/2015	General Motors	\$20,355	1/1/2015		\$66,576
4	1/1/2015	Ford	\$24,341	1/2/2015		\$70,080
5	1/2/2015	General Motors	\$20,765	1/3/2015		\$67,497
6	1/2/2015	Toyota	\$25,727	1/4/2015		\$66,337
7	1/2/2015	General Motors	\$23,588	1/5/2015		\$70,897
8	1/3/2015	General Motors	\$21,161	1/6/2015		\$67,442
9	1/3/2015	Ford	\$25,078	1/7/2015		\$67,857
10	1/3/2015	Ford	\$21,258	1/8/2015		\$70,784
11	1/4/2015	Toyota	\$21,876	1/9/2015		\$66,066
12	1/4/2015	Toyota	\$23,536	1/10/2015		\$72,821
13	1/4/2015	General Motors	\$20,925	1/11/2015		\$68,865
14	1/5/2015	General Motors	\$22,877	1/12/2015		\$67,604
15	1/5/2015	Toyota	\$24,053	1/13/2015		\$69,544
16	1/5/2015	General Motors	\$23,967	1/14/2015		\$70,685
17	1/6/2015	Toyota	\$22,263	1/15/2015		\$69,198
18	1/6/2015	Ford	\$24,035	1/16/2015		\$67,941
19	1/6/2015	General Motors	\$21,144	1/17/2015		\$71,675
20	1/7/2015	Ford	\$22,549	1/18/2015		\$67,879
21	1/7/2015	Ford	\$24,870	1/19/2015		\$71,417
22	1/7/2015	General Motors	\$20,438	1/20/2015		\$70,014

**FIGURE 8-11:**  
You want to group the daily sales into months.

	A	B	C	D	E	F	G	H
1	Sales	Product	Sales					
2	Date		Revenue	Row Labels	Sum of Sales Revenue			
3	1/1/2015	General Motors	\$20,355	1/1/2015		\$66,576		
4	1/1/2015	Ford	\$24,341	1/2/2015		\$70,080		
5	1/2/2015	General Motors	\$20,765	1/3/2015		\$67,497		
6	1/2/2015	Toyota	\$25,727	1/4/2015		\$66,337		
7	1/2/2015	General Motors	\$23,588	1/5/2015				
8	1/3/2015	General Motors	\$21,161	1/6/2015				
9	1/3/2015	Ford	\$25,078	1/7/2015				
10	1/3/2015	Ford	\$21,258	1/8/2015				
11	1/4/2015	Toyota	\$21,876	1/9/2015				
12	1/4/2015	Toyota	\$23,536	1/10/2015				
13	1/4/2015	General Motors	\$20,925	1/11/2015				
14	1/5/2015	General Motors	\$22,877	1/12/2015				
15	1/5/2015	Toyota	\$24,053	1/13/2015				
16	1/5/2015	General Motors	\$23,967	1/14/2015				
17	1/6/2015	Toyota	\$22,263	1/15/2015				
18	1/6/2015	Ford	\$24,035	1/16/2015				
19	1/6/2015	General Motors	\$21,144	1/17/2015				
20	1/7/2015	Ford	\$22,549	1/18/2015				
21	1/7/2015	Ford	\$24,870	1/19/2015				
22	1/7/2015	General Motors	\$20,438	1/20/2015				

**Grouping** ? x

Auto

Starting at: 1/1/2015

Ending at: 1/1/2017

By

- Seconds
- Minutes
- Hours
- Days
- Months
- Quarters
- Years

Number of days: 1

OK Cancel

**FIGURE 8-12:**  
You can nest periods by selecting more than one grouping factor.

Because you want to summarize your revenues by month, you can accept the default selection, Months. When you click OK, the pivot table changes and looks like the one shown in Figure 8-13.

You always need to know how you're going to get out. If you want to ungroup the records, do this:

	A	B	C	D	E	F
1	Sales	Product	Sales			
2	Date		Revenue	Row Labels	Sum of Sales Revenue	
3	1/1/2015	General Motors	\$20,355	Jan	\$4,306,332	
4	1/1/2015	Ford	\$24,341	Feb	\$4,013,180	
5	1/2/2015	General Motors	\$21,880	Mar	\$4,354,322	
6	1/2/2015	General Motors	\$20,765	Apr	\$4,208,318	
7	1/2/2015	Toyota	\$25,727	May	\$4,347,012	
8	1/2/2015	General Motors	\$23,588	Jun	\$4,147,662	
9	1/3/2015	General Motors	\$21,161	Jul	\$4,327,593	
10	1/3/2015	Ford	\$25,078	Aug	\$4,360,702	
11	1/3/2015	Ford	\$21,258	Sep	\$4,187,668	
12	1/4/2015	Toyota	\$21,876	Oct	\$4,343,942	
13	1/4/2015	Toyota	\$23,536	Nov	\$4,202,288	
14	1/4/2015	General Motors	\$20,925	Dec	\$4,315,534	
15	1/5/2015	General Motors	\$22,877	Grand Total	\$51,114,553	
16	1/5/2015	Toyota	\$24,053			

**FIGURE 8-13:**  
Now you can see the total revenue for each month.

### 1. Right-click in a cell with grouped records.

In Figure 8-13, that's a cell in the Row area.

### 2. Choose Ungroup from the shortcut menu.

Now you have the row field back to its original state.



TIP

If you want to change a grouping level from, say, Months to Quarters, you don't need to ungroup first. Just right-click a grouped cell, deselect Months, and select Quarters. Then click OK.



TIP

In forecasting, grouping in pivot tables tends to occur with date fields. But the capability extends to any numeric field. You could group on discount percentages to view sales in which the price was discounted less than 5%, from 5% to 10%, 10% to 15%, and so on. Or you could group by length of vehicle leases.

## Avoiding Grief in Excel Pivot Tables

You need to watch out for a few things if you're grouping on a date field in an Excel pivot table: You might get an error message that doesn't tell you what the problem is; you want to orient your table correctly; you want to make sure that you've chosen enough grouping levels. The next few sections talk about these issues.

### Don't use blank dates

Have another look at the table that the pivot table in Figure 8-13 summarizes. All the records in the table have dates. Suppose that one (or more) of the records in the table was missing a date value — that the record just had a blank cell rather than a date.

You could create a pivot table from that table. And one of the rows in the pivot table would show (*blank*) rather than a date. In some versions of Excel prior to 2016, as soon as you tried to group that field, Excel would whine at you. You'd see a message box that would say *Cannot group that selection*.

If you see that message, it's almost certain that the problem is a missing value (or even a text value) in the field you're trying to group. Here's how to fix it:

**1. Go back to your table and locate the record with the missing value.**

Maybe you can figure out what the missing value should be.



TIP

If you can't figure out what the missing value should be, you may be better off just removing it from the table.

**2. Fill in the missing data with your best guess.**

**3. With a pivot table cell selected, go to the Ribbon's Analyze tab and choose Refresh in the Data group.**

Refreshing the pivot table's data forces it to go back to its source — in this example, the table — and replace the missing value that kept it from grouping.



TIP

Any time the underlying data changes, you need to refresh the data to force the pivot table to reflect the change. It doesn't happen automatically, the way a worksheet formula recalculates automatically.



REMEMBER

The same thing happens if you are trying to group a standard number field that doesn't have dates in it. Regardless of the type of data you're grouping on, and regardless of the Excel version in use, the thing to remember is that Excel can't accurately group a field that has missing values.

## Making multiple groups

What if your table has dates that span more than one year? If your data spans more than one year, and you choose to group just on month, Excel will put March 2015 in with March 2016 and call them both March.

You might decide to group your dates from different years just by their month, because it could be interesting to see how your sales revenues vary by month, regardless of the year. If you sell parkas, for example, you might expect to see your sales spike in the fall and winter — and you might have to go on the dole in the spring and summer. (I show you more about forecasting and seasons in Chapter 18.)

But normally you'll want to look at the monthly (or quarterly) results for 2016 separately from the monthly results for 2015. That's easy to handle. Just follow these steps:

1. Right-click any cell in the date field of the pivot table report and choose **Group** from the shortcut menu.
2. As usual, accept the default **Months** grouping.
3. Click **Years** at the bottom of the list box.  
You might need to scroll down the list box to find Years.
4. Click **OK**.

Now you'll be able to see months within years, as in Figure 8-14.

	A	B	C	D	E	F
1	Sales	Product	Sales			
2	Date	General Motors	\$20,355	Row Labels	Sum of Sales Revenue	
3	1/1/2015	Ford	\$24,341	2015		
4	1/1/2015	Ford	\$21,880	Jan	\$2,153,166	
5	1/2/2015	General Motors	\$20,765	Feb	\$1,970,486	
6	1/2/2015	Toyota	\$25,727	Mar	\$2,177,161	
7	1/2/2015	General Motors	\$23,588	Apr	\$2,104,159	
8	1/3/2015	General Motors	\$21,161	May	\$2,173,506	
9	1/3/2015	Ford	\$25,078	Jun	\$2,073,826	
10	1/3/2015	Ford	\$21,258	Jul	\$2,163,796	
11	1/4/2015	Toyota	\$21,876	Aug	\$2,180,351	
12	1/4/2015	Toyota	\$23,536	Sep	\$2,093,834	
13	1/4/2015	General Motors	\$20,925	Oct	\$2,171,971	
14	1/5/2015	General Motors	\$22,877	Nov	\$2,101,139	
15	1/5/2015	Toyota	\$24,053	Dec	\$2,157,767	
16	1/5/2015	General Motors	\$23,967	2016		
17	1/6/2015	Toyota	\$22,263	Jan	\$2,153,166	
18	1/6/2015	Ford	\$24,035	Feb	\$2,042,694	
19	1/6/2015	General Motors	\$21,144	Mar	\$2,177,161	
20	1/7/2015	Ford	\$22,549	Apr	\$2,104,159	
21	1/7/2015	Ford	\$24,870	May	\$2,173,506	
22	1/7/2015	General Motors	\$20,438	Jun	\$2,073,836	
23	1/8/2015	Ford	\$23,399	Jul	\$2,163,797	

**FIGURE 8-14:**  
The years appear in an outer row field and the months in an inner row field.



TIP

You can deselect any grouping level that's already selected just by clicking it in the Grouping dialog box. For example, Months is the default grouping level for a date field. If you don't want to group Months, just click on it to deselect it, and then group the level you want to use — like Quarters or Years.



TIP

You can't group specifically on Weeks. You need to select Days, and then specify 7 with the Number of Days spinner. Excel enables the spinner as soon as you select Days. (A *spinner* is a type of control in a dialog box. You can see one — although it's dimmed — above the Cancel button in Figure 8-12. You use it to increase or decrease the number of days to group by.)





Using dates and times in Excel

Figuring out what's going on in your baseline

Using pivot charts to make your data look good

Using more than one scale

## Chapter 9

# Charting Your Baseline: It's a Good Idea

**T**o make good forecasts — forecasts that your colleagues think of as useful guides to the future — you need to know as much as you can about the baselines that determine those forecasts. You can run all sorts of highfalutin statistical tests on a baseline, but one of your best sources of information is the lowly chart. For example, if your baseline describes an upside-down U, simple regression statistics or correlations aren't going to tell you that. But charting the baseline often results in the well-known interocular impact effect: It hits you right between the eyes.

If you use a fairly recent version of Excel, you can use a special kind of chart called a pivot chart. Pivot charts are similar to pivot tables in two primary ways:

- » You can easily change the role of a variable — for example, moving it from a chart's horizontal axis to its vertical axis.
- » You can display a data field as any one of a set of summary options: Count, Sum, Average, and so on.

Sometimes you want to chart more than one data series at once, so you can see how two baseline variables behave over time. If those data series have very different scales, such as monthly sales and monthly sales commissions, the differences in the scales can make it hard to see what's happened during the course of the baseline. Putting two value axis scales into your chart can help in this situation.

## Digging into a Baseline

Sometimes you can stare at the numbers in a baseline until they start to swim up at you out of the worksheet and tell you absolutely nothing. But you have to find a way to get a read on what's going on in that baseline. It might be rising, falling, holding steady, or just out for a walk.

At these times — really, at all times — you should think about putting the baseline into a chart. Because of the nature of forecasting, many charts that help you understand the baseline use dates on one chart axis or another. So you may as well start by getting your arms around how Excel keeps track of dates.

## Using date and time data in Excel

I feel as though I should apologize for this section but I'm not going to. It's a little dull, and I've looked in vain for ways to spice it up with some jokes that are in questionable taste. The thing is, you have to have a grasp on this stuff if you're going to understand how dates and times work in Excel — and, therefore, how they work in regression forecasts and in charts.

### How Excel keeps track of dates

Excel assigns a numeric value to each date since January 1, 1900 (January 1, 1904, on Macintosh computers). You can see this for yourself by following these steps:

- 1. Enter 1 in a worksheet cell.**
- 2. Right-click that cell.**
- 3. Select Format Cells from the shortcut menu.**
- 4. Select the Number tab.**
- 5. In the Category list box, choose Date.**
- 6. In the Type list box, choose any date format that shows the month, the day, and the year.**

## 7. Click OK.

Depending on which specific format you selected in Step 5, you'll see something such as 1/1/1900 or January 1, 1900.

Working the other direction, if you enter **1/2/1900** in a worksheet cell and format it as Number (as in Step 5), you'll see the value 2, or 2.00, or however many decimal places you chose when you formatted the value as a number. January 2, 1900, is the second day in the sequence.

So, despite their format on the worksheet, dates in Excel are actually numbers. This fact has implications for using, say, date of sale as a predictor variable in a regression forecast, and for how you view date of sale in an Excel chart.

## How Excel keeps track of time

Excel also assigns a number to time of day. It's a fractional value, ranging from 0.000 for midnight, to 0.999 for 11:59 p.m. And Excel adds that fractional value to the integer date value, so you can specify a particular time of day on a particular date. For example, 17804.72917 represents September 28, 1948, at 5:30 p.m.

You probably won't have much use for fractional time values in sales forecasting, but if you do happen to come across a date/time value with a fractional component, you'll know what it's about.

## Charting dates and times in Excel

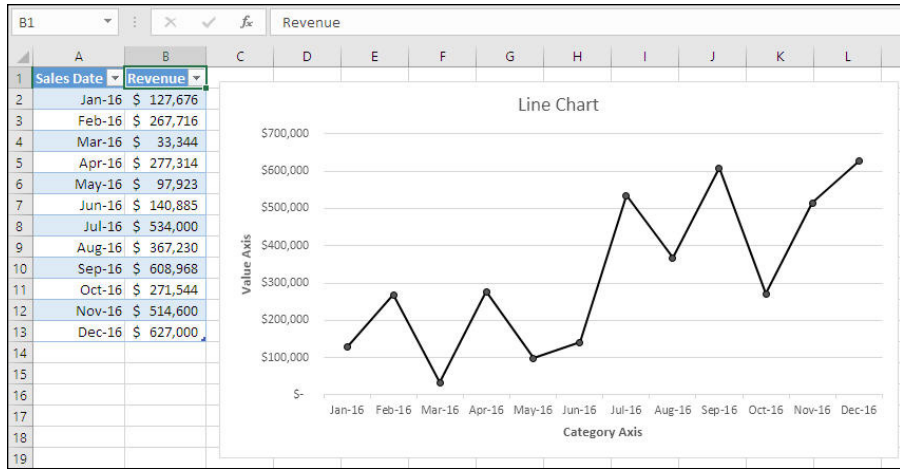
Chapter 7 dips briefly into the topic of using charts in forecasting, just enough to whet your appetite. In this section, I tell you more about the importance of choosing the right chart type when you chart your baseline.

There are at least two axes on most Excel standard chart types (the exceptions are Pie and Doughnut charts). Three-dimensional charts have three axes. The charts that have two axes, as you'd expect, have a vertical axis and a horizontal axis. Figure 9-1 shows the two axes on a Line chart.

A chart axis can display variables of two different types: *value* and *category*. These are Excel's terms, and unfortunately the term *value* is misleading.

A category axis is meant to display labels, like London, Paris, New York, and Boise. There's no intrinsic value in a label. All a label does is name a category, such as London, that's different in some way (perhaps in many ways) from another category, such as Paris.

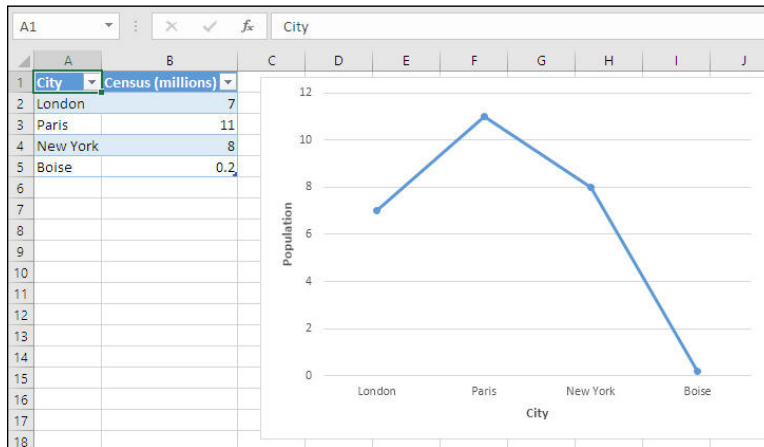
**FIGURE 9-1:**  
The horizontal axis is often termed the *X axis* and the vertical axis the *Y axis*.



If there were an intrinsic value to a category, you could put the labels in an order that made sense. But you typically can't do that with categories, except with alphabetical sorting, and now you're dealing not with categories but with letters.

On a Line chart, the horizontal or X axis is designed to display categories, which means that the different categories are equally spaced. If a category has no intrinsic value, neither do the differences between categories, and Excel just puts them on the axis. Figure 9-2 shows an example.

**FIGURE 9-2:**  
The X axis is a category axis in a Line chart.



The other kind of axis, the so-called *value axis*, takes account of intrinsic differences among values. All Line charts, such as the one in Figure 9-2, have a vertical axis that is a value axis. Notice in Figure 9-2 that the vertical distances between the data points reflect their relative values.

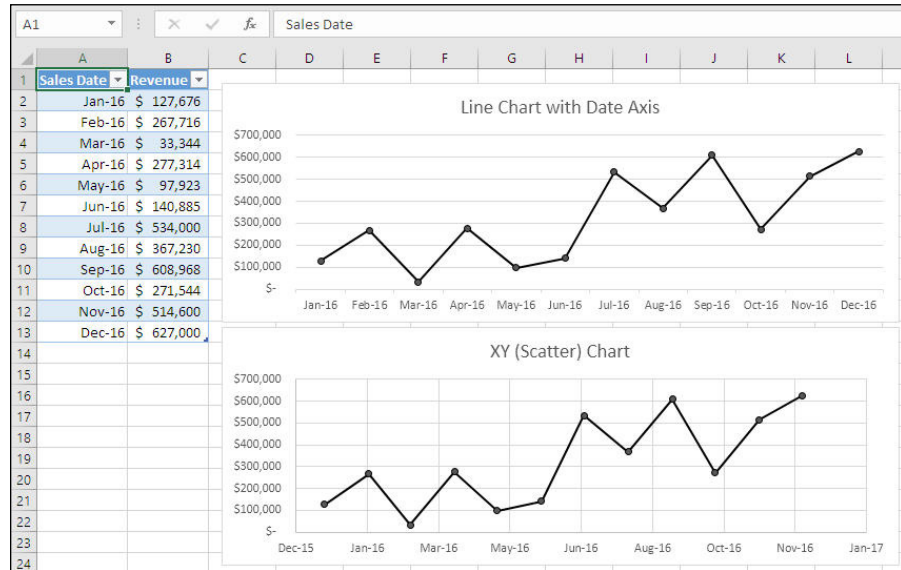
I dislike the term *value axis* because it implies that text values (such as London and New York) are not values. They are values; they're just not numeric values. A much better term would have been *numeric axis*. If it sounds like I'm kvetching here, I'm not. Things need to be given names that are descriptive and accurate, or else people get misled. But that's how Excel uses the term, so I'll do the same, under protest.

Excel has 13 standard chart types. I list them here, along with the nature of their axes and some comments:

- » **Cylinder, Cone, and Pyramid charts:** One category axis and one value axis. These charts are just 3-D variations on Column and Bar charts. These charts don't truly have three dimensions, but the data markers are formatted to look like 3-D.
- » **Pie and Doughnut charts:** A value axis only. Not suited for forecasting due to design and single axis.
- » **Radar chart:** One category axis and one value axis. Not suited for forecasting due to nonlinear layout.
- » **Area chart:** One category axis and one value axis. The problem is that the Area chart's visual design draws your eye to the area below the data points rather than to the height of the points themselves.
- » **Surface chart:** Two category axes and one value axis. A true three-dimensional chart. However, a 3-D chart, if you're forecasting, gives you too much visual information.
- » **Column and Bar charts:** One category axis and one value axis. These charts are identical except that a Column chart's X axis is a category axis and a Bar chart's Y axis is a category axis. The data markers — the columns or the bars — draw your eye away from the main issue: the value of the data series at each point on the category axis.
- » **Bubble charts:** Two value axes. There is actually a third value axis, represented by the area occupied on the chart by each data marker, so this type is unsuitable for forecasting.
- » **Line charts:** One category axis and one value axis. Excellent for forecasting.
- » **XY (Scatter) charts:** Two value axes. Excellent for forecasting.

In forecasting, you want the predictor variable to run from left to right along a horizontal axis, and the value of the forecast variable to be tied to the vertical axis. That arrangement is visually the most informative, and it's the one you get with Line charts and XY (Scatter) charts.

One problem with XY (Scatter) charts is that they tend to line up the X-axis labels poorly with the data points for the forecast variable. Figure 9-3 shows what happens (despite the presence of the lines between the points, the second chart is a scatter chart).



**FIGURE 9-3:** The Line chart does a good job of aligning labels with data points; the XY (Scatter) chart does not.



TIP

You don't have as much control over the appearance of a value axis and its labels as you do over a category axis. So, I recommend you consider using a Line chart if you're charting a variable such as sales revenue or number of units sold over time. However, be wary of calling for a trendline or its equation on the chart unless you feel comfortable with setting the horizontal axis type as a Date Axis. See the next section for more on that matter.

## Using Line charts

A closer look at the category axis on a Line chart shows you that things are a little more complicated than they seem. To get a closer look, take these steps:

- 1. Click somewhere in a Line chart to activate it.**

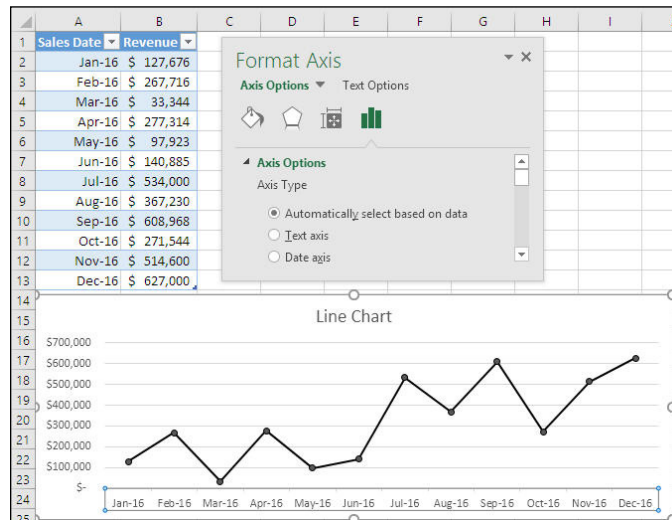
It doesn't matter what part of the chart you select, but make sure that you've activated a Line chart.

- 2. Right-click the chart's category axis.**

That's the chart's horizontal axis.

### 3. Choose Format Axis from the shortcut menu.

See Figure 9-4.

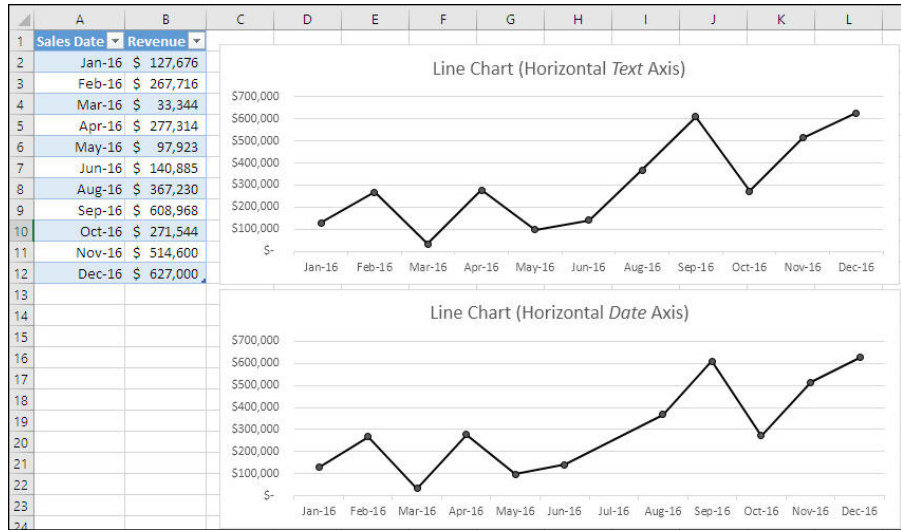


**FIGURE 9-4:** In Excel terminology, the points in the chart (here, connected by lines) are called a *data series*.

A category axis can have three subtypes (and this is true of the category axis in any type of chart, given that the chart has a category axis). Suppose you want to use whatever's in A2:A13 as the values for your category axis. On the Axes tab, you can choose from:

- » **Automatically select based on data:** This is the default. If you select this option, you let Excel decide for you. Excel uses some persnickety rules to decide whether to treat the category axis as a series of text values (true categories such as Urban, Suburban, and Rural) or as dates. The rules involve looking for text values in A2:A13, and whether any numeric values are formatted as dates. You can override Excel's decision using either the Text or the Date option.
- » **Text axis:** Selecting this option forces Excel to treat the values as categories — that is, as though they were text labels. Doesn't matter if the values in A2:A13 are genuinely numeric, either with or without a date format. If you choose this option, Excel treats the values in A2:A13 as text labels, and we're back to London, Paris, and so on.
- » **Date axis:** The only reason you'd select this option is that you have dates in A2:A13, you want to use them on the category axis, and those dates are in a Number rather than a Date format. Figure 9-5 shows the difference between an X axis that's a true category axis, and an X axis that Excel has scaled for dates.

**FIGURE 9-5:**  
The true category axis spaces the values evenly, although July is missing. The time-scaled category axis leaves room for the missing data.



What does all this have to do with forecasting? Suppose your Line chart's X axis is a true category axis. Excel charts can show you a trendline that depicts the relationship between the predictor variable and the forecast variable (see Chapter 7). Charts can also show you the R-squared value and the regression equation itself.

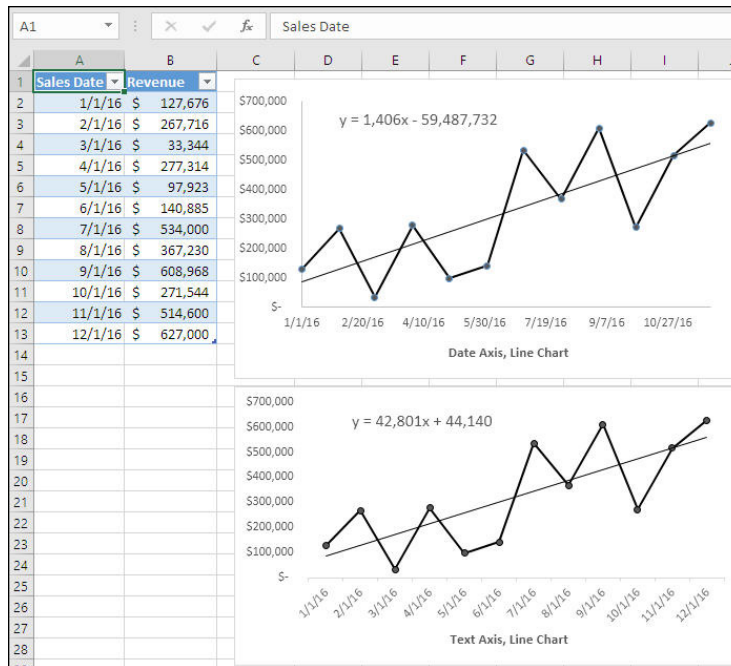
In a chart with a date-scaled category axis, the trendline and the regression information are based on the relationship between the predictor and the forecast variable.

In a chart with a true category axis, Excel has to use some numeric value to represent the categories for the trendline and regression, and it uses 1, 2, 3, and so on for its calculations. In other words, because Excel can't calculate a regression where the predictor variable is London, Paris, New York, and Boise, it would treat London as the numeral 1, Paris as the numeral 2, and so on. Figure 9-6 shows you the result of this if you have dates on the category axis, first as a date-scaled axis and second as a true category or text axis.

The chart with the time-scaled category axis determines the regression equation by calculating the relationship between the sales figures and the date figures (the dates, in Number format, begin with 42,370 and end with 42,705). The regression equation as you'd use it to get the forecast for January 2017 would be:

$$\text{January 2017 Sales} = 1406 * 42,736 - 59,487,732$$





**FIGURE 9-6:** The regression equations are wildly different in the two charts.

The chart with the true category axis determines the regression equation by calculating the relationship between the sales figures and the category numbers (that is, 1, 2, 3, . . . 12). The regression equation as you'd use it to get the forecast for January 2007 would be:

$$\text{January 2017 Sales} = 42,801 * 13 + 44,140$$

So if you're going to make use of these equations, you'd better know whether your next predictor value is the 39,083rd day from January 1, 1900, or whether it's the 13th category ordinal number from the beginning of the series. For example, you don't want to be confusing the date values with the categories and entering in your worksheet an equation like this:

$$\text{January 2007 Sales} = 1406 * 13 - 59,487,732$$

which results in a forecast of \$59,469,454 for January 2017. If you turned that forecast in, you'd probably be reassessing your career goals soon after.



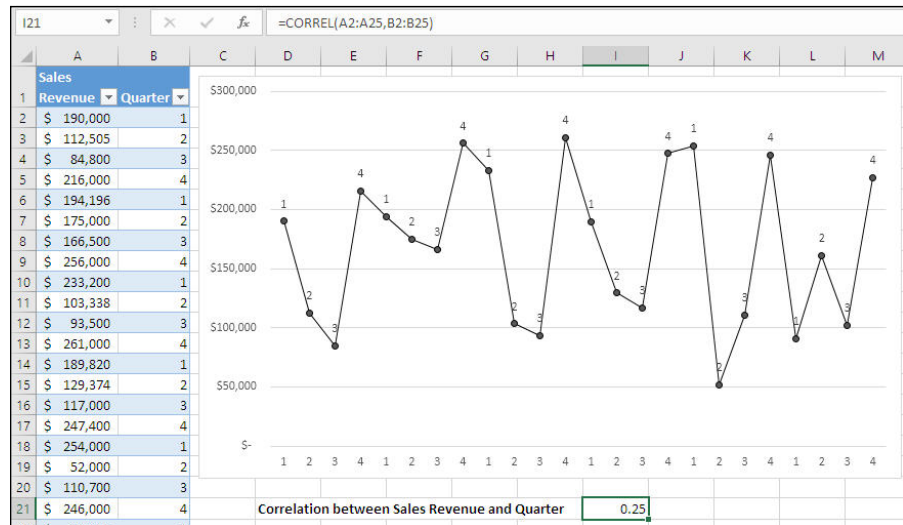
WARNING

Depending on the scale of measurement, the decimal places in the coefficients may or may not be important. If they are important, you should probably get the regression equation on the worksheet with **LINEST** — or, you should bypass the equation and work with the **TREND** function instead, which will give you the

forecast values without going through a lot of formulaic hand waving. (See Chapter 12 for information about these functions.)

Here's another example. There's some real seasonality going on in Figure 9-7.

**FIGURE 9-7:**  
A linear correlation often obscures the seasonal regularity in a baseline.



Nearly every time the fourth quarter comes along, it has the highest sales revenue for the current year. But you wouldn't suspect such regularity in the relationship between quarter of the year and sales if all you looked at was the correlation — 0.25 is marginal at best.

On the other hand, if you look at the baseline sales and quarters in a chart, the seasonality is hard to miss.



TIP

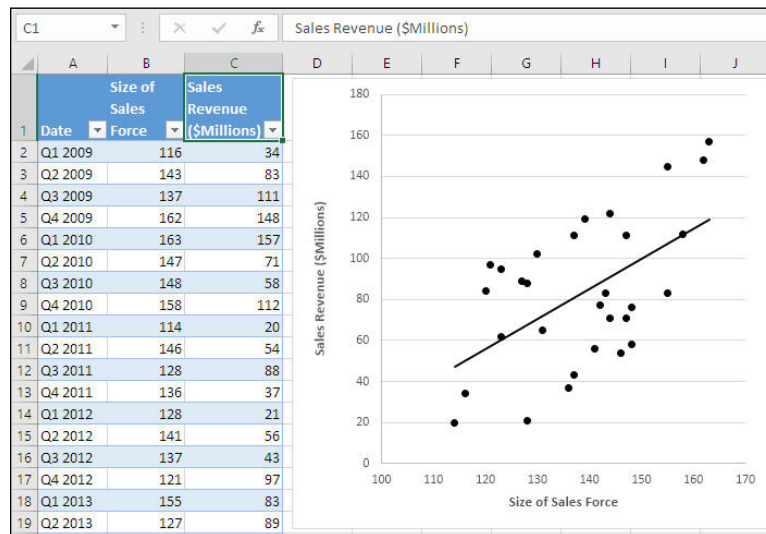
The numbers above the individual data points in the chart in Figure 9-7 are called *data labels*. In this case, each data label shows the quarter of the year that the revenue amount belongs to. There are different ways to get data labels. One is to right-click the data series in the chart and choose Add Data Labels from the shortcut menu. You'll get the charted values (here, sales revenues) as data labels. If you want to show different values as the labels, as here, right-click a data label and choose Format Data Labels from the shortcut menu. Fill the Values From Cells check box and drag through the range of your preferred labels. Then clear the Value check box.

You can find information about forecasting baselines that have this sort of seasonality in Chapter 18.

## Using XY (Scatter) charts

An XY (Scatter) chart has two value axes: the X axis is a value axis, as is the Y axis. (The rest of this section simplifies things by just calling them *XY charts*.) XY charts are best used when you're forecasting on some basis other than date. Refer to Figure 9-3 to see why Line charts are better than XY charts if you're forecasting according to some date predictor. The dates on the XY chart's horizontal axis don't line up quite right with the charted revenues. Admittedly, it's a small downside.

But if you have a different sort of predictor variable, such as size of sales force or advertising dollars, consider using an XY chart. Figure 9-8 shows an example.



**FIGURE 9-8:** Here, you're more interested in how well size of sales force predicts sales than in date as a predictor variable.

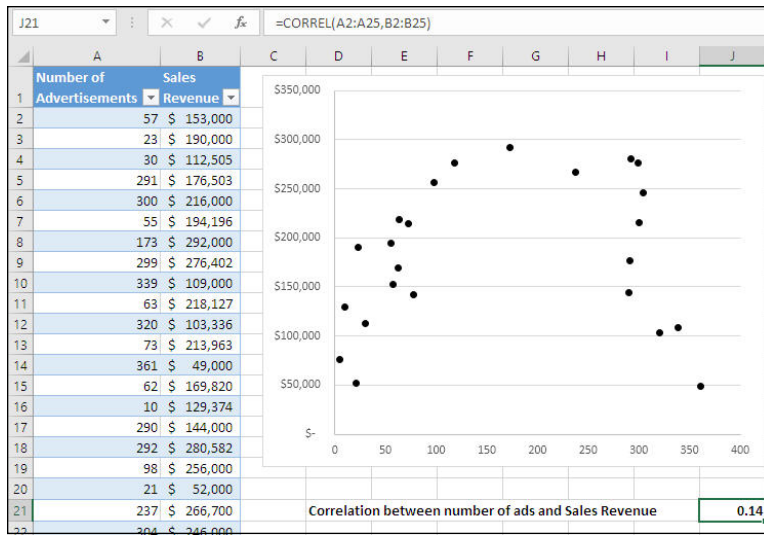
XY charts help you judge the strength of a relationship between a predictor variable and a forecast variable. People often take a look at the correlation between the two variables and, if the correlation is, say, between  $-0.2$  and  $0.2$  then they decide there's not much to it.



REMEMBER

A correlation's possible values run from  $-1.0$  to  $1.0$ . The closer to one of those two values, the stronger the relationship. The closer to  $0.0$ , the weaker the relationship. The R-squared value is the square of the correlation and measures how much variation in one variable is attributable to the other variable.

But look at the XY chart in Figure 9-9.



**FIGURE 9-9:** This pattern is often termed an *inverted U function*. The name doesn't matter; going up and then back down does.

If you just looked at the correlation between number of advertisements and number of sales, you'd ignore it. A correlation of 0.14 is really too small to concern yourself with unless it's based on hundreds or thousands of cases.

But when you look at the chart of the number of advertisements against sales, a different picture emerges. You can see that, up to a certain number of ads, sales increase fairly sharply. Then the sales flatten out and finally drop back off. This inverted U can come about if there's a relatively fixed advertising budget. Then, when the company starts buying a relatively large number of ads, the per-advertisement expense decreases, and so do things like the desirability of the placement, the size of the ad, whether or not color is used, and so on. As a result, sales decline.

So there's a useful relationship to take advantage of. But because a simple correlation assumes a linear relationship, it won't tell you what you want to know. This is only one reason that you should always chart your baselines.

## Making Your Data Dance with Pivot Charts

Chapter 8 has a lot to say about how you can use Excel's PivotTable feature to summarize individual sales records in a table. The pivot table then could show you total sales revenues according to date, region, product line — any variable that you might want to use to break down the revenue results.

Since all the way back in Excel 2000, an added feature has been included with the pivot table capability: a pivot chart. You can still build a pivot table, just as you always could. But now you can decide to build a pivot chart, and a pivot table comes with it. You have available all the standard chart types, from Column to Pyramid, and of course including Line and XY (Scatter).

If you've looked at Chapter 8, you know the basics of setting up a pivot table in Excel. The process of setting up a pivot chart is the same in concept, just a little different in execution.

Why the difference? At heart, it's because a pivot table expresses the size of its value field — that is, the field that the pivot table is summarizing — with a number. For example:

- » The sum of sales dollars during each month, for each product line, for each branch
- » The average number of units sold of each product, in each branch, for each sales rep in that branch
- » The sum of revenue, for goods separately from services, for each product line in each branch

But a chart can't show sums, averages, counts, and so on in a single cell. That's both the advantage and the disadvantage of a chart. The disadvantage: You have to reserve an entire dimension, up for Line (and other) charts, up and across for XY charts, to show the magnitude that you see in a single cell of a pivot table. The advantage (at least as to sales forecasting): You can see how the magnitude grows, falls, rises and falls seasonally, or remains generally static over time.

Start with the baseline data laid out in a table structure, as in Figure 9-10.

Then take these steps:

- 1. Select any cell in the table.**
- 2. Go to the Ribbon's Insert tab and click Pivot Chart in the Charts group.**  
The Create PivotChart dialog box appears (see Figure 9-11).
- 3. Select the Existing Worksheet option button.**
- 4. Click in the Location edit box and then click in the worksheet cell where you want the results to begin.**
- 5. Click OK to start defining the pivot chart. See Figure 9-12.**

	A	B	C	D	E
1	Month	Branch	Sales Rep	Product	Sales
2	Dec	NE	Timms	Notebook	\$ 13,435
3	Jul	NE	Norris	Notebook	\$ 5,801
4	Apr	SW	Finney	Tablet	\$ 24,217
5	Dec	NW	Timms	Tablet	\$ 17,121
6	Dec	NW	Timms	Notebook	\$ 12,461
7	Nov	NW	Smith	Notebook	\$ 6,748
8	Jul	SW	Norris	Notebook	\$ 14,263
9	Feb	SW	Anderson	Tablet	\$ 23,768
10	Sep	SW	Roberts	Notebook	\$ 10,854
11	Sep	SW	Roberts	Tablet	\$ 6,576
12	Dec	SW	Timms	Tablet	\$ 29,620
13	Feb	NW	Anderson	Tablet	\$ 25,594
14	Mar	NW	Edwards	Notebook	\$ 27,158
15	Nov	NE	Smith	Tablet	\$ 22,948
16	Jan	SE	Allen	Tablet	\$ 10,238
17	May	SE	Johnson	Notebook	\$ 10,529
18	Mar	NE	Edwards	Tablet	\$ 27,963
19	Sep	NE	Roberts	Notebook	\$ 3,467
20	Oct	SW	Rodgers	Notebook	\$ 21,470
21	Apr	NE	Finney	Tablet	\$ 11,776
22	May	SE	Johnson	Tablet	\$ 11,307
23	Sep	SE	Roberts	Notebook	\$ 21,806
24	Oct	SW	Rodgers	Tablet	\$ 12,129

**FIGURE 9-10:** When you're preparing a pivot table or pivot chart, the order of the data in the table doesn't matter.

The screenshot shows the 'Create PivotChart' dialog box in Microsoft Excel. The background is a table with columns labeled 'Month', 'Branch', 'Sales Rep', 'Product', and 'Sales'. The dialog box has the following options:

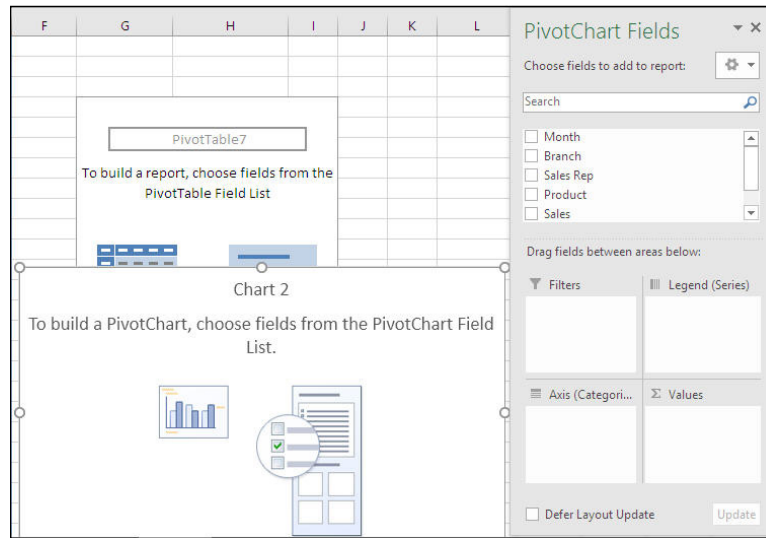
- Choose the data that you want to analyze:**
  - Select a table or range: Table14
  - Use an external data source
- Choose where you want the PivotChart to be placed:**
  - New Worksheet
  - Existing Worksheet
- Choose whether you want to analyze multiple tables:**
  - Add this data to the Data Model

Buttons for 'OK' and 'Cancel' are visible at the bottom right of the dialog box.

**FIGURE 9-11:** Because you began by selecting a cell in the table, Excel fills in the complete reference to the data source for you.

At this point you start designing the pivot chart by dragging fields into the four areas at the bottom of the PivotChart Fields pane. Take these steps to complete the current example:

**FIGURE 9-12:**  
At this point you can use the PivotTable Field List box to drag fields into the areas where you want them on the chart.



1. Click on **Month** in the Field List, hold down the mouse button, and drag **Month** down into the **Axis (Categories)** area. Release the mouse button.
2. Click on **Product** in the Field list and, as you did with **Month**, drag it into the **Axis (Categories)** area, just below **Month**.

When you release the mouse button, you'll have established an *inner* field: Each value of **Product** will be repeated within each value of **Month**.

3. Click on **Branch** in the Field List and drag it into the **Legend (Series)** area.

Now you'll have a different data series, each representing a different branch, in the pivot chart for each combination of **Month** and **Product** in your table.

4. Drag the **Sales** button from the Field list into the **Σ Values** area.

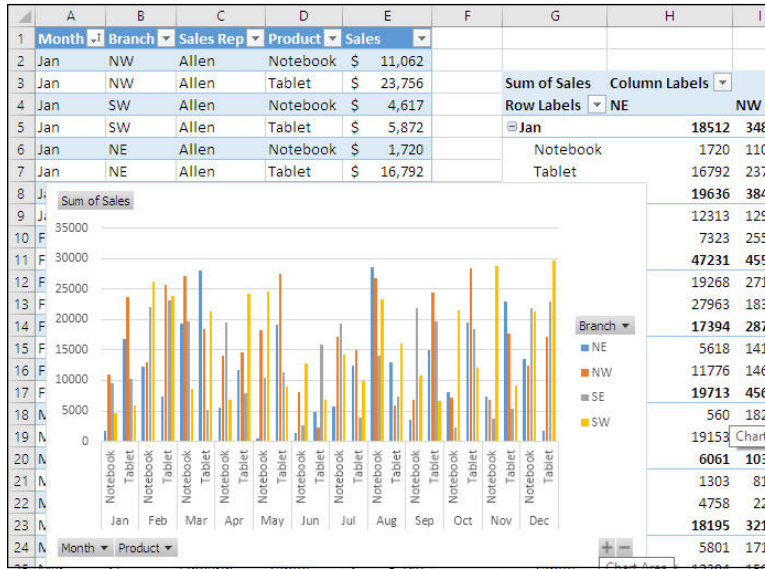
The result appears as in Figure 9-13.

You can convert the Column chart to a Line chart easily:

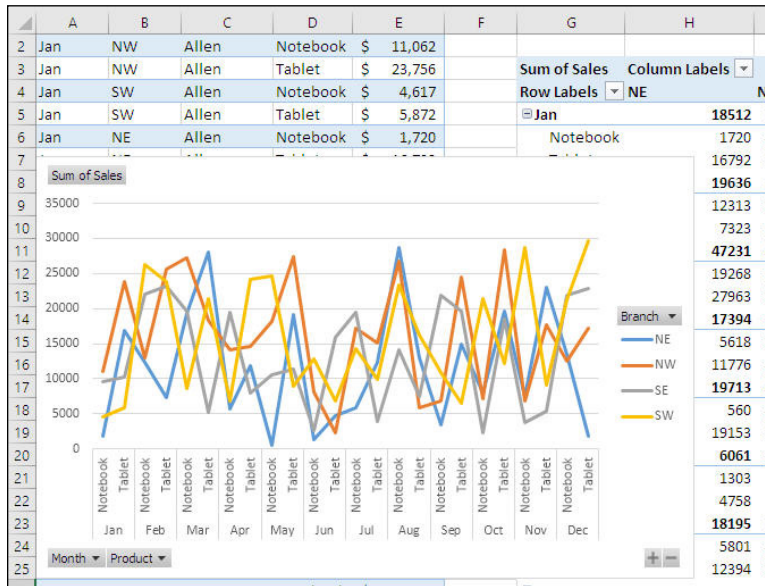
1. Select the pivot chart.
2. Go to the Ribbon's **Design** tab.
3. Click **Change Chart Type**.
4. Choose **Line Chart** and click **OK**.

Figure 9-14 shows the result.





**FIGURE 9-13:** A Column chart is normally not very useful for depicting a baseline. A Line chart or an XY chart is a better choice.



**FIGURE 9-14:** This is far too muddled. You should probably move Branch into the Page area and look at each Branch one by one. Or click the Branch field button and look at them two by two.

After you've created the pivot chart, if you think a different layout might make better visual sense, you can drag a field button from one area to another. For example, in Figure 9-14, you might try dragging the Branch button into the Category area along with Month, and the Product Line button from the Category area into the Series area. You might think that's an improvement (I don't), but whether



or not you do, bear in mind that you *can* do so. That should be easy to bear in mind: It's why they call it a pivot chart.



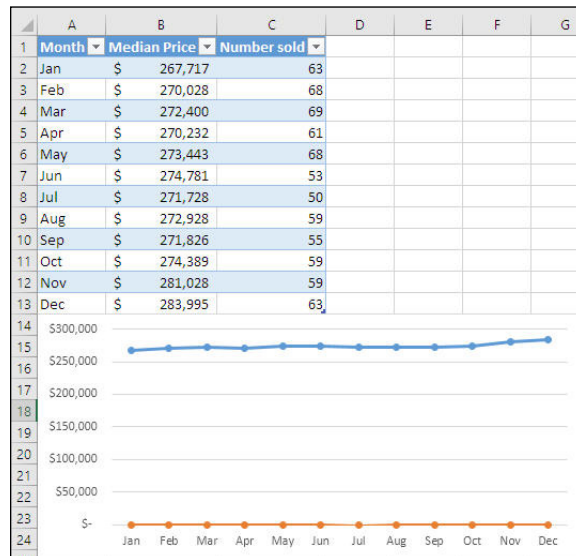
TIP

Take a look at the pivot table that accompanies the pivot chart. You can control where the pivot table is placed. It often makes good sense to put the chart on the same worksheet as your table occupies, assuming you're using a table as the basis for the report.

## Using Two Value Axes

There are times when you have put two data series in a chart, and they're on very different scales of measurement. One series can bury the other at the bottom of the chart.

For example, suppose your company builds and sells residential houses. You want to build a baseline of number of housing units you've sold per month, as well as the median regional price of single family detached dwellings each month. Figure 9-15 shows what that baseline might look like, along with a chart of the baseline.



**FIGURE 9-15:** Both data series — housing units and median price — appear on the same Y axis scale.

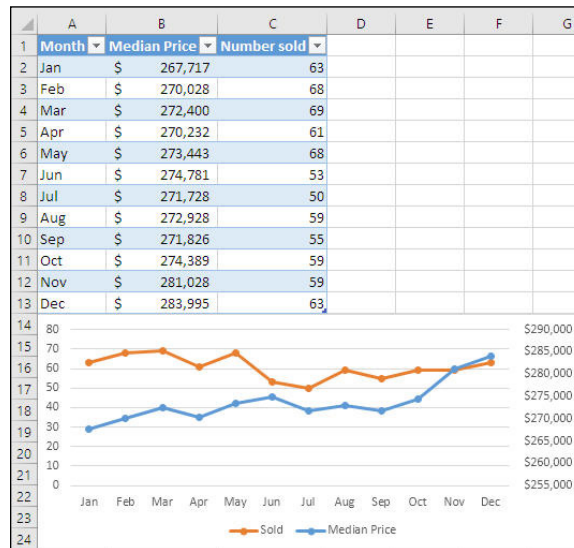
The fact that one data series, Units Sold, gets no larger than 69, and the other data series, Median Regional Price, ranges from \$267,717 to \$283,995 means that you can't make out differences in either.

Excel charts provide a good solution with two value axes. If you run into a situation like this, follow these steps:

1. Click on the chart to activate it.
2. Right-click on either data series to select it.
3. Choose Format Data Series.
4. Under Series Options, click the Secondary Axis button.

The Primary and Secondary Axis buttons are not available unless you have more than one data series in your chart.

The result appears in Figure 9-16.



**FIGURE 9-16:** With a primary and a secondary scale, you can see that Units Sold is holding steady as Median Regional Price is rising.



TIP

When you have two scales in one chart, labeling them can be helpful. After you've created the secondary scale, go to the Ribbon's Design tab and click Add Chart Element and choose Axis Titles. Then supply a name for the Primary Vertical axis and the Secondary Vertical axis.

## Chapter 10

# Forecasting with Excel's Data Analysis Add-in

**W**ay back in the mid-1990s, Microsoft arranged to collect some statistical analysis tools in a single package, to accompany Excel 95 — the first version of Excel to take advantage of Windows 95, Microsoft's then new operating system for PCs.

Perhaps inspired by the overwhelming commercial success of such brand names as Kwik Kar Wash, Tastee Freez, and Rite Aid, Microsoft decided to give its collection of statistical tools a catchy name. They chose *Analysis ToolPak*, which was often abbreviated as *ATP*.

Microsoft apparently now judges that the earlier name was an aberration, an attack of the quaints and the cutes. It has renamed its collection of tools the *Data Analysis add-in*. The term *Analysis ToolPak* still appears here and there in the Excel application, such as in a list box of available add-ins. But you get to its tools by clicking *Data Analysis* on the Ribbon's *Data* tab, and that's the important part. It's also important to bear in mind that if you ever spent much time learning how to use the *ATP*, its functionality is still available in the *Data Analysis add-in*. Only the name has changed.

*Add-ins* are collections of BASIC-like code — code that, fortunately, you never have to see. The idea behind add-ins is that they can extend Excel's reach, usually to do specialized tasks for you, such as forecasting using moving averages, exponential smoothing, and regression.

Because these tasks are not standard ones that Excel handles directly, such as inserting columns or constructing PivotTable reports, you have to make special provisions. In particular, this means installing add-ins, both onto your computer and into Excel. Installing add-ins is a straightforward process, but it can be hard to find out how to do it. I show you how in this chapter.

The collection of Data Analysis tools is an add-in, one of the few that is distributed with the retail version of Excel. It has several tools that help you do forecasting. After you've installed the add-in, you have choices. It's a tool kit, after all, so it has a bunch of tools in it, and you need to know which one to use when you're ready to make a forecast.

There are three basic methods of making quantitative forecasts: moving averages, exponential smoothing, and regression. (More advanced methods use these three basic approaches as building blocks.) In this chapter, I give you some recommendations about how to choose between these methods to do your forecasts. There's no pat answer: Much depends on the nature of your baseline. But this chapter does offer you some guidance on what to pay attention to when you're choosing among the methods.

## Installing Add-ins: Is the Add-in Even There?

Add-ins are not at the top of the food chain at Microsoft. The tasks that add-ins perform may be important enough to automate, but they're not regarded as important enough to become a full-fledged part of the Excel application. (If add-ins did enjoy that degree of positive regard, there'd be a Data Analysis option on the Ribbon's Data tab right out of the box, just like Sort or Filter.)



TIP

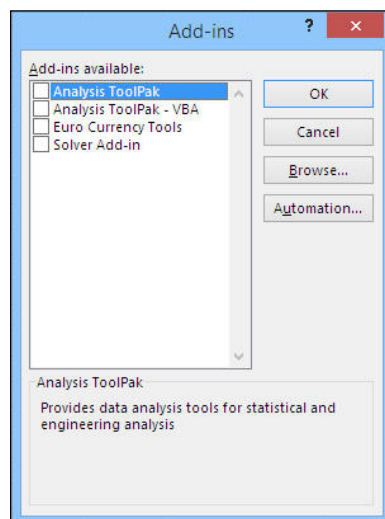
Developers other than Microsoft *do* offer lots of add-ins. A Google search in 2016 for the words *Excel* and *add-in* returned nearly 30 million hits; that's compared to half a million hits in 2005. Lots of these pages offer add-ins for sale. If you know how to code using Visual Basic for Applications and have a copy of Excel, you can create an add-in and post it for sale on a website. If you're looking for some sort of specialized capability that Excel doesn't offer, but that it could, check out the net — but be prepared to get something less than what you're looking for.

First, you have to get the add-in onto your computer. Then you have to get the add-in into Excel. The following sections describe how to do that for the Data Analysis add-in.

Do a quick check, first, by making sure that the Data Analysis add-in isn't installed already. Start Excel and go to the Ribbon's Data tab. Look in the Analyze group for an icon labeled Data Analysis. If you see it, you're probably good to go. (I weasel with "probably" just as a CYA maneuver. With add-ins, you always run a certain risk that someone has installed something that isn't the Data Analysis add-in at all, but that nevertheless puts the Data Analysis item in the Analyze group. Don't worry about it. If people wanted to put ransomware on your computer, they'd choose a better way. Skip ahead to the section titled "Using Moving Averages.")

If you don't see Data Analysis in the Analyze group, you have a little added work to do. The add-in may still be on your computer, but no one told Excel. Take these steps:

1. In Excel, click the File tab.
2. Choose Options from the navbar at the left of the Excel window.
3. Choose Add-Ins from the navbar at the left of the Excel Options window. Click OK.
4. Make sure that the Manage drop-down near the bottom of the Excel Options window contains *Excel Add-ins*. Click Go.
5. The Add-ins dialog box appears as in Figure 10-1. Make sure that the check box next to Analysis ToolPak (*sic*) is checked, and click OK.



**FIGURE 10-1:** There's an installation problem if you don't see Analysis ToolPak in the list box.

If the list box shown in Figure 10-1 doesn't show an Analysis ToolPak item, your best bet is to get in touch with whoever installed Excel on your computer and complain bitterly.



TIP

As long as you're here at the Add-ins dialog box, you might as well select the Solver check box if it isn't already. Excel's Solver is a powerful utility that's absolutely indispensable when it comes to forecasting with the exponential smoothing methods.



TIP

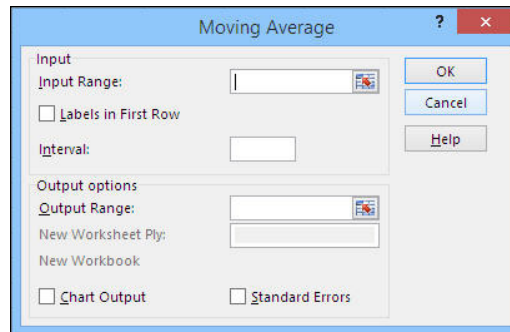
If you think you may want to use some of the special functions in the Data Analysis add-in *in your own VBA code*, select both the Analysis ToolPak and the Analysis ToolPak – VBA check boxes. Otherwise, select just the Analysis ToolPak check box. This is one of the few locations in Excel that, in Excel 2016, still make reference to the Analysis ToolPak with that name.

## Using Moving Averages

After you have the Data Analysis add-in installed and you've made it available to Excel, you can select any one of its analysis tools and run that analysis on the input data that you supply. In the world of forecasting, that means the baseline that you've gathered and structured properly on a worksheet.

The first tool you might consider — if only because it's the easiest to use and understand — is the Moving Average tool. As always with the add-in, begin by going to the Ribbon's Data tab and choosing Data Analysis. In the Analysis Tools list box, select Moving Average and click OK.

The Moving Average dialog box, shown in Figure 10-2, appears. You can find instructions on how to use this dialog box in Chapter 13.



**FIGURE 10-2:** The Interval is the number of actuals from your baseline to use in each moving average.

## Moving day: Getting from here to there

As easy as moving averages are to set up and understand, you take on an additional responsibility when you decide to forecast with them. The issue is how many time periods from your baseline you should include in each moving average.



TIP

It may go without saying, but I'll say it anyway: Use the same number of actual observations in calculating each moving average. If the first moving average that you have Excel calculate uses three periods from the baseline, then all the moving averages in your forecast use three periods.

You want to select the right number of periods:

- » If you use too few, the forecasts will respond to random shocks in the baseline, when what you're after is to smooth out the random errors and focus on the real drivers of your sales results.
- » If you use too many, the forecasts lag behind real, persistent changes in the level of the baseline — maybe too far for you to react effectively.

When you decide to use the Moving Average tool — or, more generally, to use moving averages regardless of whether you use the tool or enter the formulas yourself — you're taking a position on the effect of recent baseline values versus the effect of more distant baseline values.

Suppose you have a baseline that extends from January 2016 to December 2016, and you use a three-month moving average of sales results for your forecasts. The forecast for January 2017 would be the average of the results from October, November, and December 2016. That forecast is dependent entirely on the final quarter of 2016 and is mathematically independent of the first three quarters of 2016.

What if instead you had chosen a six-month moving average? Then the forecast for January 2017 would be based on the average of July through December 2016. It would be entirely dependent on the second half of 2016, and the first half of 2016 would have no direct influence on the January 2017 forecast.

It could well be that either of these situations — or another one, such as a two-month moving average — is exactly what you want. For example, you may need your forecast to emphasize recent results. That emphasis can be especially

important if you suspect that a recent event, such as a significant change in your product line, will have an effect on sales.

On the other hand, you may not want to emphasize recent sales results too much. Emphasizing recent sales results can obscure what's going on with your baseline in the long term. If you're not sure how much to emphasize recent results, you have a couple of good options:

- » **Experiment with different numbers of time periods to make up your moving averages.** This approach is often best. A way of evaluating different moving average lengths is in Chapter 15; it's tailored for exponential smoothing, but it's easily applied to moving averages.
- » **Use exponential smoothing, which uses the entire baseline to get a forecast but gives greater weight to the more recent baseline values.** Exponential smoothing gives a little less weight to the next-to-last baseline value, a little less weight to the one before that, and so on all the way back to the first baseline value, which has the least amount of influence on the next forecast. (See "Using Exponential Smoothing," later in this chapter, for more information.)

## Moving averages and stationary baselines

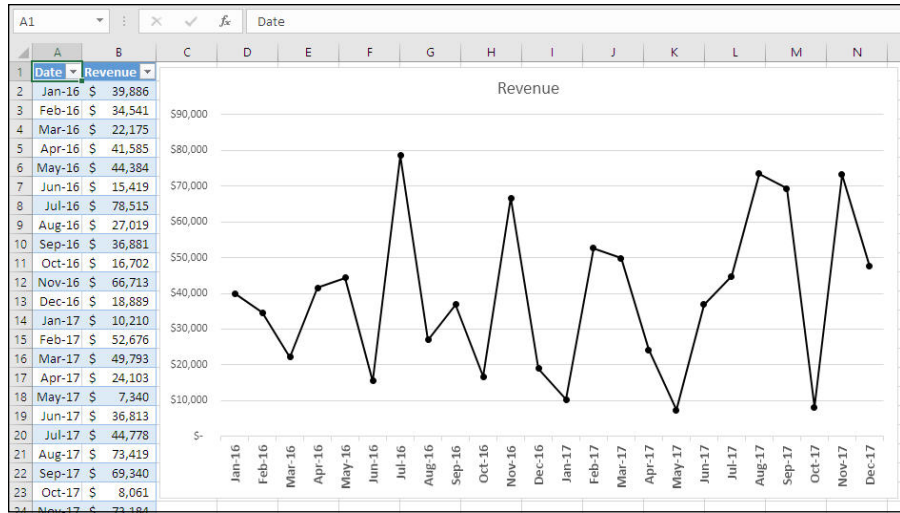
Moving averages are well suited to *stationary baselines* (baselines whose levels do not generally increase or decrease over a long period of time). You can use moving averages with baselines that trend up or down, but you should usually detrend them first (see Chapter 17 for more information) or else use one of the more complicated moving-average models, which I don't cover in this book.

How do you tell a stationary baseline from one that is trending up or down? One way is to look at it. Figure 10-3 has an example. The baseline in Figure 10-3 certainly looks stationary. It has spikes and peaks and valleys, but overall the baseline doesn't appear to trend up or down.

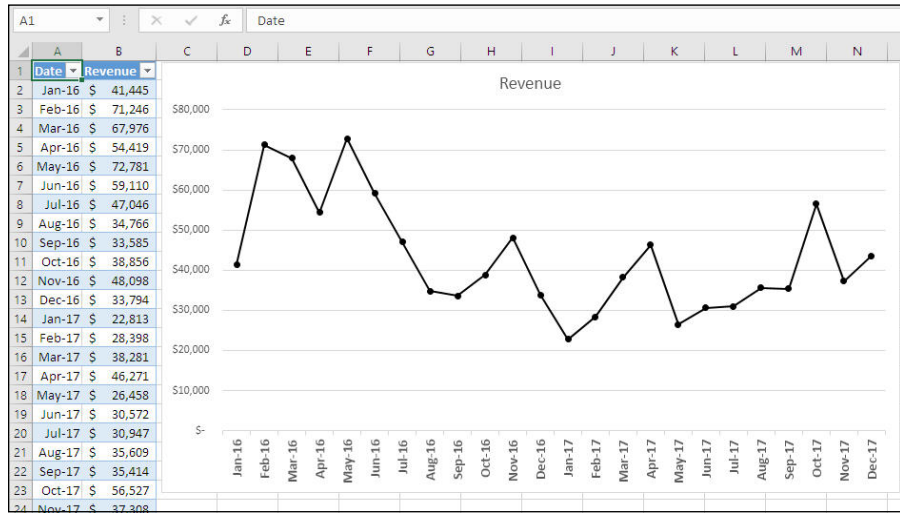
The problem with just looking at the baseline is that sometimes it's not entirely clear whether it's stationary or trended. What do you think about the baseline in Figure 10-4? Looking at the chart, it's hard to say whether the baseline is stationary. It might be, but then again it might really be drifting gradually down. You can make a quick test by checking the correlation between date and revenue. (See Chapter 4 for details.)



**FIGURE 10-3:**  
Over a longer period of time (say, six years rather than two), this baseline may turn out to be part of a cycle. But for shorter-term purposes, this is a stationary baseline.



**FIGURE 10-4:**  
This baseline looks as though it may be gently heading down. Adding a trendline to it can help you interpret what's going on.



## Using Exponential Smoothing

The preceding section on moving averages implies that you can solve the problem of how many baseline values to include in a moving average by using exponential smoothing. And that's true, as far as it goes: You don't have to make that decision, because the entire baseline is involved to one degree or another.

Take, for example, the forecast for December. It may depend 30 percent on November's actual, 20 percent on October's actual, 15 percent on September's actual, and so on all the way back to the start of the baseline. The older the actual, the less its influence on the next forecast.

But you're saddled with a new problem: How much do you want the most recent actual to influence the subsequent forecast? The best way to make this decision — at the very least, a preliminary decision — is to try different amounts of influence and see how much error each amount causes.

There's a lot packed into that last sentence. Seeing its meaning is easier in a worksheet than in words. Figure 10-5 has an example.

		Smoothing Constant						
	A	B	C	D	E	F	G	H
2	Date	Revenue	0.3	0.4	0.5	0.6	0.7	0.8
3	Jan-16	\$ 41,175	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
4	Feb-16	\$ 47,504	\$ 41,175	\$ 41,175	\$ 41,175	\$ 41,175	\$ 41,175	\$ 41,175
5	Mar-16	\$ 45,632	\$ 43,074	\$ 43,707	\$ 44,340	\$ 44,972	\$ 45,605	\$ 46,238
6	Apr-16	\$ 59,351	\$ 43,841	\$ 44,477	\$ 44,986	\$ 45,368	\$ 45,624	\$ 45,753
7	May-16	\$ 78,320	\$ 48,494	\$ 50,426	\$ 52,168	\$ 53,758	\$ 55,233	\$ 56,631
8	Jun-16	\$ 98,703	\$ 57,442	\$ 61,584	\$ 65,244	\$ 68,495	\$ 71,394	\$ 73,982
9	Jul-16	\$ 28,346	\$ 69,820	\$ 76,432	\$ 81,974	\$ 86,620	\$ 90,510	\$ 93,759
10	Aug-16	\$ 20,000	\$ 57,378	\$ 57,197	\$ 55,160	\$ 51,656	\$ 46,995	\$ 41,429
11	Sep-16	\$ 29,347	\$ 46,165	\$ 42,318	\$ 37,580	\$ 32,662	\$ 28,099	\$ 24,286
12	Oct-16	\$ 20,898	\$ 41,119	\$ 37,130	\$ 33,463	\$ 30,673	\$ 28,972	\$ 28,335
13	Nov-16	\$ 49,398	\$ 35,053	\$ 30,637	\$ 27,181	\$ 24,808	\$ 23,320	\$ 22,385
14	Dec-16	\$ 24,830	\$ 39,356	\$ 38,141	\$ 38,289	\$ 39,562	\$ 41,575	\$ 43,995
15	Jan-17	\$ 21,218	\$ 34,999	\$ 32,817	\$ 31,560	\$ 30,723	\$ 29,853	\$ 28,663
16	Feb-17	\$ 21,122	\$ 30,864	\$ 28,177	\$ 26,389	\$ 25,020	\$ 23,809	\$ 22,707
17	Mar-17	\$ 48,695	\$ 27,942	\$ 25,355	\$ 23,755	\$ 22,681	\$ 21,928	\$ 21,439
18	Apr-17	\$ 20,000	\$ 34,168	\$ 34,691	\$ 36,225	\$ 38,289	\$ 40,665	\$ 43,244
19	May-17	\$ 20,000	\$ 29,917	\$ 28,815	\$ 28,113	\$ 27,316	\$ 26,199	\$ 24,649
20	Jun-17	\$ 20,901	\$ 26,942	\$ 25,289	\$ 24,056	\$ 22,926	\$ 21,860	\$ 20,930
21	Jul-17	\$ 23,081	\$ 25,130	\$ 23,534	\$ 22,479	\$ 21,711	\$ 21,189	\$ 20,907
22	Aug-17	\$ 70,793	\$ 24,515	\$ 23,353	\$ 22,780	\$ 22,533	\$ 22,513	\$ 22,646
23	Sep-17	\$ 45,690	\$ 38,399	\$ 42,329	\$ 46,786	\$ 51,489	\$ 56,309	\$ 61,164
24	Oct-17	\$ 79,861	\$ 40,586	\$ 43,673	\$ 46,238	\$ 48,010	\$ 48,876	\$ 48,785
25	Nov-17	\$ 68,942	\$ 52,368	\$ 58,148	\$ 63,050	\$ 67,120	\$ 70,565	\$ 73,646
26	Dec-17	\$ 34,473	\$ 57,341	\$ 62,466	\$ 65,996	\$ 68,213	\$ 69,429	\$ 69,883
27	Jan-18 Forecast:	\$ 50,480	\$ 51,269	\$ 50,234	\$ 47,969	\$ 44,960	\$ 41,555	
29	Error Summary	\$ 23,494	\$ 23,398	\$ 23,429	\$ 23,566	\$ 23,796	\$ 24,128	

**FIGURE 10-5:**  
This sort of analysis is much easier to set up if you enter the formulas yourself, instead of relying on the add-in.

These components of an exponential smoothing analysis are in Figure 10-5:

- » The baseline itself in cells A3:B26. (Here I'm using the more relaxed usage of the term *baseline*, to include both the actual sales results and the associated dates.)
- » Several different constants, ranging from 0.3 to 0.8, in cells C2:H2.

- » The constants in row 2 are used to create the forecasts in cells C4:H26 and C28:H28. Each column from C to H contains a series of forecasts that is based on the constant at the top of that column. Chapter 15 shows you how to create the forecasts using the constants.
- » The #N/A values in row 3 are due to the fact that, without earlier baseline data, you can't make a forecast for January 2016.
- » The values in row 30 are measures of the overall amount of error in the forecasts. For example, there's an error associated with the forecast value in cell C5: It's the difference of \$2,558 between the forecast in cell C5 and the actual value observed for March 2016.

The errors in forecasting — the differences between the forecasts and the actuals — are massaged and totaled until you wind up with a summary of the errors given a particular smoothing constant. That summary appears for each smoothing constant in C30:H30. You can also find the method for getting these summaries in Chapter 15.

One primary goal of forecasting is to minimize the errors of your forecasts: All other things being equal, the closer your forecast comes to the actual result the better. In Figure 10-5, the smallest measure of total forecast error is in cell D30: \$23,398. So, you'd use the constant that created the forecasts in that column: the value in cell D2, or 0.4.

This technique is useful not only for deciding how large a constant to use, but also for deciding between using moving averages or exponential smoothing. You can calculate the error summary for both a moving average of a certain length and exponential smoothing with a given constant. Then select the approach that provides the smaller measure of forecast error.



TECHNICAL  
STUFF

Instead of creating a series of forecasts, as in columns C through H of Figure 10-5, you can also use Excel's Solver to help you choose the best constant. (This is one reason I recommend, earlier in this chapter, that you install the Solver add-in along with the Data Analysis add-in.) Although Excel's built-in Goal Seek tool is another possibility, you can establish better control using Solver. You would have Solver minimize the forecast error summary by selecting the optimum value of the constant for the smoothing forecast.



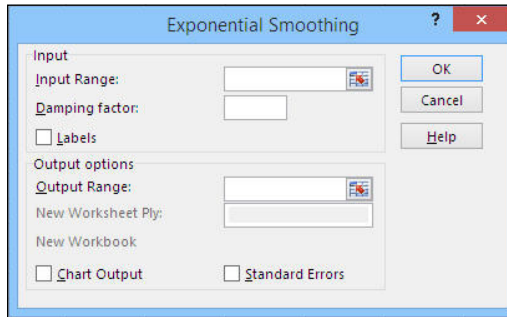
TIP

You can find step-by-step instructions for using the Exponential Smoothing dialog box in Chapter 15. As an overview, you start the whole process by taking these steps:

1. **Go to the Ribbon's Data tab.**
2. **Click Data Analysis.**

3. **Select Exponential Smoothing in the Analysis Tools list box.**
4. **Click OK.**

The Exponential Smoothing dialog box, shown in Figure 10-6, appears.



**FIGURE 10-6:** If you include a label, such as *Revenue*, at the top of the Input Range, select the Labels check box to tell Excel not to try to forecast from that label.

Here's some brief information on the controls shown in Figure 10-6:

- » For the Input Range, drag through the range where you have your baseline. It should include the actual results only, in one column only. In other words, if you have an adjacent column of dates as in Figure 10-5, *don't* include it.
- » The Damping Factor is 1.0 minus the smoothing constant. I know, that's irritating, but you have to make the best of it.
- » In contrast to other Data Analysis add-in tools, Exponential Smoothing requires that you put the output on the same worksheet as the Input Range.
- » If you want to chart the actuals and the forecasts, select the Chart Output check box. As shown in Chapter 15, though, you can do a better job of charting if you do it yourself.
- » Don't bother selecting the Standard Errors check box.

## Using the Regression Tool

Suppose you have one or more other variables in your baseline, along with your sales results, that you have reason to believe may be associated with those results. Neither moving averages nor exponential smoothing provides for using those other variables: Each of these approaches relies on using the forecast variable as its own predictor.

In contrast, an approach termed *regression* is designed to make use of these other variables in forecasting future values. The process of regression is not as intuitive as moving averages or exponential smoothing, and some people avoid it for this reason.

If you have additional data, at least give some thought to using it. Some things to think about:

- » **Sales results are often related to the time period in which the sales were made.** That is, as your baseline moves forward in time, the sales results may improve (with a maturing product) or decline (with a mature product). If the sales results are stable over time, knowledge of the date of sale won't help your forecast much.
- » **You need to know future values of the additional variables.** Time period is easy: You know what the next month and year are going to be. Finding out how many sales reps you'll have next month or next quarter, or how many dollars your company will put into advertising and other marketing programs, may be more difficult. If you can't get your hands on that future data in the form of plans or budgeted amounts, then you won't be able to use them to forecast sales.
- » **You need a longer baseline than is absolutely necessary with moving averages or exponential smoothing.** The very mathematics of regression requires that you have more time periods in your baseline than variables in the forecast equation. Some forecasters are willing to use regression with as few as 10 periods times the number of predictor variables — so, using 3 predictors, a baseline that's at least 30 periods long. I feel much more comfortable with a baseline that has 30 time periods for each predictor variable, but I admit that baselines that long can be difficult to come by. I also admit that there's no truly reliable rule of thumb here. Much depends on factors other than the length of the baseline, particularly the strength of the relationships between the variables.
- » **Regression does not work well with baselines that have sudden and prolonged changes in level.** Suppose your baseline chugs along at \$500,000 in monthly revenue for a couple of years and then (perhaps because of a new product launch) jumps to \$1,000,000 per month and stays at roughly that level. Regression has a difficult time dealing with that situation (it *can*, but the necessary management is beyond the scope I adopt here).

You can find much more detail on using regression for forecasting in Chapters 11 and 16. To get a feel for what you need to provide, a brief tour of the Data Analysis add-in's Regression tool is a reasonable place to start.

Suppose you have a worksheet laid out as in Figure 10-7. Here are two points to note about that layout:

- » There are three variables in the baseline — Revenue, which you want to forecast, and Date and Sales Reps, which you want to use to make the next forecast.
- » You can't see them all in the figure, but there are 40 time periods in the baseline. A baseline this long is probably, if barely, sufficient to support a regression analysis with two predictor variables.

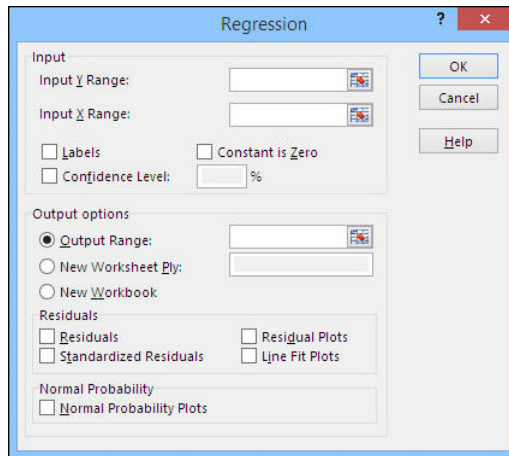
	A	B	C	D	E
	Sales				
1	Date	Reps	Revenue		
2	1/1/2014	11	\$ 461,753		
3	2/1/2014	10	\$ 421,771		
4	3/1/2014	12	\$ 441,563		
5	4/1/2014	13	\$ 460,760		
6	5/1/2014	10	\$ 435,918		
7	6/1/2014	12	\$ 490,642		
8	7/1/2014	10	\$ 461,047		
9	8/1/2014	11	\$ 419,476		
10	9/1/2014	8	\$ 488,784		
11	10/1/2014	12	\$ 479,800		
12	11/1/2014	11	\$ 465,455		
13	12/1/2014	10	\$ 397,685		
14	1/1/2015	12	\$ 404,979		
15	2/1/2015	10	\$ 440,276		
16	3/1/2015	11	\$ 444,571		
17	4/1/2015	11	\$ 469,068		
18	5/1/2015	12	\$ 411,992		
19	6/1/2015	10	\$ 408,022		
20	7/1/2015	9	\$ 494,317		
21	8/1/2015	8	\$ 486,399		
22	9/1/2015	11	\$ 434,797		
23	10/1/2015	8	\$ 499,630		
24	11/1/2015	10	\$ 440,602		

**FIGURE 10-7:**  
Be sure to put your predictor variables (here, Date and Sales Reps) in contiguous ranges.

To get things going, take these steps:

1. Go to the Ribbon's Data tab and click Data Analysis in the Analyze group.
2. Scroll down the Analysis Tools list box and select Regression.
3. Click OK.

The Regression dialog box, shown in Figure 10-8, appears.



**FIGURE 10-8:**  
The Regression  
dialog box.

**4. Supply the address of a worksheet range in the Input Y Range box.**

This range must occupy a single column. For the baseline in Figure 10-7, you'd enter **\$C\$1:\$C\$41**. (You can also just click in the Input Y Range box and drag through the range on the worksheet.)

**5. Supply the address of a worksheet range in the Input X Range box.**

This range contains your predictor variable(s). This range need not have just one column, although it could if you have just one predictor variable. It should have the same number of rows as the Input Y Range.

**6. If you included the variable labels at the top of each column in the Input Range boxes, select the Labels check box.**

**7. Do *not* select the Constant Is Zero check box.**

**8. If you add a confidence level, it shows up in the output *in addition* to the default 95%.**

If you later want to delete the level you added, you'll need to do so specifically in the Regression dialog box by clearing the check box.

**9. Choose an output option — that is, where you want the Regression tool's output to appear.**

**10. If you chose Output Range, *click in the box to its right* and fill in a cell address. If you chose New Workbook Ply, *click in the box to its right* and fill in the name for the new *ply* (an old term for a worksheet).**

**11. Click OK.**

Chapter 11 has information about the charts that you can create from this dialog box.

All you need to perform a regression forecast is a baseline with one forecast variable, such as sales revenue or units sold, and one or more predictor variables. The Data Analysis add-in refers to the forecast variable as the  $y$  variable, and you supply the worksheet range that has the baseline for the forecast variable in the Input Y Range box.

The variable that you want to use as a predictor is the  $x$  variable, and its worksheet address goes into the Input X Range box. Keep in mind that you can use more than one predictor variable, but if you do, the columns that the predictors occupy must be contiguous: They should be in adjacent columns, and their first and last values should be in the same rows. All variables (forecast and predictors) must have the same number of baseline periods.



REMEMBER

This usage — referring to the forecast variable as  $Y$  — is, I'm happy to report, consistent with the syntax of the `LINEST` worksheet function. The `LINEST` function refers to known  $Y$  values as the values for the forecast variable, and to known  $X$  values as the values to use for the predictor variable(s). Bearing this in mind can be helpful when you read Chapters 12 and 16.



Knowing whether to use the  
Regression tool

Getting familiar with the  
Regression tool

Taking the Regression approach

## Chapter 11

# Basing Forecasts on Regression

**R**egression is a standard technique in forecasting, whether sales revenues or sunspots. (And yes, meteorologists and astronomers have used regression for years in forecasting sunspots.)

This chapter introduces forecasting with regression. The idea is to get your hands on one variable (say, the price you charge for your product) that is strongly related to another variable (say, your unit sales), and then use what you know about the first variable to forecast what will happen to the second variable. Of course, “what you know about the first variable” ranges from the obvious (“Next month is April”) to the mysterious (“Is the Fed about to raise interest rates?”). Generally, you’d prefer to avoid a situation in which you find yourself forecasting your predictor variable.

That’s simple regression: One variable forecasts another. You can also make use of multiple regression, where you use more than one variable to forecast another. A typical example is to use both product price and the index of consumer confidence to forecast sales. Within limits, you can use as many predictors as you can lay your hands on — often (again, within limits) the more predictors you use, the more accurate your forecast. It’s important to keep in mind that the more predictor variables you have, the more records you need. A regression equation based on too many variables and too few records becomes unstable.

# Deciding to Use the Regression Tool

Regression sounds a lot more complicated than it really is — one of the problems is that the very word *regression* sounds intimidating. Truth to tell, regression *can* be intimidating, but more often than not it's really pretty straightforward.

Let's get a basis. You're taller now than you were when you were 5 years old. Up to a point, you can forecast a person's height if you know how old the person is. I say "up to a point" because, eventually, people quit getting taller, but, alas, not older.

Here's an example. Up to the age of 18 or so, you can predict with reasonable accuracy how tall a person is, given that you know the person's age. Even if you don't know whether the person is male or female, this equation is a pretty fair guide:

$$(0.14 * \text{Age In Months}) + 38$$

In words, multiply someone's age in months times 0.14, and add 38. This will give you a decent approximation of that person's height in inches. At least until the person reaches 18 years of age.

This is what regression is all about. It's a way to develop an equation that predicts one variable (height, poppy-seed sales, sunspots, whatever) from another variable (age, poppy-seed advertising, calendar year, whatever).

The idea is to use not just one but two variables. If you were interested in forecasting height from age, you would get hold of both the height and the age of a bunch of kids. Both the heights and ages would be, in a real sense, baselines. You'd then know how tall a child was at a given age. Figure 11-1 shows a couple of baselines. Figure 11-2 shows the chart that results from the data in the baselines.

Now, here's what Excel can do for you if you let it. If you haven't already installed Excel's Data Analysis add-in, turn to Chapter 10 for step-by-step instructions.

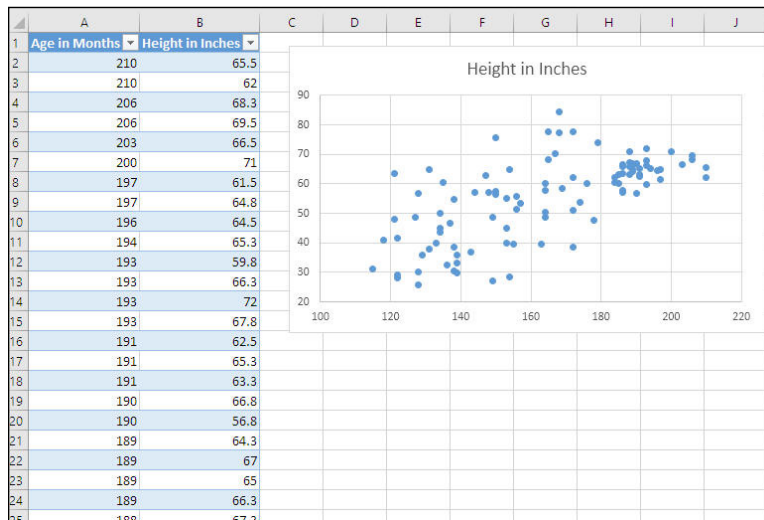
After you have that add-in installed, you have a new item, Data Analysis, in the Analyze group of the Ribbon's Data tab. Go to that tab, click Data Analysis, and scroll down the Analysis Tools list box until you see Regression. Select Regression and click OK. You'll see the Regression dialog box shown in Figure 11-3.

The Regression tool gives you certain standard options for developing a forecast based on regression. You have somewhat less control over what's going on than if you entered the formulas yourself, but using the tool is certainly easier.

**FIGURE 11-1:**  
Judging the relationship between age and size just by looking at the numbers is difficult.

	A	B	C	D	E	F	G	H
1	Age in Months	Height in Inches						
2		210	65.5					
3		210	62					
4		206	68.3					
5		206	69.5					
6		203	66.5					
7		200	71					
8		197	61.5					
9		197	64.8					
10		196	64.5					
11		194	65.3					
12		193	59.8					
13		193	66.3					
14		193	72					
15		193	67.8					
16		191	62.5					
17		191	65.3					
18		191	63.3					
19		190	66.8					
20		190	56.8					
21		189	64.3					
22		189	67					
23		189	65					
24		189	66.3					
25		188	67.2					

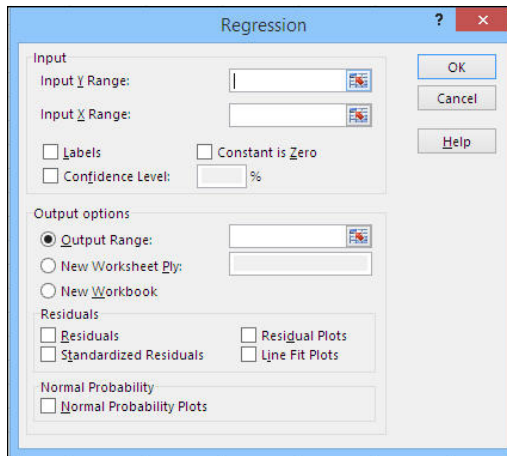
**FIGURE 11-2:**  
After you've got the relationship charted, telling what's going on is much easier.



## Adopting the Regression approach

When you're deciding whether to use regression to forecast sales, you want to keep a few things in mind.

**FIGURE 11-3:** The Regression tool helps you develop a forecast without having to enter formulas.



## Using related variables

In the preceding section, I use a couple of variables — age and height — that are related up to the point that people stop growing. Before you consider using regression to forecast sales, you want to be sure you have one or more variables that are related to sales levels.

A good place to start is with sales drivers, such as dollars that your company spends on advertising or number of sales representatives that your company employs. If you can lay your hands on historical information for, say, advertising dollars and sales dollars, you're ready to see whether a dependable relationship exists.

You do need to be able to pair up individual values on the variables. In this example, you need to know advertising and sales dollars for January, advertising and sales dollars for February, and so on. Having an uninterrupted baseline of values is necessary, so if you have one or two missing months (or quarters or years, depending on how you're building your baselines), you might need to estimate a couple of missing values. As in many forecasting situations, a good, long baseline is your best protection against being misled by your estimate of a missing value.

Later in this chapter, in “Understanding the Data Analysis Add-in’s Regression Tool,” I show you how you can evaluate the relationship between the variables.

## Using a variable you can predict

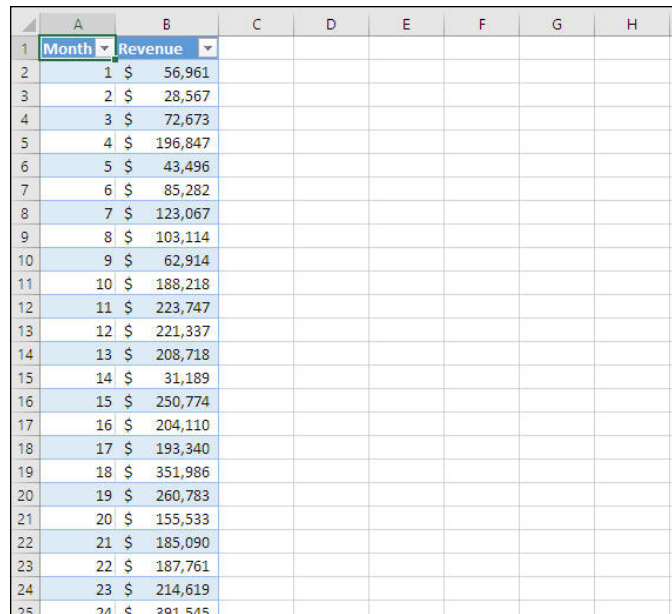
When you're preparing a sales forecast based on regression, you need to choose a variable that is itself predictable. For example, suppose that you use advertising dollars to forecast sales. From looking at your baselines, you might see that a strong relationship exists between advertising and sales. And by using the Data

Analysis add-in's Regression tool on the baselines, you can get an equation that takes in advertising dollars and puts out sales forecasts.

But if you don't know how much your company is going to spend on advertising during the next time period, you can't plug that value into the equation, and you can't make a forecast. Before you spend a lot of time and energy collecting baseline data and analyzing it, make sure you're using a variable with a future you know about. In this example, you might have information from Product Marketing about how much it intends to commit to its next ad buy.

## Using time periods

If your sales data show a trend over time periods, you can use those periods themselves to forecast. Figure 11-4 shows an example.



	A	B	C	D	E	F	G	H
1	Month	Revenue						
2	1	\$ 56,961						
3	2	\$ 28,567						
4	3	\$ 72,673						
5	4	\$ 196,847						
6	5	\$ 43,496						
7	6	\$ 85,282						
8	7	\$ 123,067						
9	8	\$ 103,114						
10	9	\$ 62,914						
11	10	\$ 188,218						
12	11	\$ 223,747						
13	12	\$ 221,337						
14	13	\$ 208,718						
15	14	\$ 31,189						
16	15	\$ 250,774						
17	16	\$ 204,110						
18	17	\$ 193,340						
19	18	\$ 351,986						
20	19	\$ 260,783						
21	20	\$ 155,533						
22	21	\$ 185,090						
23	22	\$ 187,761						
24	23	\$ 214,619						
25	24	\$ 291,545						

**FIGURE 11-4:** You know that another time period is going to come along, so you can use that to get your next sales forecast.

You don't need to use actual date values to create the forecast, although you can if you want. It doesn't really matter if you use, say, 1998, 1999, 2000, . . . 2005 to forecast from, or 1, 2, 3, . . . 8.

## Using more than one predictor variable

So far, I've given you a look at what's called *simple regression*: using the relationship between one variable and another to forecast one of them. You can also use

two or more predictor variables at once, and the Data Analysis add-in's Regression tool can manage that situation.

Suppose you're able to get your hands on three baselines: monthly sales dollars, monthly advertising dollars, and the number of sales reps working for you during each month. The Regression tool might report back that the equation you should use to forecast a month's sales is:

$$\text{Sales} = (25 \times \text{Advertising Dollars}) + (802 \times \text{Sales Reps}) + 37,268$$

Here's where things can start to get a little sticky. The more predictor variables you use, the more important it is to have a lengthy baseline. Otherwise, your forecasts are going to jump all over the place like popcorn. And if the number of variables you're using even comes close to the number of records in the baseline, the forecasts will become completely unreliable.

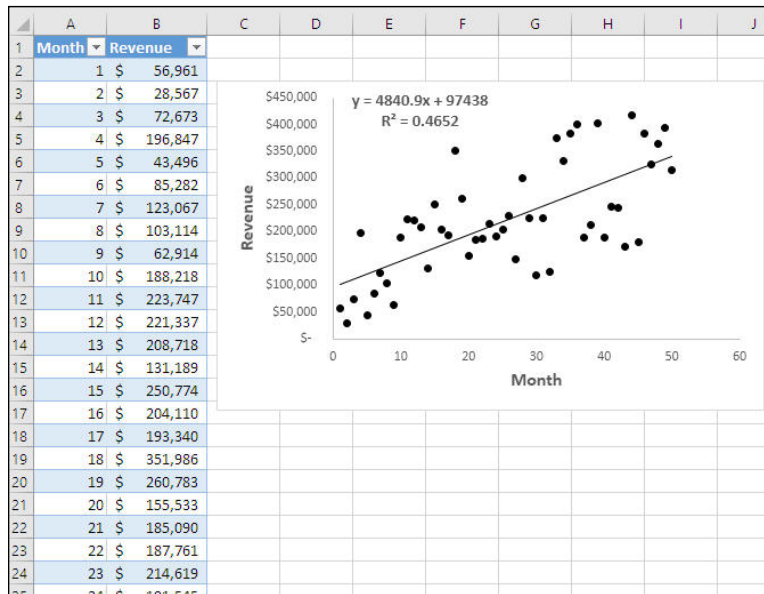
This is why I like to forecast using one month as the time period in my baseline. A month is often a useful period to forecast into ("How many widgets are we going to sell next month?"), and all you need is 4 years' worth of data to have a baseline almost 50 periods long. Just choose your predictor variables well, so you're forecasting sales using advertising or sales-rep counts or the time periods themselves, rather than a goat's entrails. When you use variables that correlate well with sales, a baseline of 50 will often do just fine.

## Understanding the Data Analysis Add-in's Regression Tool

In this section, I show you how to use the regression method to forecast sales. Figure 11-5 shows a baseline of data. Charting the data first, as shown in Figure 11-5, helps you decide whether to continue with the analysis. If you see something like a diagonal line, continuing probably makes sense. If you see something more like a circle, probably no relationship exists and you may as well find a different variable as your predictor.

With the data laid out as an Excel table, you can use the Data Analysis add-in's Regression tool to help you get a forecast. Take these steps:

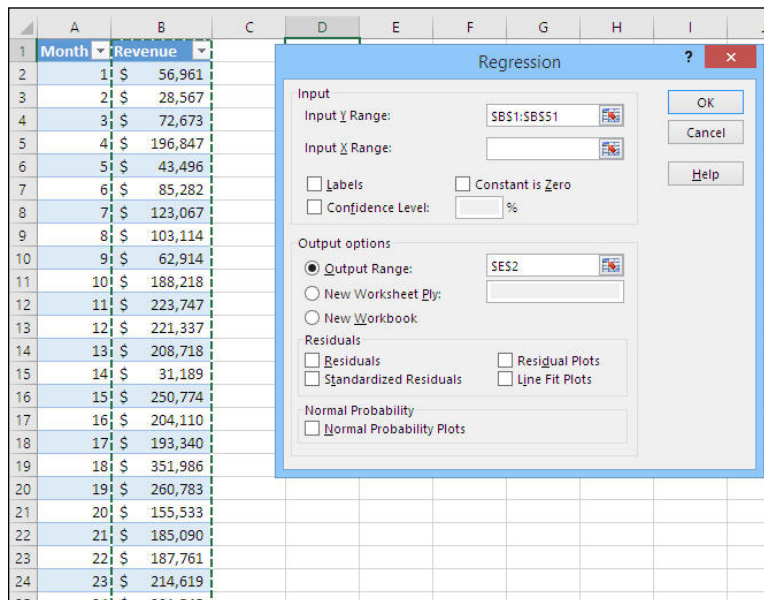
1. **Go to the Ribbon's Data tab.**
2. **Click Data Analysis in the Analyze group.**



**FIGURE 11-5:**  
This baseline is used to forecast sales on the basis of time period.

**3. Click the Regression tool in the list box and click OK.**

The Regression tool's dialog box appears along with your data (see Figure 11-6).



**FIGURE 11-6:**  
In the Regression dialog box, a *worksheet ply* is just an old term for a worksheet.

4. **Click in the Input Y Range box and drag through the cells that contain the values you want to forecast.**

Here, that's Revenues. Figure 11-6 doesn't show it all, but that range extends from B1:B51.



TECHNICAL  
STUFF

In regression and, in fact, in most forecast methods, you often find the variable that's being forecast termed the *y* variable. The variables that are being used as predictors are often termed the *x* variables.

5. **Click in the Input X Range box and drag through the cells that contain the values you want to forecast *from*.**

Because you're forecasting using the Month number, drag through A1:A51.

A quick way to select a range of values is to use your mouse to select all the cells in the top row, and then simultaneously press Ctrl+Shift+Down Arrow. Excel selects the range for you, but it stops when it runs into an empty cell — another good reason to avoid missing data.



TIP

6. **Because you included the variable labels in row 1 as part of the input ranges, select the Labels check box.**

Otherwise, Excel will find nonnumeric data in the input ranges and will throw a warning message at you.

7. **If necessary, select the Output Range option button.**

When you select any of the option buttons in the Output Option frame, Excel perversely makes the Input Y Range box active. If you're not watching what you're doing, you can easily overwrite the range address you entered as the Input Y Range with the address you want to use for your output. Not a huge deal, but it can be really annoying.



WARNING

8. **When you're sure you have the Output Range box selected, click a cell where you want the output to start, and then click OK.**

Excel creates the regression information shown in Figure 11-7.

I know that output looks pretty intimidating, but you can ignore most of it if you want. Here are the things you want to pay attention to:

- » **The number that's labeled *Multiple R*:** That's in cell E4 in Figure 11-7. The closer it is to 1.0, the stronger the relationship between the predictor variable (or variables) and the variable you want to forecast — and therefore the more confidence you can have in the forecast. The closer it is to zero, the worse the relationship and the less confidence you have. It can run only from 0.0 to 1.0, and the value of 0.571 shown in Figure 11-7 is reasonably good. In simple regression, the Multiple R is really just the correlation between the predictor



and the forecast variable (although in contrast to a simple correlation, a Multiple R is always positive).

» **The two numbers labeled Intercept and Month:** They're in cells E17 and E18, respectively. You use them in your forecast equation. (The one in E18 is labeled *Month* only because that's the name used for that variable in the list, in cell A1.)

	A	B	C	D	E	F	G	H	I	J
1	Month	Revenue		SUMMARY OUTPUT						
2	1	\$ 56,961								
3	2	\$ 28,567		<i>Regression Statistics</i>						
4	3	\$ 72,673		Multiple R	0.571					
5	4	\$ 196,847		R Square	0.326					
6	5	\$ 43,496		Adjusted R Square	0.312					
7	6	\$ 85,282		Standard Error	91901.258					
8	7	\$ 123,067		Observations	50					
9	8	\$ 103,114								
10	9	\$ 62,914		ANOVA						
11	10	\$ 188,218			<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	11	\$ 223,747		Regression	1	196406644310	1.96E+11	23.25484	0.00001	
13	12	\$ 221,337		Residual	48	405400377457	8.45E+09			
14	13	\$ 208,718		Total	49	601807021766				
15	14	\$ 31,189								
16	15	\$ 250,774			<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	16	\$ 204,110		Intercept	109615.238	26388.46317	4.15391	0.00013	56558	162673
18	17	\$ 193,340		Month	4343.108	900.62490	4.82233	0.00001	2532	6154
19	18	\$ 351,986								
20	19	\$ 260,783								
21	20	\$ 155,533								

**FIGURE 11-7:** There are only a few numbers shown here that you need to make your forecast.

So, using the information you get from the Regression tool, you can round the numbers a little and write this equation:

$$\text{Revenue} = (4,343 \times \text{Month}) + 109,615$$

In words, the equation says that, given these baselines, the best estimate of Revenue comes by multiplying the month number (here, 1 through 50) by 4,343 and adding 109,615.

And if you want to forecast revenue for month 51, you'd use this formula in a worksheet cell:

$$= (4343 * 51) + 109615$$

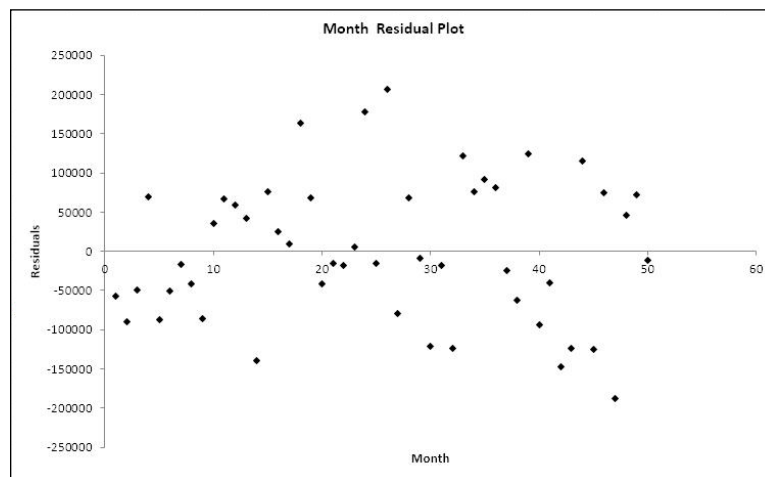
which equals 331,108 — your regression-based forecast for month 51.

## Checking the forecast errors

As I keep complaining in this book, when you forecast you make errors. And regression is no different from any other approach to forecasting when it comes to errors.

Regression does use a different term, though: *residuals*. In this sense, residuals are errors. Have another glance at the formula at the end of the preceding section. Instead of putting 51 in the formula, you could put in any month number from 1 to 50, and the formula would predict the revenue for that particular month.

You could then subtract that predicted revenue amount from the actual revenue for that month — the result is a residual. Having the Regression tool chart the residuals against the month's number, shown in Figure 11-8, is useful.



**FIGURE 11-8:**  
When the Regression tool creates this chart, it embeds it in the worksheet and puts it to the right of the numeric output.

You get this chart by selecting the Residual Plots check box in the Regression dialog box (refer to Figure 11-6).

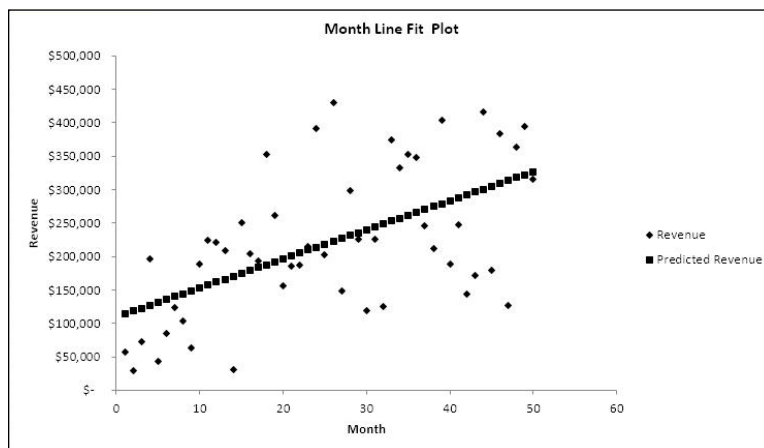
The pattern of the markers in the chart is important. You'd like to see them form a shape similar to a rectangle, paralleling the chart's horizontal axis. In Figure 11-8, it's close, but I'd judge that it's close enough to a rectangle — even though the markers tend to be closer together (vertically) on the left, in the early months, than on the right, in the later months. Sometimes you also see a shallow U shape or a rectangle that's tilted from the horizontal axis.

When you see that sort of non-random distribution of the residuals in a chart, it can mean that there's some kind of dependable pattern to the data that your

forecast hasn't accounted for. In that case, you might want to add a predictor variable to the analysis. Suppose you sell plywood in a region that suddenly starts getting knocked around by an El Niño storm. After that sort of event becomes part of the long-term weather pattern, you're going to sell a lot of plywood to cover windows in some years, and in other years you'll sell a lot less than your forecast estimated. If you don't account for that effect in your forecast model, by adding a variable that's sensitive to the occurrence of storms, you're likely to see more variation after the pattern has changed than you did before. Your forecasts will tend to be more accurate if you can account for that source of variation.

## Plotting your actual revenues

Another very useful chart that the Regression tool will create for you is the Line Fit Plot. Using the current example, it shows your actual sales revenues and the predicted revenues, charted against Month. You can see one in Figure 11-9.

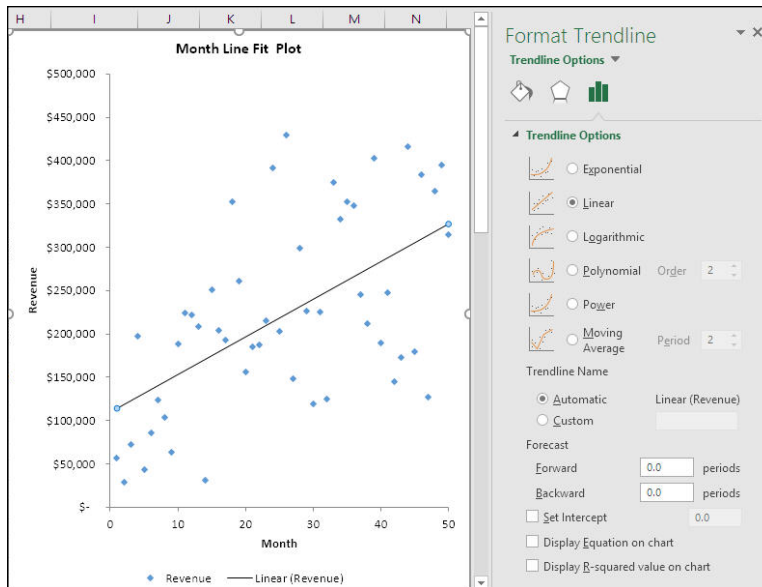


**FIGURE 11-9:** The line showing the predicted revenues is the result of applying the regression equation to each actual month number.

The straight line made of dots in Figure 11-9 is called a *regression line* or *trendline*. As the Regression tool creates it, it's not very useful. I like to click on it to select it and then press Delete. Since Microsoft started including the Data Analysis add-in with Excel, it has improved Excel's charting tools. Right-click any marker in the chart that represents a revenue value in your baseline. Then choose Add Trendline from the shortcut menu. You'll see the Format Trendline pane at the right of the Excel window, as shown in Figure 11-10.

Scroll down to the bottom of the Format Trendline pane if necessary. You can call for Excel to extend the trendline forward, into the future, or backward, into the past, by the number of periods that you specify in the edit boxes.

**FIGURE 11-10:**  
You'll usually want to choose the Linear Trendline.



Still in Figure 11-10, notice that you can also call for the chart to display the regression equation and the  $R^2$  value.  $R^2$  is the square of the multiple R that you met earlier in this chapter, and it represents the percent of variability that the predicted variable shares with the predictor variable (or variables).  $R^2$  is an important diagnostic statistic.

The trendline is much more useful for forecasting than the line of predicted values, created by the Regression tool, that I deleted near the start of this section.

## Understanding confidence levels

Confidence levels are closely related to probabilities expressed as percentages. At the right of Figure 11-7, you see a couple of columns with numbers that are labeled *Lower 95%* and *Upper 95%*. Notice that the Month predictor's Lower is 2532 and its Upper is 6154. Those two figures together create a *confidence interval*. We are 95 percent confident that the Month coefficient, if calculated on the full population from which we took this sample baseline, is between 2532 and 6154.

In forecasting, we often get samples of data for baselines. You might think that your sample constitutes a long enough baseline, but the picture might change drastically if you'd gone back another year. And the regression results — everything you see in the Regression tool's numbers — would also change. The question is how much it would change.

Suppose you had 100 different samples from the same population of data, each somewhat different. Here's what the 95 percent confidence figures mean: Using the analysis shown in Figure 11-7, we believe that 95 of those 100 samples would result in a coefficient for the Month variable between 2532 and 6154.

Of course you'd never do that — collect 100 different samples and calculate a Month coefficient for each of them — but you can tell by looking at the upper 95% and lower 95% figures how precisely the combination of the baseline and the regression analysis has estimated the coefficients you'll use to make your forecast.

Why is it important to know that? Suppose that the Month coefficient, if it were calculated using the entire relevant population of revenues, were 0.0. In that case, the month in which the revenues occurred has no effect on the amount of the revenue. In the regression equation, you multiply the Month by its coefficient. With a coefficient of 0.0, nothing gets added to or subtracted from the forecast revenue by virtue of the month in which the revenue occurred. You might as well leave Month out of the equation.

But in this example the 95 percent confidence interval runs from 2352 to 6154. The interval does *not* span 0.0. Therefore you can conclude with 95 percent confidence that the Month coefficient is not 0.0 and you can leave it in the regression equation. Keeping Month as a predictor may not add much accuracy to the forecasts, but you can test for that. And in the meantime, the fact that the confidence interval does not span 0.0 tells you that Month adds at least *some* accuracy to the equation.

There's nothing magic about a 95 percent confidence level. It's just a traditional level of probability that researchers have used for decades. If you select the Confidence Level check box in the Regression tool's dialog box, its associated box becomes enabled and you can enter another confidence level. That confidence level is included in the output along with the default 95 percent. In general, a lower confidence level such as 90 percent results in a narrower interval between the lower and upper figures. That means you can estimate the actual value of the coefficient more narrowly, but with less confidence in your conclusion.

## Avoiding a zero constant

The Regression tool's dialog box has a check box labeled Constant Is Zero. The term *constant* is just another term for the intercept that the Regression tool includes in its output. In the equation shown at the end of the "Understanding the Data Analysis Add-in's Regression Tool" section, earlier in this chapter, it's the value 109615. It's called an intercept because if you extended the regression line to the vertical axis, the line would intercept that axis at 109615.

Some people who use regression know that, in reality, the intercept based on their full data set is zero, and if the Regression tool calculates a different value it's because of sampling error. So they select the Constant Is Zero check box.

In sales forecasting, that's almost certainly not the case, unless your company didn't sell anything during the first time period. The problem is that by setting the constant to zero, because of the math that underlies regression, you can wind up getting really screwy results. For just one example, the Multiple R, which you look at to judge the strength of the relationship between what you're forecasting and your predictors, can get seriously inflated and you'll be misled about how accurate your forecasts are.



TIP

In general you're much better off leaving the Constant Is Zero check box unchecked. If the constant in the population is really zero, a decent sample will result in an estimate of the constant that's close to zero anyway.

## Using Multiple Regression

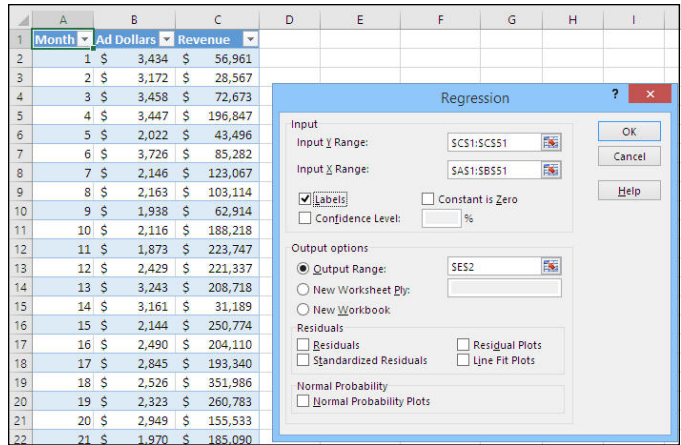
If you have two or more different predictor variables, you can often improve your forecasts. It's called *multiple regression* because you use multiple predictor variables. This chapter gives a light once-over to that approach in the "Using more than one predictor variable" section. In this section, I show you how to use multiple regression to forecast sales using month and advertising dollars.

Suppose that, in addition to sales revenues and month number, you have advertising expenses (this is mildly unrealistic because ad expenses are often leading indicators of sales, but not always).

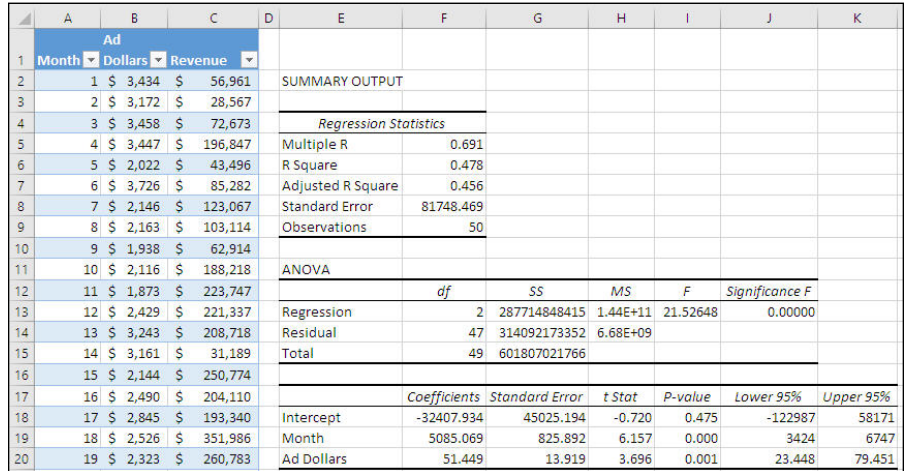
The data in Figure 11-11 is identical to earlier figures in this chapter, except that an additional column, Ad Dollars, has been added to the list and is referenced in the Regression dialog box Input X values. Running the Data Analysis add-in's Regression tool on this data requires only a slight change to the Regression dialog box so that the Input X values are found in A1:B51 (see Figure 11-12 for the resulting output).

What characterizes a good predictor to add to an existing equation? There are two main characteristics for a good added predictor, one having to do with the relationship between the new predictor and the forecast variable, and the other with the relationship between the new predictor and the existing predictors.

**FIGURE 11-11:**  
Make sure your predictor variables occupy adjacent columns. Don't lay it out as, say, Month, Sales Revenues, Ad Dollars.



**FIGURE 11-12:**  
The basic layout of the output is the same regardless of the number of predictor variables.



## New predictor with forecast variable

A new predictor, such as Ad Dollars in this example, should bear some relationship to the value that you want to forecast — here, Sales Revenues. The correlation in this example between Ad Dollars and Sales Revenues is a small one: only 0.24. (You can check that if you want by using the CORREL() function on the worksheet for Figure 11-11 or 11-12.) That's right on the cusp between what most people would regard as a weak correlation and a moderate one, and you might not think it worth including in the regression equation.

The Multiple R value in Figure 11-7 is 0.571. But look at the Multiple R in Figure 11-12: It's bounced all the way up to 0.691. That's a respectable bounce for a predictor that correlates only 0.24 with the forecast variable. The reason lies in the next section.

## New predictor with existing variable

What has happened is that Month, as measured by the month's number in the baseline, has already accounted for much of the Sales Revenues. What's left are the residuals in the Sales Revenues, discussed in the section of this chapter titled "Checking the forecast errors."

It turns out in this case (and in many others) that the added variable — here, Ad Dollars — correlates well with the *residuals*. That means there's some systematic variation left in the Revenues after accounting for Month. Including Ad Dollars as a predictor accounts for some of that leftover variation and moves it out of the Residual category and into an Explained category. So the two predictors together can explain more about the behavior of the forecast variable than the single predictor can by itself, and your multiple R gets more of a bounce than a presidential candidate at a nominating convention.

And as you've seen, the higher the multiple R, the better your forecast. All other things being equal, of course. But all other things are seldom equal, and it wouldn't hurt to read some more about forecasting with regression in Chapters 12 and 16.



TIP

By the way, you're not limited mathematically to any particular number of predictor variables, as long as you have enough observations (or a long enough baseline) to handle them. As a matter of practical fact, though, Excel limits you to 64 predictor variables, no matter how long your baseline is.



TECHNICAL  
STUFF

You're not limited to numeric predictors, either. For example, you could add sales region and product line as predictors. But to do so, you have to convert names such as Northwest Region and Dee's Archery into numeric codes. There are special ways to do this, sometimes termed *dummy coding*. I don't go into dummy coding in this book — there aren't enough pages, and you can find several more-advanced books specifically on multiple regression that cover dummy coding. Just keep in mind that it's both possible and feasible.



# **4** **Making Advanced Forecasts**

### **IN THIS PART . . .**

Here you find out how to use Excel to make more-advanced forecasts. By that, I mean forecasts that you can control more closely. For example, although the Data Analysis add-in is useful, you're largely resigned to what it tells you. But if you know how to enter the formulas yourself, you have more control over what's going on. In this part, I also show you how to account for seasonality when you forecast your sales.

## Chapter 12

# Entering the Formulas Yourself

**T**his chapter starts with a brief review — or overview — of the rules for Excel formulas. As you get more and more comfortable with quantitative forecasting, you're likely to find yourself wanting to rely less on tools such as the Data Analysis add-in, and more on writing your own formulas.

Excel has hundreds of prefabricated formulas, called *functions*. A good example is the AVERAGE function. All you need to do is refer to the function in a worksheet cell and point it at a range of cells with numbers in them. The AVERAGE function adds up the values in that range of cells and divides by the number of values.

You get more information about Excel functions that are useful in forecasting in this chapter. Besides the AVERAGE function, there are some functions that are important in forecasts that use regression.

Some functions require that you array-enter them if they're to return the results you're after. This chapter goes into some detail about how to array-enter a formula, whether or not the formula contains a function.

# About Excel Formulas

Formulas are at the heart of Excel, which makes it difficult to understand why 80 percent of Excel worksheets contain no formulas, only static values like *14* or *Smith* (that's what a market research study determined a few years back). Formulas, together with functions, help you summarize data with totals, analyze data with averages, find data, rearrange data, transform data — the list of things you can do with worksheet formulas is a long one.

This section recommends that, sometimes anyway, you should consider using a formula that you create yourself rather than relying on the Data Analysis add-in. But it's not an introduction to formulas. If you're comfortable with entering a formula, by all means continue. If you harbor a suspicion that I'm talking about liquid food for infants, you may want to look through *Excel 2016 For Dummies* by Greg Harvey (published by Wiley).

## Doing it yourself: Why bother?

As is probably apparent by now, this book is about quantitative forecasting, and particularly quantitative sales forecasting, using Excel. Chapters 6 and 7 discuss setting up your baseline data in an Excel worksheet. With the data set up as an Excel table, you can deploy the Data Analysis add-in and get a forecast using the Moving Average, Exponential Smoothing, or Regression tool.

So, why would you bother with entering the formulas yourself? Well, if you use formulas, you have more control over what's going on in the development of the forecast.

For example, suppose you use the Data Analysis add-in's Moving Average tool to develop a moving-average forecast. Although the moving averages that the tool calculates are formulas, those formulas use a constant: the length of the subset of the baseline values that are used by a given moving average — 3, for example, as in this formula:

```
=AVERAGE(A2:A4)
```

But what if you wanted to change the number of baseline values that go into each moving average — for example:

```
=AVERAGE(A2:A5)
```

where the moving average is of length 4. In that case, you'd have to go through each moving-average formula and adjust it, or you'd have to rerun the Moving Average tool. And then what if you wanted to go back to length 3?

If you wrote your own formula, you could use something like this instead:

```
=AVERAGE(OFFSET($A2,0,0,$C$1,1))
```



TIP

When you see dollar signs in a cell or range reference, as in the preceding formula, you know that it's an *absolute* reference. When you copy a formula that has no dollar signs into a different row or column, the cell references adjust to take account of where you copied it to. But when you use dollar signs to make a formula's cell references *absolute*, you can copy and paste it anywhere on the worksheet and the cell references will *not* adjust.  $\$C\$1$  is an absolute reference.  $\$A2$  is a mixed reference: It will always point to column A, because of the dollar sign before the column reference.  $A\$2$ , another mixed reference, will always point to row 2.  $A2$  is a relative reference, and both its row and its column will adjust, depending on where you copy and paste it to.

What's all this good for? Meet the OFFSET function. In the following example, it returns a range of cells — specifically, the range that starts in cell  $\$A2$  has as many rows as whatever the number is in cell  $\$C\$1$ , and is one column wide.

Suppose that  $\$C\$1$  contains the number 3. In that case, this formula:

```
=AVERAGE(OFFSET($A2,0,0,$C$1,1))
```

if entered in row 2, returns the average value in the range of cells that you want to average: That range is in column A, starts in row 2, has as many rows as the number in cell  $\$C\$1$ , and has one column. Because  $\$A2$  is a mixed reference, you can copy the formula down into row 3, row 4 and so on, and the basis cell changes to  $\$A3$ ,  $\$A4$ , and so on. (But because the column is fixed, the basis cell  $\$A2$  would not change if you copied it into a different column.)

This approach is useful because you can change the number of cells in each average by simply changing the number in cell C1.

Here I walk you through the formula, repeated here for convenience, and assuming that it's been entered in cell B2:

```
=AVERAGE(OFFSET($A2,0,0,$C$1,1))
```

- » It refers to cell  $\$A2$ . That's the basis cell, the one that anchors the range of cells that we're after so we can average them. In this example, it would be smart for  $\$A2$  to be the first value in the baseline.

» It makes \$A2 a mixed reference. The dollar sign anchors it to column A: You can copy the formula to, say, cell Q5 and it will still use a cell in column A (specifically, it will use A5 as a basis cell). But because no dollar sign is anchoring the row, if you copy and paste the formula from B2 to, say, F20, the reference to \$A2 will change to \$A20.

» The two zeros used in the function's arguments mean that the range that OFFSET returns begins in \$A2. That is, the range will begin zero rows and zero columns away from \$A2. If the formula instead were this one:

```
=AVERAGE(OFFSET($A2,1,0,$C$1,1))
```

then the range that OFFSET returns would begin one row below \$A2, and

```
=AVERAGE(OFFSET($A2,0,1,$C$1,1))
```

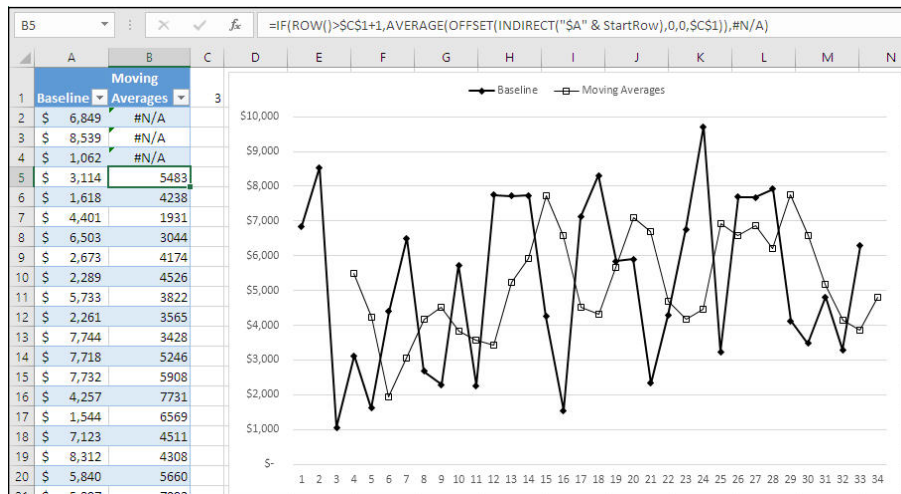
would cause OFFSET to return a range that begins one column right of \$A2 — that is, B2.

» Whatever numeric value you enter in \$C\$1 controls how many rows are in the range that OFFSET returns.

» The range will be one column wide.

» By surrounding the OFFSET function with the AVERAGE function, you get the average of the values in the range that OFFSET returns.

So, if the number 4 were in \$C\$1, the OFFSET function would return the range A2:A5. Figure 12-1 shows how this works in practice. (Bear with me: I get into what INDIRECT is, and how StartRow is used, very shortly.)

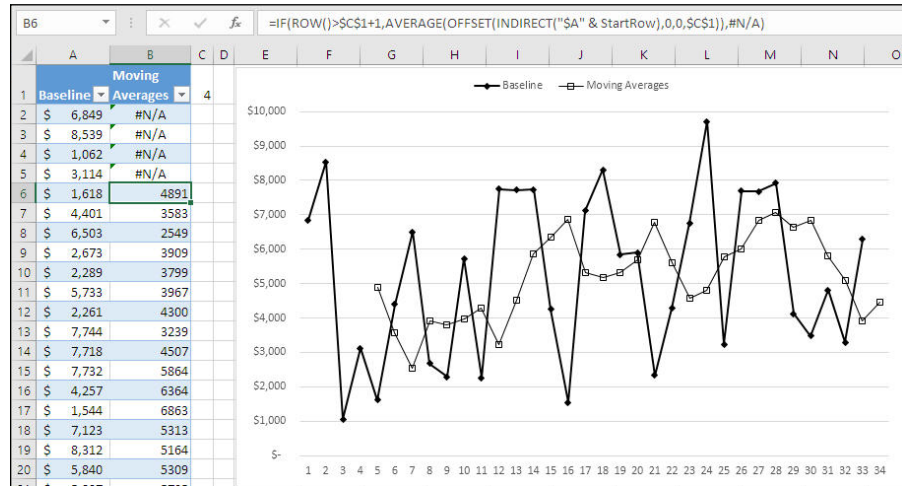


**FIGURE 12-1:** This is one way to change the length of the baseline section that's used by each moving average.

In Figure 12-1, you can tell the number of values that are averaged to create each moving average by looking at cell C1, which contains 3. Three values are combined into each moving average.

If you want to get a look at the moving-average series with a moving-average length of 4, all you do is change the value in C1 from 3 to 4 (see Figure 12-2).

**FIGURE 12-2:**  
Changing just one number, in C1, is a lot faster and easier than rerunning the Moving Average tool.



Although the basic idea is the same, the formula actually used in column B of Figures 12-1 and 12-2 is somewhat more complicated than the one we've been discussing. That's because we want to make the starting point of the moving-average series in the chart depend on the length we want to use for the moving average. Here's the formula:

```
=IF(ROW()>$C$1+1, AVERAGE(OFFSET(INDIRECT("$A"&StartRow),0,0,$C$1,1)),#N/A)
```

It's best to take something like this in small doses, and to work inside out:

» There's a defined name in the workbook, StartRow. It's defined as:

```
=ROW()-$C$1
```

Wherever you enter the name StartRow, it will return the number of the row where you entered it minus the value in cell C1. This tells Excel how far up to look to find the baseline's starting value for the moving average. If 3 is in C1 and you enter the formula in B15, then it returns 15 - 3 = 12. The current moving average should average baseline values beginning in row 12 — that's the StartRow.

- » In Excel, if you define a name such as `StartRow` to refer to a formula, as here, you can use the name rather than the formula, as in this segment of the complete formula:

```
INDIRECT(“$A”&StartRow)
```

Let’s continue to assume that `StartRow` returns 12, given the value in C1 and that you enter this formula on row 15. Then using the `INDIRECT` function results in a cell address — in this case, that’s A12. (You define a name in Excel by going to the Ribbon’s Formulas tab and clicking Define Name in the Defined Names group. Type the name in the Name box, and the formula, or a cell address, in the Refers To box.) At this point, the `INDIRECT` fragment just given resolves to the cell address A12.

- » Expanding and simplifying the segment a little further, you’ve got this segment of the formula:

```
OFFSET(A12,0,0,$C$1,1)
```

In other words, return the range that starts zero rows and zero columns from cell A12. Give that range as many rows as the number in `$C$1` — which this example assumes is 3 — and one column.

- » Get the average of the values in that range:

```
AVERAGE(OFFSET(A12,0,0,3,1))
```

or:

```
AVERAGE(A12:A14)
```

- » Finally, the full formula as entered on row 15 (taking into account the simplifications you’ve already made:

```
=IF(ROW()>$C$1+1,AVERAGE(A12:A14),#N/A)
```

and simplifying further:

```
=IF(ROW()>4,AVERAGE(A12:A14),#N/A)
```

In words, you’re telling the formula to look at the number of the row where you’re entering the formula. If that number is 15 (and therefore greater than 4), return the average of the values in A12:A14 — that is, the current moving average. Otherwise, return the Excel error value `#N/A`.

A couple of issues are wrapped up in that formula. One is that you’ll get a `#REF!` error if you put this formula in rows 1, 2, 3, or 4, because it will start referring to cells A0, A-1, and so on (these last references are just for illustration — there’s no such thing as row zero or row -1. Excel would actually show you the `#REF!` error value). Instead, you use the `IF` function to make sure that the formula hasn’t been entered too far up on the worksheet. If it has, return `#N/A`.



The other issue is the use of #N/A. That keeps those cells out of the chart and ensures that the series of moving-average values are in the correct rows vis-à-vis the baseline. See Chapter 13 for information about how the baseline values and the forecast values should line up so that your forecast will be for the correct period, and so that the baseline and forecast will chart correctly.

There are at least three other reasons to consider entering your own forecasting formulas rather than relying on the Data Analysis add-in.

## Until they get it right

Some Data Analysis add-in tools don't get the analysis right, and the Moving Average tool is one of them. Not to beat a dead horse, but the Data Analysis add-in's Moving Average tool doesn't correctly chart the forecasts against the actuals in the baseline. The way that the Moving Average tool lines up the forecast values against the baseline values, it treats the actual value for the current period as part of the forecast for that period.



WARNING

Another problem with the Data Analysis add-in is the standard errors that the Exponential Smoothing tool returns. You can find more information about this in Chapter 15.

In words: If you're looking for a moving-average forecast of length 3 for July, you want that forecast to average the baseline for April, May, and June. But the Data Analysis add-in's Moving Average tool forecasts July by averaging May, June, and July, and that's clearly wrong. If you're going to include July in the forecast for July, then you have to wait for July's actual. In that case, what's the point of forecasting July? You avoid this tail-chasing by calculating and placing the moving averages yourself.

## Static electricity

Every tool in the Data Analysis add-in other than Moving Average and Exponential Smoothing puts static values on your worksheet. (Before I wrote this book, I hadn't counted and checked every Data Analysis add-in tool to see whether it returns formulas or values. I was a little surprised that, of the three forecasting tools this book discusses, two of them are the only two tools in the Data Analysis add-in that return formulas.)

Formulas are usually better than static results: With formulas, you can change the inputs and the results will recalculate. Using a tool that returns static results means that you have to rerun the analysis when your inputs change, or when the next period's actuals become available.

## Managing the layout

Using your own formulas, you can lay out the worksheet however you want, and recalculate with new inputs without disturbing the layout. With the Data Analysis add-in, though, if you have new inputs and you're using a tool that returns static values, you'll need both to rerun the tool and then move the results around again to get the layout the way you want it.

## Getting the syntax right

Functions like the ones used in the last section (such as AVERAGE and OFFSET) soothe a lot of headaches on your behalf. All you have to do is supply the right values, or point them at the right cells, and they'll take care of the rest.

But sometimes you have to be careful to supply those values and cells in a particular order, and sometimes you don't. If you're using the SUM function, for example, you need to tell it which values to total, or which cells contain the values you want it to total, but order doesn't matter. You could use this:

```
=SUM(5,4,3,2,1)
```

or, equivalently, this:

```
=SUM(1,2,3,4,5)
```

Or, if you want to get the sum of several cells (and that's really the way to do it, instead of putting constants inside the parentheses), you could use this:

```
=SUM(A1,B2,C3,D4,E5)
```

or, equivalently, this:

```
=SUM(E5,C3,A1,B2,D4)
```

But other functions are persnickety about the order of the numbers and cell addresses that you supply them. Returning to OFFSET, consider that

```
=OFFSET(A1,1,0,1,2)
```

returns the range A2:B2. But just swapping the order of the first two numbers and of the second two numbers,

```
=OFFSET(A1,0,1,2,1)
```

returns the range B1:B2.

By the way, the values inside the parentheses following the name of a function are called its *arguments*. The arguments just give the function the information it needs to operate properly.



TIP

An Excel function cannot handle more than 255 arguments. So if you want to get the sum of 256 individual numbers, you're out of luck. Excel won't react properly if you enter =SUM(1,2,3,4 . . . 256). But — and it's a big but — Excel treats a range of cells as just one argument. So if you had the number 1 in cell A1, the number 2 in cell A2, and so on down to the number 256 in cell A256, this would work just fine:

```
=SUM(A1:A256)
```

## Using Insert Function

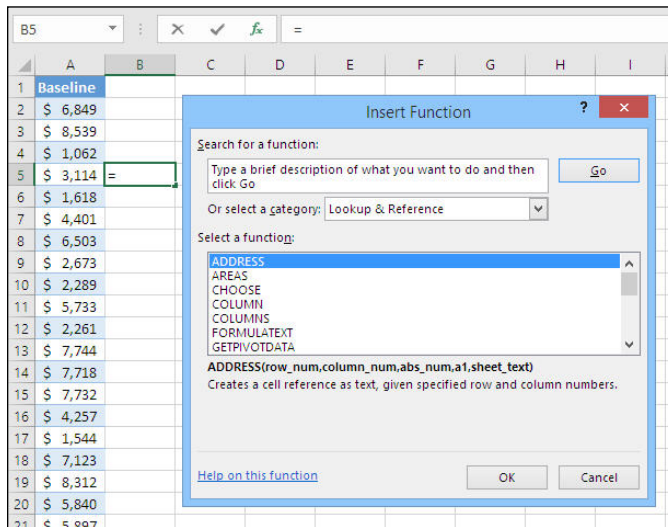
Without years of experience (and even *with* them), keeping the necessary order of a function's arguments in mind is difficult. Fortunately, you don't have to. Use Insert Function by clicking the  $f_x$  button, just to the left of the Formula Bar. When you do, two things happen:

- » The Formula Bar gets itself ready to accept a formula and to put it in the active cell. (Notice in Figure 12-3 that both the Formula Bar and cell B3 now contain equal signs.)
- » The Insert Function dialog box appears.

Suppose you want to enter the OFFSET function on your worksheet, and you'd like to use the Insert Function button to help you out with OFFSET's arguments. After you've clicked the Insert Function button to open its dialog box, there are four ways to get that assistance:

- » Choose All in the Select a Category list box. Scroll down the Select a Function list box until you find OFFSET, select it, and then click OK.
- » You may know from experience that OFFSET belongs to the Lookup and Reference category. So, select Lookup and Reference in the Select a Category list box, scroll down the Select a Function list box, select OFFSET, and click OK. The virtue of starting by selecting Lookup and Reference first is that, in this case, you have fewer than 20 Lookup and Reference functions to deal with in the Select a Function list box, rather than over 200.

**FIGURE 12-3:**  
Over 200 English language worksheet functions are available to you, so breaking them up into categories makes pretty good sense.



- » If you used the OFFSET function recently (at least, recently as far as Excel is concerned), you may start by choosing Most Recently Used from the Select a Category list box. Excel puts the ten functions that you used most recently in the Select a Function list box — so, you may not have to scroll at all. Select OFFSET and click OK.
- » If you're really in the dark (you don't know the function's category, you don't know its name, and it doesn't show up when you choose the Most Recently Used category), you can try typing a question in the Search for a Function box and click Go. Because you're relying on finding a keyword in the function's description, this is the least reliable way to locate a particular function. But sometimes it's all you've got.



TIP

When you can't find what you're looking for using Insert Function, try Google. Enter some pertinent words in Google's search box.

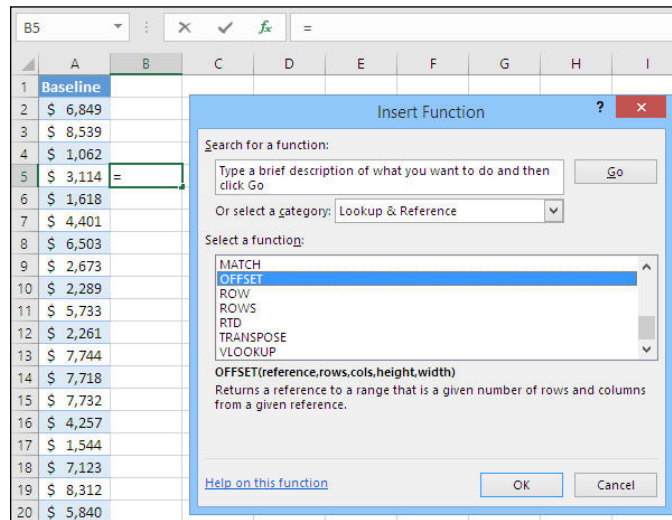
After you've selected a category, the Select a Function list box adjusts its entries (all the available functions, all the functions that belong to a smaller category such as Lookup and Reference, or the ten most recently used), as shown in Figure 12-4.



TIP

The category that a function belongs to is a little arbitrary. For example, Excel puts the TRANSPOSE function into the Lookup and Reference category. This isn't unreasonable, because TRANSPOSE puts a horizontal range of cells (such as A1:E1) into a vertical range (such as F1:F5) — or vice versa, and so it sort of belongs to Lookup and Reference. Putting it there is a bit of a stretch, though, because TRANSPOSE neither looks something up nor provides a reference, as COLUMN

does. I'd expect to find it in the Math and Trig category (TRANSPOSE is used in matrix algebra) or conceivably in the Statistical category (it can be used in regression analysis). The point is that if you don't find what you're looking for in one category, try another possible category, or go back to the All category.



**FIGURE 12-4:** Clicking a function causes the names of its arguments and its purpose to appear in the Insert Function dialog box.



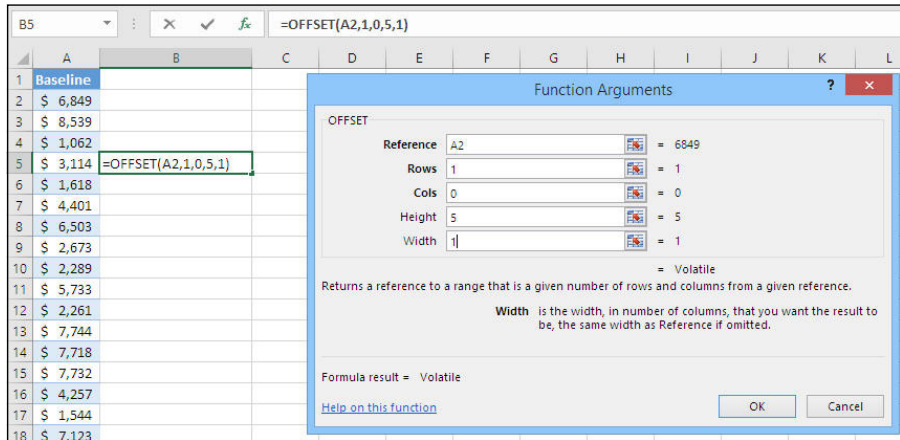
**TIP**

I've been a little coy about the number of functions that are available in Excel, because it depends. If you haven't installed the Data Analysis add-in, you have 235 functions available to you. The Data Analysis add-in adds some functions, though, and it's both possible and feasible for you to write your own functions if you're comfortable with Visual Basic. If you install the Data Analysis add-in, you have an additional 103 functions available, largely in the areas of imaginary numbers and coupons (you use the coupon functions when you're analyzing securities, not in the express checkout line — at least, not when you're ahead of me, you don't). And these days there's a bunch of functions in languages other than English, such as RisolutoreAggiungi, most of which appear to be Solver functions.

After you've selected the OFFSET function and you've clicked OK, you see the Function Arguments dialog box, shown in Figure 12-5.

Now that you're at the Function Arguments dialog box, a lot of problems resolve themselves. The dialog box shows you the name of each argument; it gives you those names in the order that the function expects them to come. In each box, you just put the value you want, or first click in the box you want to use next, and then click a worksheet cell that contains that value.

**FIGURE 12-5:**  
The Function Arguments dialog box shows the names of the arguments that the function *requires* in boldface.



To the right of each argument box is a label that shows what kind of information you're expected to provide. It's hard to see on the printed page, so Figure 12-5 shows the actual argument values rather than the type of argument that's expected. In that figure, the Reference argument is expected to be a worksheet address (to the right of the box is the word *reference* in a gray font on a light-gray background). The rest of the arguments to OFFSET are all *number*.

If you see *reference* next to an argument box, you'll need to enter a cell or range address in the box. If you see *number* or *text*, you can enter a numeric or a text value, or you can enter a cell reference that contains the value you want to use for that argument.

As soon as you've entered the final required argument, the result of the function appears in the dialog box to the right of *Formula result* =.

Excel has a few functions (including OFFSET, RAND, and NOW) that recalculate every time the worksheet is recalculated. The Insert Function dialog box shows their value as *Volatile* regardless of the value they return at the time you use Insert Function.



TIP

There's a difference between an Excel formula and an Excel function. *Formula* is a more inclusive term. It might contain a function all by itself:

```
=OFFSET(A1,5,3,10,1)
```

or fixed values only (this is pretty rare in a well-designed worksheet):

```
=96/12
```

or a combination of a function with fixed values:

```
=RAND()*1000
```

*Function* refers to the function itself, with its arguments. For example, you could refer to

```
=OFFSET(A1,5,3,10,1)
```

as either a function or a formula. But

```
=96/12
```

is not a function.

As the Function Arguments dialog box implies, you can get to detailed Help on the function by clicking the Help on This Function hyperlink in the dialog box's lower-left corner.

There's a button at the right end of each argument box. It's called a *collapse dialog* button. If you click it, the dialog box collapses and shows only the argument box whose button you clicked. This gives you more room on the worksheet (see Figure 12-6).

**FIGURE 12-6:** Click the button again to restore the Function Arguments dialog box.

The screenshot shows an Excel spreadsheet with a sales data table. The table has columns for months (January to November) and rows for sales representatives (Brown, Buzzl, Gibson, Hawm, Johnson, Martin, Owens, Rowan, Tomlin, Worley). A 'Function Arguments' dialog box is collapsed over the data, showing only the 'Sales' argument box for the 'Gibson' row. The dialog box title bar is visible, and a small 'x' button is on the right side of the argument box.

	Sales	January	February	March	April	May	June	July	August	September	October	November	
1 Rep	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	
2 Brown	\$ 17,446	\$ 18,963	\$ 33,824	\$ 19,868	\$ 29,875	\$ 15,766	\$ 21,226	\$ 3,353	\$ 26,834	\$ 23,045	\$ 5,365		
3 Buzzl	\$ 29,231	\$ 5,766	\$ 13,757	\$ 29,610	\$ 10,496	\$ 24,057	\$ 24,122	\$ 28,233	\$ 3,715	\$ 30,663	\$ 30,486		
4 Gibson	\$ 7,225	Function Arguments										700	\$ 21,464
5 Hawm	\$ 6,343											977	\$ 35,934
6 Johnson	\$ 31,495											000	\$ 20,140
7 Martin	\$ 25,517	\$ 17,704	\$ 15,442	\$ 11,231	\$ 21,714	\$ 9,502	\$ 7,401	\$ 24,771	\$ 26,926	\$ 10,844	\$ 25,334		
8 Owens	\$ 9,631	\$ 25,440	\$ 23,057	\$ 14,062	\$ 13,610	\$ 30,492	\$ 29,967	\$ 35,922	\$ 12,994	\$ 30,750	\$ 18,804		
9 Rowan	\$ 22,342	\$ 30,357	\$ 19,752	\$ 12,138	\$ 31,752	\$ 20,312	\$ 31,006	\$ 15,263	\$ 32,720	\$ 23,105	\$ 32,557		
10 Tomlin	\$ 8,921	\$ 35,175	\$ 16,549	\$ 21,110	\$ 11,233	\$ 19,155	\$ 34,407	\$ 32,892	\$ 29,430	\$ 31,501	\$ 15,453		
11 Worley	\$ 23,968	\$ 12,667	\$ 3,901	\$ 8,454	\$ 10,921	\$ 33,860	\$ 32,601	\$ 13,975	\$ 18,193	\$ 25,012	\$ 32,010		

You don't really need that button, though. You can always click the dialog box's title bar and drag it out of the way if it's hiding cells that you want to click in.

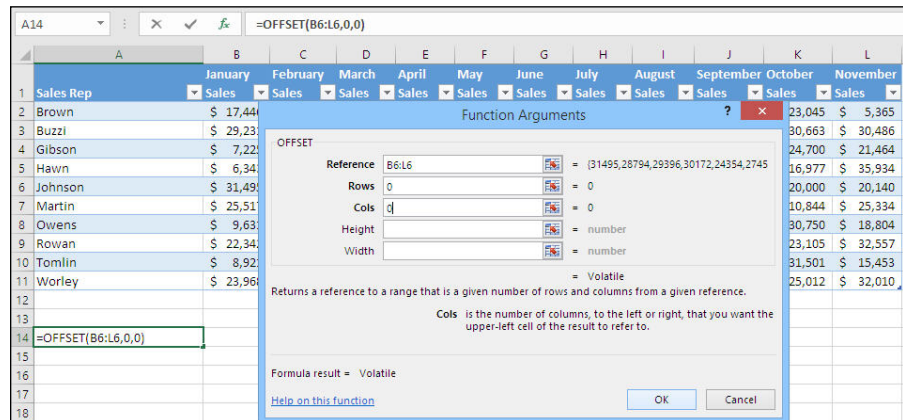


TIP

You can enter the address of a range of cells that are on a different worksheet. Just click that worksheet's tab to activate it and drag through the cells you want. However, if the tab isn't visible — perhaps because you have so many worksheets that there isn't room to show all the tabs — you'll have to arrange to make it visible

before you start this whole process by clicking the Insert Function button. One way to make the hidden worksheet tab visible is to scroll until you can see it, then right-click its sheet tab and choose Move or Copy Sheet from the shortcut menu.

Assuming you want the monthly sales results for the sales rep named Johnson, using the OFFSET function, Figure 12-7 shows how the Function Arguments dialog box would look after you filled in the argument boxes.



**FIGURE 12-7:** Notice that the arguments appear in the correct order in the Formula Bar.



**TIP**

If you omit the fourth and fifth arguments from the OFFSET function, it defaults to the number of rows and columns in OFFSET's first argument, *Reference*. So, given that you've begun OFFSET's arguments with B6:L6, a one-row eleven-column reference, OFFSET returns a reference with those dimensions, just as though you had supplied 1 and 11 as its fourth and fifth arguments. Put another way, these two instances of OFFSET return the same reference:

```
=OFFSET(B6,0,0,1,11)
=OFFSET(B6:L6,0,0)
```

You might find it easier to supply the B6:L6 argument by clicking and dragging rather than to enter the 1 and 11 arguments by counting the rows and the column.

Now, you click OK in the Function Arguments dialog box, to get the result shown in Figure 12-8.

A couple of things to bear in mind if you're using Insert Function to put a function in several cells at once, as shown in Figure 12-3 through Figure 12-8:



- » Begin by selecting all the cells you want the function to occupy. In Figure 12-8, that's B14:M14. There are 12 months, preceded by the sales rep's name, and M is the 13th column from column A.
- » When you're ready to click OK on the Function Arguments dialog box, first press Ctrl and continue pressing it as you click OK. This signals Excel that the function is to occupy all the selected cells. You do this directly on the worksheet if you're not using Insert Function. Select a range of cells, enter a formula, and finish with Ctrl+Enter rather than just Enter. (This is different from array formulas, which often also occupy multiple cells, but which you enter with Ctrl+Shift+Enter. See "Understanding Array Formulas.")

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Rep	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales	Sales
2	Brown	\$ 17,446	\$ 18,963	\$ 33,824	\$ 19,868	\$ 29,875	\$ 15,766	\$ 21,226	\$ 3,353	\$ 26,834	\$ 23,045	\$ 5,365	\$ 7,654
3	Buzzi	\$ 29,231	\$ 5,766	\$ 13,757	\$ 29,610	\$ 10,496	\$ 24,057	\$ 24,122	\$ 28,233	\$ 3,715	\$ 30,663	\$ 30,486	\$ 26,373
4	Gibson	\$ 7,225	\$ 7,766	\$ 32,877	\$ 16,221	\$ 30,051	\$ 24,174	\$ 30,679	\$ 15,878	\$ 21,814	\$ 24,700	\$ 21,464	\$ 18,883
5	Hawn	\$ 6,343	\$ 29,394	\$ 15,398	\$ 7,271	\$ 7,921	\$ 34,231	\$ 31,914	\$ 16,727	\$ 12,112	\$ 16,977	\$ 35,934	\$ 9,251
6	Johnson	\$ 31,495	\$ 28,794	\$ 29,396	\$ 30,172	\$ 24,354	\$ 27,454	\$ 17,374	\$ 32,658	\$ 21,566	\$ 20,000	\$ 20,140	\$ 10,067
7	Martin	\$ 25,517	\$ 17,704	\$ 15,442	\$ 11,231	\$ 21,714	\$ 9,502	\$ 7,401	\$ 24,771	\$ 26,926	\$ 10,844	\$ 25,334	\$ 22,533
8	Owens	\$ 9,631	\$ 25,440	\$ 23,057	\$ 14,062	\$ 13,610	\$ 30,492	\$ 29,967	\$ 35,922	\$ 12,994	\$ 30,750	\$ 18,804	\$ 22,665
9	Rowan	\$ 22,342	\$ 30,357	\$ 19,752	\$ 12,138	\$ 31,752	\$ 20,312	\$ 31,006	\$ 15,263	\$ 32,720	\$ 23,105	\$ 32,557	\$ 23,953
10	Tomlin	\$ 8,921	\$ 35,175	\$ 16,549	\$ 21,110	\$ 11,233	\$ 19,155	\$ 34,407	\$ 32,892	\$ 29,430	\$ 31,501	\$ 15,453	\$ 12,866
11	Worley	\$ 23,968	\$ 12,667	\$ 3,901	\$ 8,454	\$ 10,921	\$ 33,860	\$ 32,601	\$ 13,975	\$ 18,193	\$ 25,012	\$ 32,010	\$ 27,154
12													
13													
14		31495	28794	29396	30172	24354	27454	17374	32658	21566	20000	20140	10067

**FIGURE 12-8:** A function that picks up existing values and shows them elsewhere, such as OFFSET or INDEX, picks up the values only, not the formats,

## Understanding Array Formulas

You can use array formulas to get a lot of unusual results, such as:

- » Reversing the order of the letters in a word or phrase (so that *Washington* becomes *notgnihsaW*).
- » Separate a last name from a full name. Suppose you have a table of sales reps with their full names in column A, and you want to sort the list by last name. A fairly complicated array formula will strip the last names out of the list and put them in column B, which you can then use to sort the table. (This approach often works better than Text to Columns when some people use two names and others use three names, or an honorific such as Ms. or Dr.)
- » Determining if two different sets of data have duplicate values.

“Understanding Array Formulas” is without doubt the most grandiose, arrogant heading I’ve ever written. I’ve been using Excel for roughly 20 years and I still don’t fully understand array formulas, so how can I expect to explain them in one section of a chapter? But I do know when I should *use* an array formula.

An overbroad and oversimplified guide is that you use an array formula rather than a regular formula when:

- » You’re using a function whose purpose requires that you *must* array-enter it.
- » You’re using a function that doesn’t normally expect to get an array of worksheet cells as an argument. (For example, the IF function doesn’t normally expect to get an array of worksheet cells as its condition, but that structure can be very useful if you have multiple values to test in one argument.)
- » You’re manipulating two different arrays of worksheet cells. (For example, multiplying B2:B5 by C2:C5 so as to get four different products in D2:D5.)

As I just indicated, some functions, such as TRANSPOSE, LINES, and TREND require you to array-enter them; if you don’t, you won’t get the result you’re after, and you may even get an error value instead. LINES and TREND are of special interest in forecasting, particularly if you’re entering the formulas yourself rather than relying on the Data Analysis add-in’s Regression tool. Further, TREND makes for an easy forecast of values if all you’re interested in is the forecast and not at all in the regression equation, or the statistical analysis itself.



Despite what the prior sentence implies, it’s *never* a good idea to unthinkingly accept the results of using the TREND function. There are several ways to go wrong, and you can’t depend on TREND’s results to alert you. I always use LINES in conjunction with TREND to check for unexpected conditions, such as a regression coefficient with a standard error of zero.

## Choosing the range for the array formula

One reason that array formulas can require some specialized knowledge is that you need to select the range that the results will occupy before you enter the array formula. Suppose you want to transpose the range B1:F1 in Figure 12-9.

You want to transpose B1:F1 into B3:B7, and one good way is to use the TRANSPOSE function. But you need to know that transposing a range of, say, one row and five columns results in a range that occupies five rows and one column. Otherwise you won’t know the dimensions of the range where you’ll array-enter the TRANSPOSE function.

	A	B	C	D	E	F	G
1		\$ 7,956	\$ 9,343	\$ 7,841	\$ 8,348	\$ 9,851	
2							
3							
4							
5							
6							
7							
8							

**FIGURE 12-9:**  
The range B1:F1  
has the same  
number of  
columns as B3:B7  
has rows.



TIP

Another way is to go to the Ribbon's Home tab. Copy B1:F1 and right-click cell B3. Click Paste Special in the shortcut menu, select the Transpose check box, and click OK. But this pastes B1:F1 into B3:B7 as values, not as formulas. So if you later change a value in B1:F1, that change won't be reflected in B3:B7.



TIP

Transposing data comes up frequently in forums and newsgroups that involve Excel. The question usually arises when someone has a worksheet that contains a moderately large table of values, and someone wants to use it in a formal report. The problem is that the original table doesn't work well in a report — but if it were turned 90 degrees it would be perfect. For example, the original list might put sales reps' names in one column and monthly sales in adjacent columns. You'd like to present it to your corporate management team the other way around: Sales reps' names in one row and monthly sales in subsequent rows. To achieve that, the Excel community generally recommends using the TRANSPOSE function.

So you need to know that TRANSPOSE will put the values in B1:F1 into a one-column-by-five-row range, and you need to start by selecting a range with those dimensions. Then, array-enter the TRANSPOSE function with the argument B1:F1:

```
=TRANSPOSE(B1:F1)
```

How do you array-enter a function or formula? Use the three-finger salute. No, not Ctrl+Alt+Delete. I mean the three-finger salute à la Excel (see the following section).

## Excel's three-finger salute: Ctrl+Shift+Enter

You tell Excel that you want your formula treated as an array formula by a special keyboard sequence: Ctrl+Shift+Enter. In words, after you've typed the formula, hold down Ctrl and Shift at the same time, and while you're still holding them down, press Enter. Then release all three keys.

You would also use Ctrl+Shift+Enter if you've used Insert Function to build the function's argument list. After entering the arguments, you would hold down Ctrl and Shift, and then either press Enter or click OK in the Function Arguments dialog box.

See Figure 12-10 for the result of the example begun in Figure 12-9.

**FIGURE 12-10:**  
Using Ctrl+Shift+Enter is the only way to get the result in B3:B7 by means of the TRANSPOSE function.

	A	B	C	D	E	F	G	H	I	J
1		\$ 7,956	\$ 9,343	\$ 7,841	\$ 8,348	\$ 9,851				
2										
3		7956								
4		9343								
5		7841								
6		8348								
7		9851								
8										
9										



REMEMBER

Select a range with the correct dimensions before you array-enter your formula. This comes up a lot in forums and newsgroups: “I array-entered a function. The Help documents say I’m supposed to see values in several cells, but I see a value in just one cell. What’s going on?” And the answer, inevitably and invariably, is “Start by selecting the correct *range* of cells, not just a single cell.”

## Recognizing array formulas

Occasionally, you find yourself dealing with an Excel workbook that you didn’t create. It might have array formulas in it, and if you’re going to understand what the worksheet does and how it does it, then you’re going to have to recognize that a formula is an array formula.

An array formula has a special appearance in the Formula Bar: it’s surrounded by curly brackets, like this:

```
{=TRANSPOSE(B1:F1)}
```

You won’t see the curly brackets in the worksheet cells that contain the array formula.

And bear this in mind: When you *enter* an array formula, you shouldn’t type the curly brackets yourself. Leave that up to Excel. If you type the curly brackets yourself, Excel treats the formula as text.

# A special problem with array formulas

Array formulas pose a special problem that requires special handling. It's an all-or-none issue.

## Changing part of an array formula

You can't.

Suppose you've entered an array formula in A1:A5, and you later decide that you don't need to see the value in A5. So you select A5 and press the Delete key, or you try to type another formula or value into it. Excel won't let you. You'll get the terse, even testy, error message, *You can't change part of an array.*

You'll also get that message if you enter either a regular formula or an array formula in a range of cells that overlaps part of an existing array formula.

## Resistance is futile

But you don't need to let yourself be bullied by Excel — you're the one who's in control. Here's the fastest way around the *You can't change part of an array* problem. Continuing the example started in the preceding section:

- 1. Select the entire range that's occupied by the array formula.**

In this example, that's A1:A5. You'll see the formula in the Formula Bar.

- 2. Select the equal sign in the Formula Bar by dragging across it.**

- 3. Press the Delete key.**

You should now see the formula, minus the equal sign, in the Formula Bar.

- 4. Press Ctrl+Enter.**

This enters the formula in the selected range as text values only.

- 5. Select A5 and press the Delete key.**

- 6. Select A1:A4, reenter the equal sign, and press Ctrl+Shift+Enter.**

Now your array formula occupies A1:A4 and A5 is free for you to use in some other way.

# Using the Regression Functions to Forecast

When you decide that you want to use formulas with functions that calculate a regression equation and related statistics, you're usually thinking of using `LINEST` and `TREND`. If you have just one predictor variable, you can also use `SLOPE` to get the best regression coefficient and `INTERCEPT` for the constant to use with your predictor variable (you use just `Enter`, not `Ctrl+Shift+Enter`, to enter `SLOPE` and `INTERCEPT`).

## Using the `LINEST` function

The `LINEST` function shows you the following information:

- » **The coefficient that you use to multiply by the predictor variable to get the best regression estimate of the forecast variable.**
- » **The standard error of each coefficient:** You can use this information to decide whether to include a predictor variable in the regression equation.
- » **The R-squared value:** This value ranges from 0.0 to 1.0. The closer it is to 1.0, the more accurate the regression equation is as an approach to forecasting the data you handed off to `LINEST`. You can interpret the R-squared value as the percent of the variation in the forecast variable that you can attribute to the predictor variable or variables.
- » **The standard error of estimate:** This is the standard deviation of the residuals — that is, the differences between the actual values in the baseline and the values predicted by the regression equation. The smaller the standard error of estimate, the better the forecast. (If you want to verify this, be sure to use the degrees of freedom returned by `LINEST` rather than  $n - 1$  in the denominator of the standard error.)
- » **The F-ratio, the degrees of freedom, the regression sum of squares, and the residual sum of squares:** If your last statistics course was a while back, or if you haven't heard of these statistics, don't worry about it. You don't need them to do forecasting, although they can give you good information about the *reliability* of your forecasts. If you have heard of them, you already know what they're about — it's not like they come up in casual chit-chat.

Figure 12-11 shows both the `LINEST` function, array-entered, and the results of the Data Analysis add-in's Regression tool. Both are used on the same data set, in A1:B41.

**FIGURE 12-11:**  
All the statistics  
returned by  
LINEST show up  
in the Regression  
tool's output.

	A	B	C	D	E	F	G	H	I	J	K
3		2 \$	2,910		Regression Statistics						
4		3 \$	1,062		Multiple R	0.675					
5		4 \$	3,114		R Square	0.455					
6		5 \$	1,618		Adjusted R Square	0.441					
7		6 \$	4,401		Standard Error	1928.630					
8		7 \$	6,503		Observations	40					
9		8 \$	2,673		ANOVA						
10		9 \$	2,289								
11		10 \$	5,733			<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12		11 \$	2,261		Regression	1	117992002.614	117992002.614	31.722	0.000	
13		12 \$	5,010		Residual	38	141345323.361	3719613.773			
14		13 \$	5,600		Total	39	259337325.975				
15		14 \$	6,870								
16		15 \$	4,580			<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17		16 \$	4,257		Intercept	2596.608	621.504	4.178	0.000	1338.438	3854.777
18		17 \$	4,520		Time Period	148.786	26.417	5.632	0.000	95.308	202.265
19		18 \$	7,123								
20		19 \$	8,312								
21		20 \$	5,840		LINEST:	148.786	2596.608				
22		21 \$	5,897			26.417	621.504				
23		22 \$	4,450			0.455	1928.630				
24		23 \$	4,276			31.722	38				
25		24 \$	6,762			117992002.614	141345323.361				

In this case, you would enter LINEST by following these steps:

1. Select a range that's two columns wide and five rows high.
2. Type =LINEST(B2:B41,A2:A41,,TRUE) into the Formula Bar or use Insert Function to let it guide you.

Notice that you don't include row 1, because LINEST can't handle text labels.

3. Array-enter the formula using Ctrl+Shift+Enter.

You may want to have another look at that LINEST syntax, whether you enter it directly or use Insert Function to help you follow the trail of breadcrumbs:

- » The first argument, B2:B41, is where Excel looks to find the baseline values of the forecast variable — in this case, Sales.
- » The second argument, A2:A41, is where Excel looks to find the baseline values of the predictor variable — in this case, Time Period.
- » The third argument is left blank — notice the consecutive commas among the arguments. If it's blank, or if you set it to TRUE, LINEST calculates the value of the intercept (or *constant*) normally. This tells you the value of Sales at the time period zero: where the regression line intercepts the vertical axis on a chart (that's why it's often termed the *intercept* rather than the *constant*). If the third argument is 0, or if you set it to FALSE, LINEST forces the regression line to intercept the vertical axis at its zero point. See Chapter 11 for more information and for why this is seldom a good idea.

- » The fourth argument, TRUE, tells Excel whether to calculate the third through fifth rows of its output: the R-squared value, the standard error of estimate, and so on. If TRUE, Excel calculates those statistics and displays them if you started by selecting a range with five rows. If you enter FALSE, Excel does not calculate the additional statistics. You lose nothing by setting the argument to TRUE, and you can gain considerably by evaluating the size of the R-squared value.

See Chapter 11 for information on how to use the Regression tool to get the results shown in Figure 12-11.

In Figure 12-11, the Regression tool's output in the range E1:K18 contains some shaded cells. These shaded cells contain values that are provided by using the LINEST function, shown in the range E21:F25. In particular:

- » The coefficients to use in making a forecast are found in F17 and F18 (Regression tool) and in E21:F21 (LINEST). So the regression equation is:

```
Forecast, Time Period 2 = 2596.608 + 148.786 * 2  
Forecast, Time Period 2 = 2894.18
```

In words, the forecast for any time period that you get by using regression on this baseline is the intercept, 2596.608 plus the result of multiplying the coefficient for Time Period, 148.786 by the actual value for that time period. By the way, the Regression tool calls the intercept a coefficient, to keep it in the same column as the actual coefficients and, thus, to make for a neater table of statistics. But it's not really a coefficient; it's an intercept.

- » The R-squared value for the regression equation is found in F5 (Regression tool) and E23 (LINEST).
- » The standard errors for the coefficients and intercept are found in G17:G18 (Regression tool) and in E22:F22 (LINEST).
- » The standard error of estimate is found in F7 (Regression tool) and F23 (LINEST).
- » The sums of squares are found in G12:G13 (Regression tool) and E25:F25 (LINEST). The degrees of freedom are found in F13 (Regression tool) and F24 (LINEST). The F-ratio is found in I12 (Regression tool) and E24 (LINEST).

Clearly, you get a lot more information with the Regression tool than with LINEST. You can get all that additional information by using other Excel worksheet functions. For example:

- » T.DIST to get the P-value for a regression coefficient
- » The ratio of the coefficient or intercept to its standard error to get the what the regression tool calls the *t Stat*



- » Dividing the sums of squares by the degrees of freedom to get the mean square (MS)

But I'm not sure why you'd necessarily want all this other information. To make a forecast, you need the R-squared value to decide whether regression is a good tool to use on your baseline, and the intercept and coefficient(s) to actually make the forecast. If you want to push the envelope and judge the regression in terms of probabilities, *then* you'll want to make use of the F ratio and the residual degrees of freedom. But statistical analyses that determine probabilities require careful experimental design that's generally missing from sales baselines. See Chapter 16 for a very brief overview of the F ratio.

## Selecting the right range of cells

If you want to use LINEST to get the really important stuff for forecasting, you need to be aware of how many rows and how many columns to select before you enter the LINEST formula.

- » The R-squared value shows up in the third row of LINEST's results. So the range of cells you select before array-entering the LINEST function should have at least three rows (and no more than five — subsequent rows in the range that you select will just show the error value #N/A).
- » The range that you select for the LINEST function should have as many columns as you have predictor variables, plus 1. LINEST returns a coefficient for each predictor variable, plus the intercept.

So if you're using two predictor variables, you should select a range with three columns and at least three rows (at most, five rows).

## Getting the statistics right

Figure 12-12 shows an example with two predictor variables, again analyzed using both the Regression tool and the LINEST function.

The location of the key forecasting statistics is similar to the single predictor situation shown in the "Using the LINEST function" section, earlier in this chapter. One difference is that the Regression tool's output has an additional row, and the LINEST output has an additional column, because you're now using not one but two predictor variables.

**FIGURE 12-12:**  
Notice that LINEST's arguments specify that the predictor variables, Size of Sales Force and Time Period, occupy columns A and B. Multiple predictors result in multiple regression.

	A	B	C	D	E	F	G	H	I	J	K
E22											
	Size of Sales Force	Time Period	Sales	Regression Tool: SUMMARY OUTPUT							
1	89	1	\$ 2,660								
2	21	2	\$ 2,910								
3	25	3	\$ 1,062								
4	21	4	\$ 3,114								
5	46	5	\$ 1,618								
6	93	6	\$ 4,401								
7	82	7	\$ 6,503								
8	74	8	\$ 2,673								
9	85	9	\$ 2,289								
10	47	10	\$ 5,733								
11	40	11	\$ 2,261								
12	62	12	\$ 5,010								
13	52	13	\$ 5,600								
14	42	14	\$ 6,870								
15	70	15	\$ 4,580								
16	63	16	\$ 4,257								
17	51	17	\$ 4,520								
18	68	18	\$ 7,123								
19	63	19	\$ 8,312								
20	71	20	\$ 5,840								
21	56	21	\$ 5,897								
22	45	22	\$ 4,450								
23	69	23	\$ 4,276								
24	82	24	\$ 6,762								
25	81	25	\$ 9,705								
				Regression Statistics							
				Multiple R	0.711						
				R Square	0.506						
				Adjusted R Squ	0.479						
				Standard Error	1861.119						
				Observations	40						
				ANOVA							
					df	SS	MS	F	Significance F		
				Regression	2	131178094.4	65589047.21	18.936	0.000		
				Residual	37	128159231.6	3463763.015				
				Total	39	259337326					
					Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	
				Intercept	1086.393	979.190	1.109	0.274	-897.635	3070.420	
				Size of Sales Fo	23.444	12.015	1.951	0.059	-0.902	47.789	
				Time Period	160.930	26.241	6.133	0.000	107.760	214.100	
				LINEST:	160.930	23.444	1086.393				
					26.241	12.015	979.190				
					0.506	1861.119	#N/A				
					18.936	37	#N/A				
					131178094.422	128159231.553	#N/A				

Figure 12-12 hints at an extremely annoying aspect of LINEST — really, it's one of only a few annoyances in this function, but it can be a major one. Compare the order of the coefficients in LINEST to the order of the predictor variables in the baseline. For example:

- » The value in G22 (returned by LINEST) is the same as the value in F17 (returned by the Regression tool). It is the intercept for the regression equation: LINEST always puts the intercept in the first row, final column of the range it occupies. (This assumes that you started out by selecting a range with the proper number of columns.) No problem so far.
- » The value in F22 (LINEST) is the same as the value in F18 (Regression tool). It is the coefficient for Size of Sales Force.
- » The value in E22 (LINEST) is the same as the value in F19 (Regression tool). It is the coefficient for Time Period.

So LINEST puts the coefficient for Time Period in column E and for Size of Sales Force in column F. In the LINEST results, Time Period comes before Sales Force. But in the baseline, Sales Force comes before Time Period: Sales Force in column A, Time Period in column B. LINEST reverses the order of the predictor variables.

This means that if you're going to use the LINEST results in an equation that returns the predicted values, you have to remember to get the correct coefficient aligned with the correct column in the baseline. Figure 12-13 shows how you might do that.

**FIGURE 12-13:** You'll have to do it yourself, but it can be helpful to label the columns of the LINEST results according to which predictor variables they represent.

H3												
={LINEST(C2:C31,A2:B31,,TRUE)}												
	A	B	C	D	E	F	G	H	I	J	K	L
1	Size of Sales	Time	Sales		Sales				Size of Sales			
2	Force	Period	Sales		Forecast			Time Period	Force	Intercept		
3							LINEST:	Coefficient	Coefficient			
2	89	1	\$ 2,660		\$ 3,106.60							
3	21	2	\$ 2,910		\$ 2,437.09			153.817	12.108	1875.186		
4	25	3	\$ 1,062		\$ 2,639.33			36.301	15.246	1065.183		
5	21	4	\$ 3,114		\$ 2,744.72			0.420	1711.055	#N/A		
6	46	5	\$ 1,618		\$ 3,201.23			9.765	27	#N/A		
7	93	6	\$ 4,401		\$ 3,924.12			57179301.955	79048190.045	#N/A		
8	92	7	\$ 6,503		\$ 4,065.83							
9	74	8	\$ 2,673		\$ 4,001.71							
10	86	9	\$ 2,289		\$ 4,300.82							
11	47	10	\$ 5,733		\$ 3,982.43							

The formula in cell E2 in Figure 12-13 multiplies a predictor value in column A (that's the column that the table labels Size of Sales Force) by a coefficient in column I, and a predictor value in column B (that's the column that the table labels Time Period) by the coefficient in column H. This formula, copied down through cell E31, untangles the snarl introduced by the order in which LINEST returns coefficients.

There is no sensible reason to reverse the worksheet order of the predictor variables in LINEST's results: no statistical reason, no reason that pertains to coding, no reason regarding how the function is used on the worksheet.

This is a good illustration of the reason that you need to make sure you've got it right before you put a feature such as a function into a widely distributed application. It's also a good illustration of what happens when you turn a programmer loose with a textbook on statistics. By now (several Excel versions after LINEST was first included in Excel), Microsoft wouldn't dare correct the order of the coefficients. Doing so would screw up too many customer worksheets that assume the coefficients are coming out backward.

## Using the TREND function

An easier way to get forecasts from regression — easier than using LINEST — is the TREND function. The trade-off is that you don't see the R-squared value, the coefficients, or the intercept. But of course there's nothing to stop you from using both LINEST and TREND.

As with LINEST, you must array-enter the TREND function (see Figure 12-14 for an example).

	A	B	C	D	E	F
E2					{=TREND(C2:C31,A2:B31)}	
	Size of Sales	Time			Sales	
1	Force	Period	Sales		Forecast	
2	89	1	\$ 2,660		\$ 3,106.60	
3	21	2	\$ 2,910		\$ 2,437.09	
4	25	3	\$ 1,062		\$ 2,639.33	
5	21	4	\$ 3,114		\$ 2,744.72	
6	46	5	\$ 1,618		\$ 3,201.23	
7	93	6	\$ 4,401		\$ 3,924.12	
8	92	7	\$ 6,503		\$ 4,065.83	
9	74	8	\$ 2,673		\$ 4,001.71	
10	86	9	\$ 2,289		\$ 4,300.82	
11	47	10	\$ 5,733		\$ 3,982.43	
12	40	11	\$ 2,261		\$ 4,051.49	
13	62	12	\$ 5,010		\$ 4,471.68	
14	52	13	\$ 5,600		\$ 4,504.42	
15	42	14	\$ 6,870		\$ 4,537.16	
16	70	15	\$ 4,580		\$ 5,030.00	
17	63	16	\$ 4,257		\$ 5,099.06	
18	51	17	\$ 4,520		\$ 5,107.58	
19	68	18	\$ 7,133		\$ 5,167.33	

**FIGURE 12-14:**  
The TREND forecast is identical to the forecast based on LINEST.

TREND is much easier to use for forecasting than LINEST. To get the forecast values in Figure 12-14, follow these steps:

1. Select the range E2:E31.
2. Enter this formula:  

```
=TREND(C2:C31,A2:B31)
```

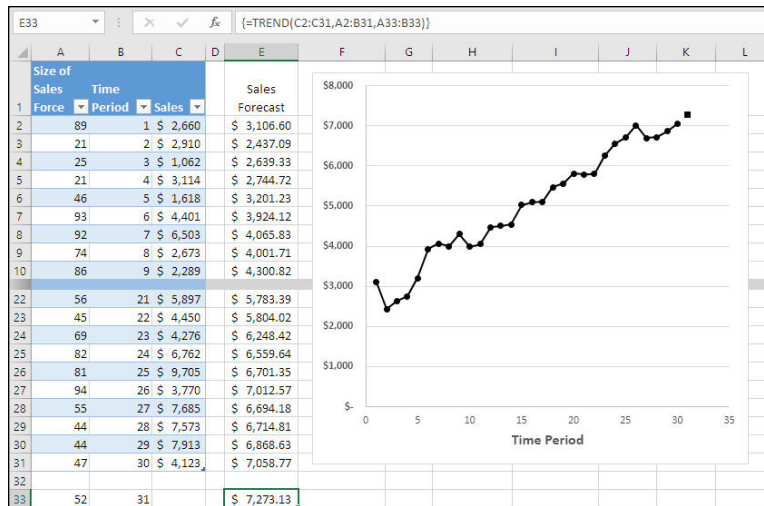
into the Formula Bar.
3. Press Ctrl+Shift+Enter.

What about forecasting the next time period, the one you don't yet have actuals for? As usual, you'll need information about the next period's values on your predictors. In the current example, you need to know (or have a good estimate of) the company's sales force next month; suppose it's 52. The next period's value for Time Period is just the next value in the baseline — here, 31.

Figure 12-15 shows these new values and how you use the TREND function to forecast the next period's sales. (To make room for the full baseline and the forecast for Period 31, I have hidden rows 11 through 21.)

In Figure 12-15, the array formula in cell E33 is:

```
=TREND(C2:C31,A2:B31,A2:B32)
```



**FIGURE 12-15:**  
The forecast for Time Period 31 appears on the chart as a square marker in the upper-right corner.

This form of the TREND function has three arguments. (*Remember:* When used in a function as here, a cell range is an argument to the function.)

- » **C2:C31:** This range of values, the actual sales, is what Excel terms the *known y*'s. The letter *y* is often used to represent the predicted variable in a regression analysis.
- » **A2:B31:** This range of values, the actual Size of Sales Force and the actual Time Period, is what Excel terms the *known x*'s. The letter *x* means the predictor variable (or, as in this case, variables) in a regression analysis.
- » **A33:B33:** Excel terms this range of values the *new x*'s. It includes the predictor variables that you expect for the next, as-yet-unobserved time period. These are in A33:B33. Time Period is easy, but you need special knowledge to determine what the size of the sales force will be next time period.

So these are the steps to use TREND to get a forecast for the upcoming time period:

1. Determine that period's values on the predictor variables.
2. Enter the predictor variables' values for the new period in a row below the existing predictor variables' values.

In Figure 12-15, you'd enter those in A33:B33.

3. Select the cell that will contain the forecast value for the upcoming period.

In Figure 12-15, that's D33.

#### 4. Type this formula:

```
=TREND(C2:C31, A2:B31, A33:B33)
```

#### 5. Array-enter the formula by pressing Ctrl+Shift+Enter.



TIP

Another worksheet function, FORECAST, also predicts the next value in a baseline. Like TREND, it uses known y's, known x's, and new x's. But FORECAST can handle only one predictor variable, whereas TREND can handle multiple predictor variables. To keep things straight and relatively simple, I recommend that you decide to always use one or the other — and I also think that the one you choose should be TREND.



TIP

Excel's tables are handy and useful for several reasons, but occasionally they trip over their own feet. Using a table as the data source for the TREND function is one of those occasions. When you're using TREND to make a forecast, you often want the *new x's* to be in the first row following your baseline. But if you put them there, Excel wants to make them part of the existing table. But you don't want that because it screws up the calculation of the regression equation.

Furthermore, if you try to include TREND forecasts as part of the table, by locating them in an adjacent column, Excel will object that multiple cell array formulas are not allowed in a table.

Here are some reasonable solutions:

1. Start the new x's two rows following the end of the baseline table, as is done in Figure 12-15.
2. Don't use a table. Use instead Excel's old list structure. Before entering your TREND formula, convert the table to a range by using the Convert to Range button on the table's Design tab. You'll lose things, at least temporarily, such as the Table's Totals row and the filter drop-downs. When you've got your forecast, you can return the list to its table status.
3. Use the coefficients and intercept from LINEST to calculate the forecasts, as illustrated in Figure 12-13. You'll need to enter LINEST as an array formula, of course, but it won't be part of the table. And the formulas that calculate the forecasts then need not be array formulas, so they can occupy the table.

Deciding on a length

Figuring out how many baseline values to use

Charting your moving-average trendline

## Chapter 13

# Using Moving Averages

**W**hen you decide to look at a baseline's moving-average values — whether to get a better idea of the baseline's behavior or to make a forecast — the number of values you choose to put into each average has consequences, some of them undesirable. For example, the more values in an average, the smaller the number of averages.

And your choice has an effect on how smooth or rough the moving averages are. Generally, the fewer the data points you include in a moving average, the more it will jump around — but the faster it will react to changes in the baseline. The greater the number of data points, the smoother it becomes — but the slower it will react. This effect can represent a difficult trade-off because we often assume that the smoother a set of moving averages, the better it represents the signal in a baseline.

Excel's Data Analysis add-in has a Moving Averages tool that you can use to put the averages into a chart, along with the original baseline. In this chapter, I show you some ways to improve the tool's effectiveness.

# Choosing the Length of the Moving Average

If you've worked with moving averages before, you may think them too basic to discuss in a book about a topic as grand as forecasting. The humble moving average?

No. Although the basic idea of moving averages is simple and intuitive, they play a starring role in some pretty complicated forecasts. These complex forecasts that combine autoregression with moving averages — and more often than not the moving-average part is the diva and the autoregression part is just the spear carrier. (Actually, if you take those two components far enough, you find that they're the same thing, but I don't take things that far here.)

So if you're going to forecast sales or anything else, it's worth your while to take moving averages seriously. Two concepts are useful when you're considering moving averages: signal and noise.

## Signaling: Left turn coming up?

A baseline that isn't completely random has what's termed *signal*. Signal is the true, credible part of the baseline, whether it's headed up or down or holding steady; whether it's moving up and down with the seasons or with a regional business cycle.

If only you could get your hands on that signal, you'd have your best estimate of what's going to happen next.

Think of what comes out of the speakers when your car begins to get within range of a radio station. As you start to hear the station, you hear a lot of static, but you can faintly hear the signal, which is trying to bury you in dittoheads if you're on AM.

As you get closer to the source of the transmission, you hear more of the announcer and less of the static, and finally you're close enough that you can't hear the static any longer, just a nice clear signal that, if you've got the FM tuned to NPR, asks you for more money.

It's similar with baselines, especially sales revenues and, to a lesser degree, units sold. The signal in the baseline is the result of all sorts of different influences, among them:



- » The size of, and the experience and skill in, the sales force
- » The increasing or decreasing desirability of the product, often due to its technology, its novelty, or its appearance
- » Its price, relative to that of the competition
- » Money spent on marketing, advertising in particular
- » Changes in social attitudes — for example, toward smoking and drinking

Fortunately, you don't have to identify all the influences on sales, list them, measure them, and then forecast them so as to get a forecast of what you're really after, which is sales revenues and units. All the influences combine in the signal and if you can get a reliable forecast of that signal, you're in good shape to name the value of the next period's sales.

So, how do you identify that signal? The idea is that over time, some special and unpredictable events — the *noise*, or the discrepancy between the signal and the actual results — averages out. Some unwelcome noisy event that pulls your actual revenues below the signal in June can be compensated for by other, welcome noisy events in July and August.

And moving averages take advantage of this. If you're getting a three-month moving average, those noisy events (the hit you take in June, the good luck in July and August) tend to average out, and the average for any given three months provides a better estimate of the signal than do the monthlies.

## A little less noise, please

Moving averages try to get the noise in the baseline to cancel itself out, and in that way to emphasize the signal. Where does the noise come from? There are many more sources of noise than of signal. Every one of your customers, every one of your sales reps, your production facilities, your distribution channel, possibly your customers' customers, is a source of noise that drags your actual results from the signal.

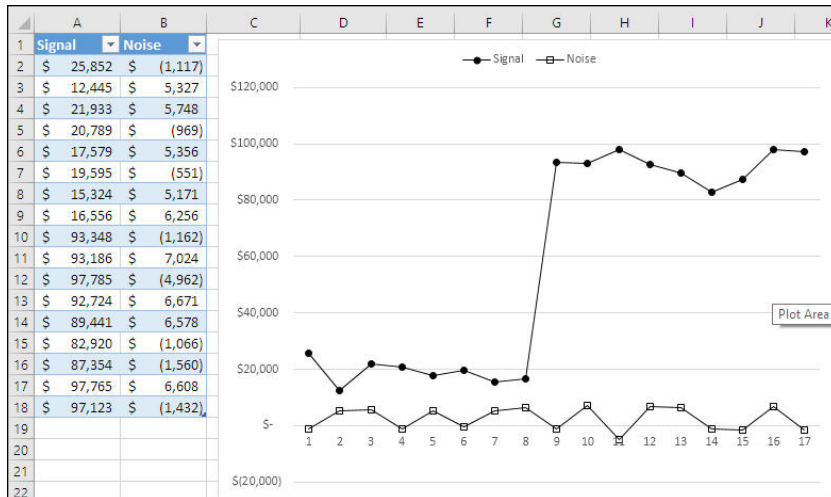
Suppose you're calculating a moving average that includes two months. Your first moving average might include January and February. The unseen and unseeable signal in those two months — the combined effects of the sales force, the product, the pricing, and so on — is \$100 million for January and \$110 million for February. But the noise steers the actual results off signal:

- » In February there's a bad flu outbreak that knocks out several of your sales reps as well as some buyers who normally purchase from the reps who are still healthy.

- » In December, a customer with a January-to-December fiscal year makes a large end-of-year purchase, to get the cost into its current fiscal year. Your company recognizes the revenue in January, which artificially swells the results for that month.
- » One of your competition's big customers gets upset by a price increase, starts shopping around, and decides it likes your pricing structure better, even though your product line isn't quite as hot. Your January revenues get a boost.
- » A moderate earthquake hits your San Andreas assembly and kitting site, disrupting production for a couple weeks during last year's fourth quarter. You can't fulfill existing orders on time and January revenues fall in consequence.

Notice that neither you nor your company can predict or control any of these events — perhaps apart from questioning the wisdom of locating a production facility next to an earthquake fault line. And that's the definition of *noise* — or, if you prefer, residuals or error. Noise is unpredictable and uncontrollable. But it may be measurable. Its long-term expected value is zero. Figure 13-1 shows an example of how this works, in theory at least.

**FIGURE 13-1:**  
This is what you'd like to see, if you could see it. You never *know* the exact value of the signal or the noise.



TIP

This is one reason it's a good idea to look at a chart of residuals. Unlike the Data Analysis add-in's Regression tool, neither the Moving Average tool nor the Exponential Smoothing tool calculates residuals, and therefore cannot chart them. See "Charting residuals," later in this chapter, for the steps to take to create a residuals chart from a moving-average analysis.

In Figure 13-1, you see what you never see in practice: the true values of the signal and the noise. The best you can really do is:

- » Calculate a moving average using your actual results.
- » Treat the moving average as if it were the signal — this is the whole point behind getting moving averages, after all: to estimate the signal as best you can.
- » Treat the difference between the actuals and the moving average as if it were the noise.



REMEMBER

As several other chapters in this book mention, the difference between the forecast and the actual is the error, or the residual. Noise is just a more general term.

You like to see noise in the baseline like that shown in Figure 13-2. The chart shows that it forms a horizontal data series. And, in cell B20, you can see that its average value is \$(489). In the context of actuals that are in the tens and even hundreds of thousands of dollars, \$(489) is effectively zero.

**FIGURE 13-2:**  
In this case, the actuals track the signal very closely, because the errors are so small.

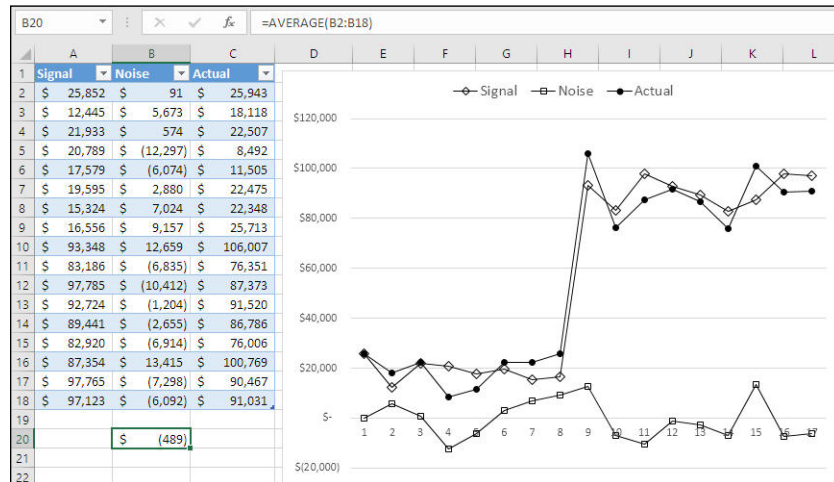
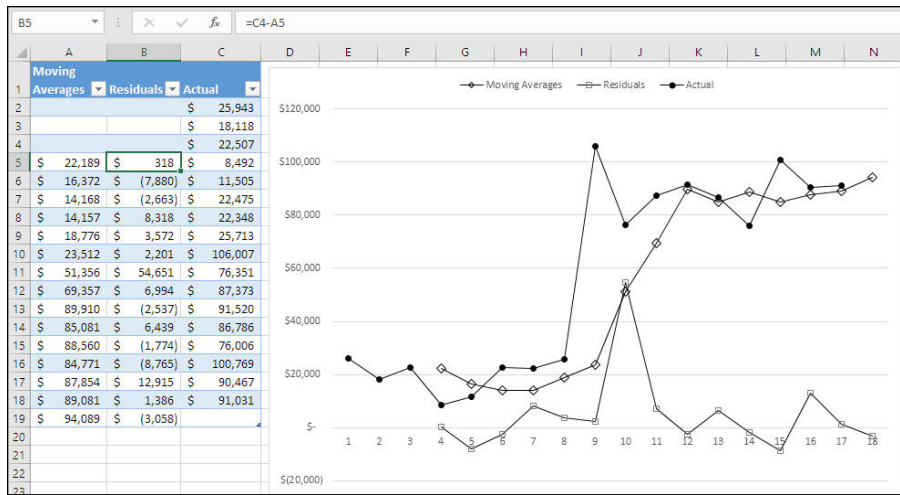


Figure 13-2 shows how the signal and the noise combine to form the actuals.

Again, you never really see the signal, you just do your best to estimate it with your forecast formula. I'm no more omniscient than you are, so I made up the signal and the actuals (and, by subtraction, the noise) shown in Figures 13-1 and 13-2. Figure 13-3 shows a more realistic situation.

**FIGURE 13-3:**  
Here, the chart estimates the signal by way of moving averages, and calculates the residual noise by subtracting the moving average from the actual.



Why doesn't the chart shown in Figure 13-3 look quite the same as the chart in Figure 13-2? Because Figure 13-2 is a fantasy: It acts as though you can know what the signal is. You can't ever know that. You can only estimate it — here, by using moving averages — and this example comes reasonably close.

The big discrepancy is at the ninth data point. The actual and the signal are very close in Figure 13-2, but the actual and the moving average are a ways apart in Figure 13-3. That's because the moving averages in Figure 13-3 are based on the prior three actuals, and those actuals drag the estimate down to \$23,512 rather than the Figure 13-2 hypothetical signal value of \$93,348.

## Stepping it up

The actuals in Figures 13-1 through 13-3 show a phenomenon that moving averages (and also exponential smoothing) deal with reasonably well. Notice the big jump in the value of the baseline at the ninth data point. In the terminology of forecasting, this is called a *random shock*. (In some contexts, it's called a *step function*.) Although the term *random shock* may make you think of something completely unplanned, it actually means a change in the level of a baseline that persists over time — intended or not. That's what you see in Figure 13-3, starting at the ninth time period.

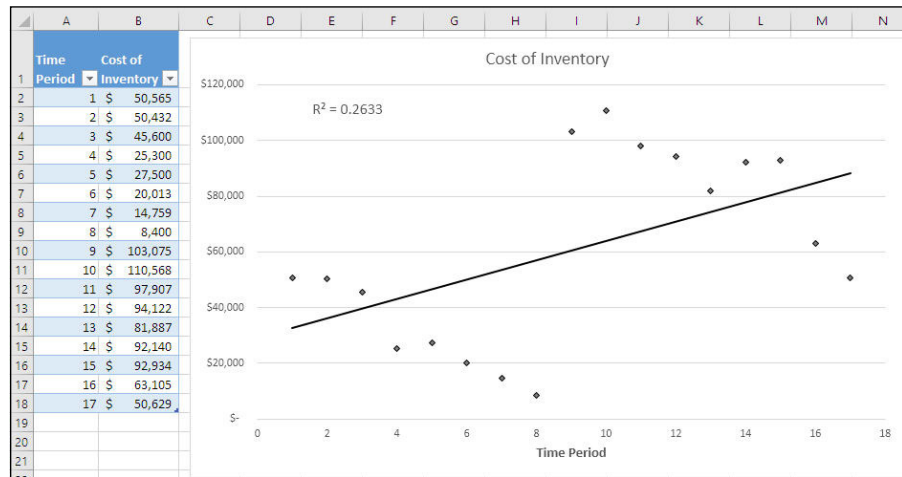
A good example of a random shock is a purchase to inventory. Perhaps your Product Management group has an opportunity to buy a large supply of materials or finished goods at a discounted unit price. The group does so, and now — if you maintain your current selling price — your company's profit margin per unit will increase. You could even discount your selling price a little, and as long as the

discount you offer is less than the discount you obtained, you'll still increase your profit margin.

But there are costs involved in that decision. There's the time cost of money, for one thing. For another, there are additional carrying costs, costs that might continue for as long as you maintain that additional inventory, depending on how you warehouse. Your sales force may need quite some time to sell down that inventory.

And that's a characteristic of random shocks: They tend to persist. Not always — some sudden increases or decreases in the level of a baseline die out rapidly, but more often they die out gently or persist for quite some time. In Figure 13-3, you see how a moving average tracks against a random shock; see Figure 13-4 for an example of why you'd want to use moving averages rather than regression for your forecast.

**FIGURE 13-4:**  
A regression forecast is seldom a good choice for a baseline that has experienced a random shock.



The best fit you can get using regression on this baseline has an R-squared value of 0.26. That means you can associate only about one fourth of the variability in the inventory costs with the time period.

And even that's misleading. Both before and after the random shock, the inventory costs are declining. But the math underlying the regression equation takes account of the sudden increase at the ninth time period, and that outweighs the declines starting at period 1 and at period 10. Linear regression analyses *always* returns a straight line — that's why they're called *linear*. Multiplying a variable such as time period (1, 2, 3, . . .) by a constant value such as a regression coefficient inevitably results in a straight line.



TECHNICAL  
STUFF

This is not to say that you can't do what's called *curvilinear* regression using Excel. You can. You just need to build an exponential component into your predictor variables. In a multiple regression context, for example, it might be right to predict using both the number of the time period (1, 2, 3, . . . 50) and the square of the number of the time period (1, 4, 9, . . . 2,500). That's just an example, one of thousands of possible examples. There's no special reason to use the square (or the cube, for that matter) of a time period's number as a predictor.



REMEMBER

Remember to chart your data early during the forecasting process. There's a lot that a chart will reveal easily that a glassy-eyed stare at a bunch of numbers will hide. And if you see a random shock like the one this section discusses, remember to think about using a moving-average (or an exponential-smoothing) forecast model.

## Reacting Quickly versus Modeling Noise

Figuring out how many actual baseline values to include in a moving average isn't easy. Some students of statistics ask similar questions when they ask their professor in Statistics 101, "How big a sample should I take?" And the prof, who as yet has no information at all about the study that the student wants to do, and who wants to dismiss the class, says "Thirty. Your sample size should be 30."

A ridiculous answer, of course, but you'd be surprised at how many people who've taken a couple of statistics courses believe that the minimum size for a sample is 30 observations.



TECHNICAL  
STUFF

There's a silly reason for that criterion. When you're building a frequency distribution and find that it's beginning to resemble a normal curve, you find that the curve begins to get steeper than 45 degrees by the time that you've got around 30 data points in the distribution.

You need to know quite a bit about the variables you're analyzing before you can estimate the right sample size. The point is that you can't just pull a number — 2, 3, 4, 5, whatever — out of the air and decide that's how many baseline values will go into each moving average.

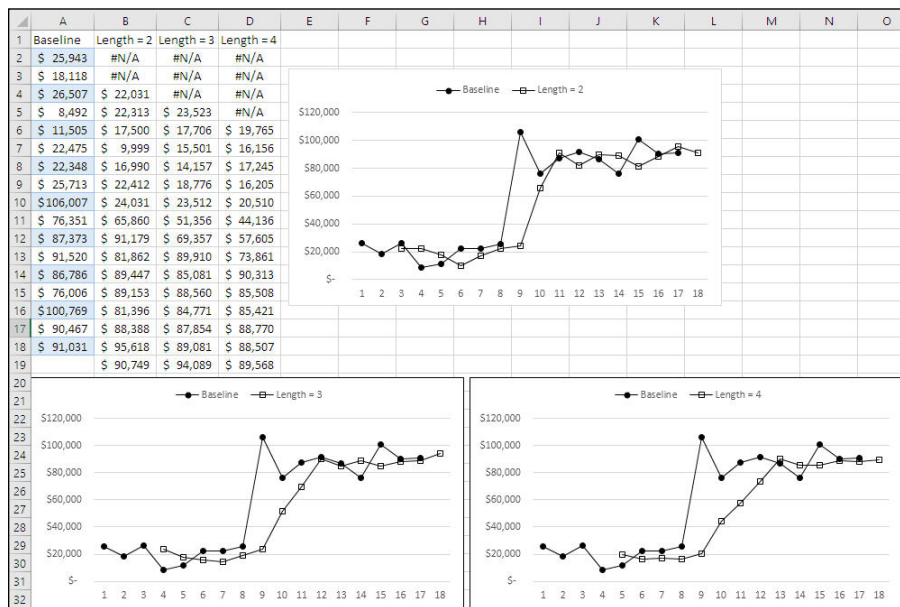
Choosing a number of baseline values is a balancing act. You're trying to achieve the right mix of reacting to changes in the level of the baseline (also called *tracking*) and letting errors average out (also called *smoothing*).

# Getting a smoother picture

The number of baseline values you put in each moving average affects how fast the moving average reacts to changes in the baseline. The fewer the baseline values in each moving average, the faster it reacts. Of course, there are two extremes:

- » **The smallest number of values in a moving average is 1.** Then the moving average reacts extremely fast — in fact, it's identical to the baseline. Every moving average of length 1 has one value from the baseline in it.
- » **The largest number of values in a moving average is all the values in the baseline.** Then you'll have only one moving average, and you can't get any smoother than that. But it makes for an unwieldy forecast, and it has little more to tell you than does the single-value moving average.

Figure 13-5 shows visually how slow smoothing versus fast tracking works.



**FIGURE 13-5:**  
The fewer the values in a moving average, the faster it tracks.

Look first at the chart that shows the baseline and the moving averages of length 2. It tracks quite closely to the baseline. It reacts quickly to changes — look particularly at period 10, when the baseline and the moving average are separated by about \$10,500, after having been about \$80,000 apart at period 9.

The chart with moving averages of length 3 has a smoother moving-average line than the length 2 moving-average line. Smoothness itself is not the goal; coming closer to the signal is. Although you can't know for certain what a line depicting the signal would look like, you do know that a length 3 moving average gives random noise more of a chance to average out than does a length 2 moving average.

Also in the Length = 3 chart, notice period 10. Here, the baseline and the moving average are separated by almost \$25,000: The moving average does not catch up with the actuals as fast as in the Length 2 chart. The difference between the Length 2 and the Length 3 charts is that the moving averages with a length of 2 react faster:

- » When the length is 2, only two values go into the average. At the tenth time period, that means \$25,713 and \$106,007, or \$65,860.
- » When the length is 3, there are three values in the average: \$22,348, \$25,713, and \$106,007, or \$51,356.

So putting a third value in the moving average, one that is relatively small, pulls that moving average down away from the baseline at point 10. This is why the more baseline data points you put in a moving average, the slower it reacts to changes in the baseline. You're putting more data points that precede the change in the baseline into the moving average. This slows the reaction time down, and at the same time smooths the moving-average line.

Finally, look at the Length = 4 chart. The line of moving averages reacts even more slowly to changes in the baseline. At the benchmark I've been using, point 10 on the chart, the difference between the baseline and the associated moving average is \$32,215, compared to about \$25,000 at length 3, and about \$10,500 at length 2. It's the smoothest of the three moving-average lines, and the slowest to react to changes in the baseline.

## Calculating and charting moving averages

I've left a few points implicit so far, and it's time to make them explicit. They have to do with which baseline values to put in a moving average, and how to line up a table's values to make the chart as clear as possible.

### Using a constant number of points



TIP

It may seem obvious, but a lot of people get it wrong: Each moving average must have the same number of baseline points as every other. If one takes the average of two baseline points, the rest of them must do the same. If one moving average involves three baseline points, so must all the others.



## Which values go into the forecast?

A moving-average forecast for a given time period involves the *previous* time periods. So, for a moving average of length 3, the forecast for April includes January, February, and March; for May, it includes February, March, and April. This makes good sense, of course. It's illogical for May's moving average to include May's baseline value — done that way, when it comes time to calculate May's moving average, you would already know May's actual.

## Where does the forecast go on the worksheet?

The forecast — the moving average — is typically in the row immediately following the final baseline data point that the average uses. In other words, it's in the same row as the time period that it forecasts. That sounds trivial, of course, but its effect is to extend the column of moving averages one row past the final baseline figure. That's your forecast for the period you haven't seen yet.

## You lose averages at the start of the moving average

In Figure 13-5, notice that the more baseline data points go into a moving average, the more moving averages you lose at the start of the series. In a moving average of length 4, you “lose” the first four periods in the moving-average series.

You can't calculate the first four moving averages if you're using Length = 4. Suppose in Figure 13-5 that you wanted to put a moving average in cell D5. That would be the average of the values in A1:A4. But you can't include the text value “Baseline” in the average; Excel just ignores it, and gives you the average of the values in A2:A4. This is equivalent to putting a different number of values in different moving averages — which is breaking the rules.

So, in a length 4 moving average, you start in the same row as the fifth baseline value. In a length 3, you start in the baseline's fourth row, and in a length 2 you start in the baseline's third row. So doing makes the structure of the worksheet match the structure of the logic, and makes it easier to match up the appropriate time periods on a chart.

# Using the Data Analysis Add-in to Get Moving Averages

The Data Analysis add-in's Moving Average tool is a handy way to get a moving average quickly. You can also get a chart of the baseline and the moving average.

It has a serious downside, though. As you'll see, it doesn't chart the moving averages correctly.

## Using the Data Analysis add-in's Moving Average tool

With the Data Analysis add-in installed in Excel, and a baseline on the active worksheet, follow these steps:

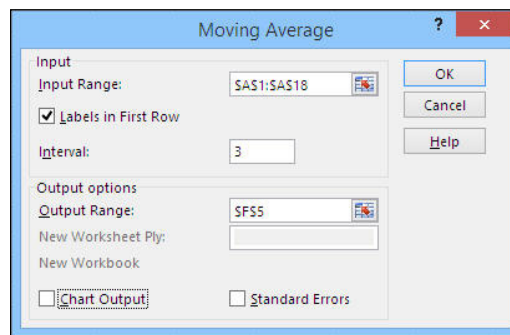
1. **Go to the Ribbon's Data tab and click Data Analysis in the Analyze group.**

The Data Analysis dialog box appears.

2. **Scroll down the Analysis Tools list box if necessary until you find the Moving Average tool, and then click on it.**

3. **Click OK.**

The Moving Average dialog box appears (see Figure 13-6).



**FIGURE 13-6:** Although you can see them, the New Worksheet Ply and New Workbook options are disabled in the Moving Average dialog box.

4. **Click in the Input Range box and drag through your baseline.**

If you've used an Excel table structure so that your baseline has a variable name in a cell at its top, include that also. In Figure 13-6, that's cell A1.

5. **If you included a variable name in the input range, select the Labels in First Row check box.**

6. **If you want a moving average of length 3, enter 3 as the interval.**

The Moving Average tool refers to the number of baseline data points in a moving average as the *interval*.

## 7. Click in the Output Range box.

You're in luck. The Moving Average tool doesn't ask you to choose among three output locations. Therefore, it does not unexpectedly snap you back to the Input Range box as soon as you've made your choice, as it does with Regression and Exponential Smoothing — and others, such as Correlation.

## 8. With the cursor still in the Output Range box, click in the cell where you want the moving average series to begin.

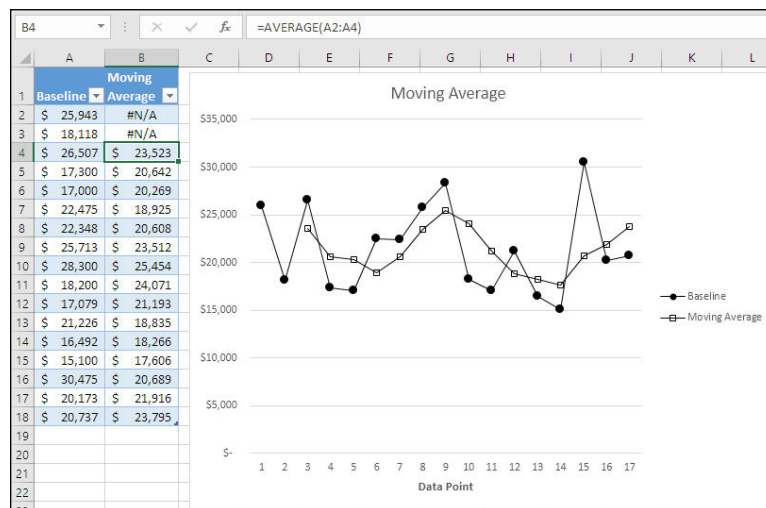
Let the way you want the worksheet to appear be your guide. The location of the results won't make a difference in the appearance of the chart. Although you'll want to make the correction that I suggest later in this section, you'll usually want to start the output in the row where the baseline starts, and one column to its right. See Figure 13-7.

## 9. Select the Chart Output check box and click OK.

The Moving Average tool takes over, calculates the moving averages, and charts both the baseline and the moving averages. (I've adjusted the size of the chart so it's easier to see what's going on.)

Notice that in Figure 13-7, the Moving Range tool has aligned the first moving average on the chart with the *third* baseline observation, rather than, correctly, with the fourth. That's also true of the way the moving ranges are aligned with the baseline: The first moving range is in row 4, the same row as the third baseline observation. But with a moving range of length 3, the first moving range should be aligned with the fourth, not the third, baseline observation.

**FIGURE 13-7:**  
I chose an interval of 3 (that is, a moving average of length 3) and an output range starting in \$B\$2 for this analysis.



Was that my fault? Shouldn't I have started the output range one row farther down, in B3?

Yes and no. Starting the output range in B3 rather than B2 would have caused the moving ranges to align properly in the worksheet cells. However, that wouldn't have helped with the chart.

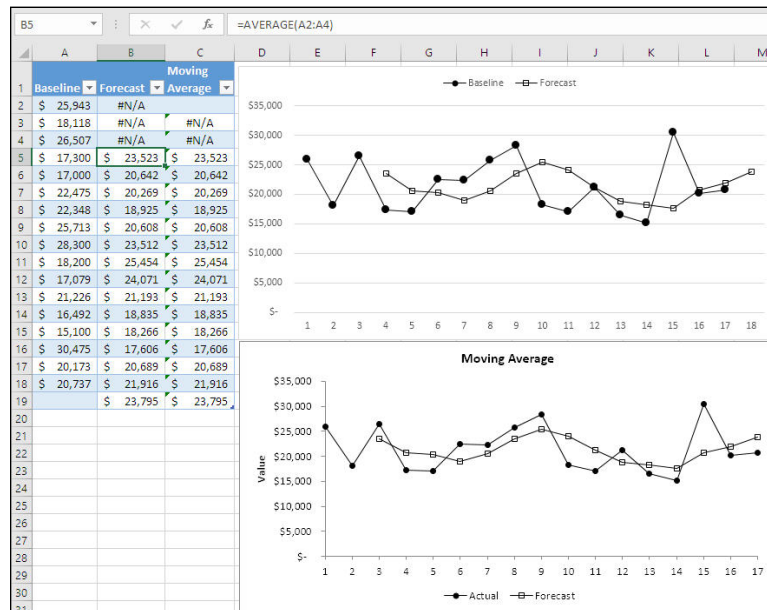


The way that two different data series are aligned on the worksheet often determines how they line up on a chart. But that's not the case here, where the range of the chart's x-axis values is not specified in the charted data series.

In Figure 13-8, column B, I entered the formulas to calculate the moving ranges by hand. It didn't take much time at all: I typed one formula:

```
=AVERAGE(A2:A4)
```

**FIGURE 13-8:** The upper chart and column B were created by using the keyboard and the mouse; the lower chart and column C were created by the Data Analysis add-in's Moving Average tool.



in cell B5 and autofilled it down through B19.



TIP

Here's how to save time by autofilling. In this case, after entering the formula in B5, I reselected B5. When a cell is selected, a small box replaces the cell's lower-right corner, and that box is called the *fill handle*. I moved my mouse pointer over the cell's fill handle. The pointer changed from a cross to crosshairs. I pressed and

held down the left mouse button, dragged into B19, and released the mouse button. That sequence copied the formula from B5 and pasted it into B6:B19, adjusting the addresses in the formula accordingly.



TIP

To save even more time by autofilling, *double-click* the cell's fill handle. Excel looks to see if there are contiguous filled cells in an adjacent column or row. If so, it autofills the contents of the selected cell down or across to the end of the filled cells in the adjacent range.

Figure 13-8 contains another range of moving averages, in column C and labeled *Moving Average*. They were entered by the Data Analysis add-in's Moving Average tool. To correct the placement of the moving averages as shown in Figure 13-7, I moved the output range down one row in the Moving Range dialog box; instead of having it start in the second row, it starts in the third. So, on the worksheet, the calculated moving ranges align properly with the baseline.

Not so in the chart, though. Regardless of where you tell the Moving Range tool to start the output on the worksheet, it aligns the moving range on the chart one row too high. With an interval of 3, the tool should have associated the first moving range value on the chart with the fourth baseline value.



REMEMBER

What the Moving Range dialog box calls an *interval* is what most forecasters — and this book — call a *length*, as in “a moving range of length 3.”

So if you're going to use the Data Analysis add-in's Moving Average tool to create the moving averages on the worksheet, and the chart of the baseline and the moving averages, you need to open the chart and click on the data series that represents the moving averages. You see something like this in the Formula Bar, immediately above the worksheet's column headers:

```
=SERIES("Forecast", , Sheet1!$C$3:$C$19, 2)
```

Change \$C\$3 to \$C\$2 and press Enter, so that the moving averages on the chart start one row higher than called for by the Moving Average tool. That pushes the charted data series one time period to the right, and now the baseline and the moving averages will line up correctly on the chart.

If you select the Standard Errors check box, you get a different standard error for each moving average. I don't want to get into a rant here, so I'll just mention that I don't subscribe to the notion that each moving average has a different standard error, so I don't ever select this check box when I'm seriously making a moving average forecast.

If you want, you can avoid the Data Analysis add-in entirely and still get moving averages into a chart. After you've charted your baseline (only), follow these steps:

- 1. Click the chart to select it.**
- 2. Right-click the charted baseline.**
- 3. Choose Add Trendline from the shortcut menu.**

A linear trendline, the default, appears on the chart and the Format Trendline pane appears in Excel's window.
- 4. In the Format Trendline pane, choose the Moving Average option button.**

Figure 11-10 shows how this looks in the Excel window.
- 5. The Format Trendline pane, in contrast to the Data Analysis add-in, terms the length of the moving average a *period*. Use the spinner to set the length to however many periods you want to include in each moving average.**
- 6. Click in the worksheet to dismiss the Format Trendline pane.**

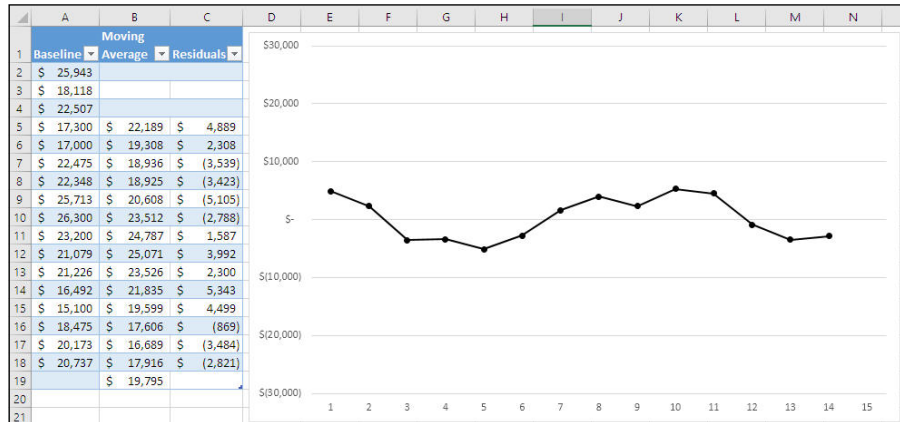
If you now click the Options tab in the Add Trendline dialog box, you see that the Moving Average tool doesn't give you the opportunity to forecast beyond the end of the baseline by means of a trendline. Furthermore, the trendline gives you no actual values on the worksheet — they can be useful for further analysis. But the Moving Average trendline can be a handy way to get a quick peek at the moving averages that would result from your baseline.

## Charting residuals

Charting the residuals from a moving average analysis is always a good idea if you think you're going to adopt the moving-average approach, and if you're reasonably happy with the results of choosing a particular length for your moving averages.

Figure 13-9 shows a chart of residuals.

The chart doesn't reveal any particular pattern to the residuals. They're distributed fairly evenly above and below their mean, which is quite close to zero given that the baseline is measured in the 10,000s.



**FIGURE 13-9:**  
The chart is of the residual values in column C.

You get the chart by taking the following steps:

- 1. With the baseline and the moving averages themselves in place, subtract a baseline value from an associated moving-average value.**

In Figure 13-9, for example, you could subtract cell A5 from cell B5 to get the first residual. Because the data is stored in a table, the formula in cell C5 appears in the Formula Bar as:

```
=[@[Moving Average]]-[@Baseline]
```

- 2. Autofill the formula down to the final row of the baseline.**

In a table, the autofill is done automatically. Unfortunately, the autofill occurs for the first three periods as well, so you'll need to replace them with #N/A values or simply delete the cells' contents.

- 3. Create a Line chart based on those residuals and examine it for any regularity in the residuals — for example, a straight line with increasing or decreasing trend.**

If you find regularity, you might want to consider a different forecasting approach, or a transformation of your baseline such as first differencing. (*First differencing* involves taking the difference between consecutive values in your baseline and analyzing those differences rather than the original values. Chapters 14 and 17 have lots of information on the differencing process.)





## Chapter 14

# Changing Horses: From Moving Averages to Smoothing

**F**or all their simplicity and intuitive appeal, moving averages have baggage. One of the problems comes with a short baseline (it's amazing how many forecasting problems a long baseline solves). Even if you choose to include only two actuals in each moving average, you lose two observations from your forecasts. Choosing a shorter or smaller length for the averages is a balancing act between tracking and smoothing, but it's also a choice of how much data you're willing to part with.

In forecasting, the notion of correlation is usually connected to regression forecasts, because correlations are the building blocks of any regression analysis. But moving averages and exponential smoothing also have to do with correlation, a special kind called *autocorrelation*. Before you can think sensibly about autocorrelation, though, you need to get a basis by looking at garden-variety correlation, and you find that basis in this chapter.

With correlation as background, the final major section in this chapter goes into autocorrelation: how it can get in the way of a good moving-average forecast, how to calculate and diagnose it, and how to make it go away and leave your forecast alone.

## Losing Early Averages

One of the problems with forecasting by means of moving averages is that you lose the opportunity to make early forecasts. The reason, when you think about how moving-average forecasts work, is really pretty clear.

A forecast based on a moving average is the average of two or more *prior* time periods. So a moving average of length 2 is the average of the two prior observations, whether the observations are sales results or traffic accidents. Using a moving average of length 2, the sales revenue forecast for March would be January revenue plus February revenue, divided by two.

If you're going to average two results to forecast the next one, you can't get a forecast for the second period. There's no result from time zero to average with the first result. The same is true with a moving average of length 3: The first forecast you can get is for the fourth period.



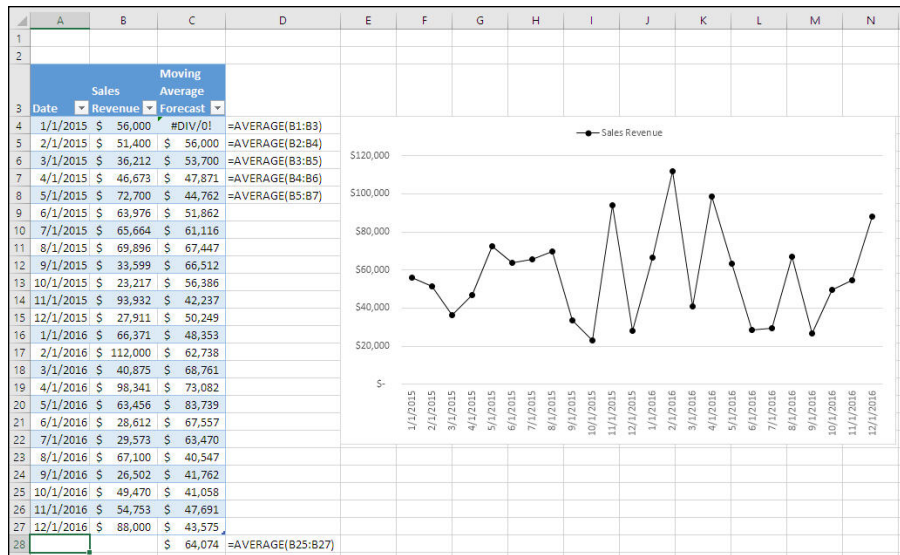
REMEMBER

The word *forecast* isn't limited to a future value that hasn't yet been observed. It also means a value that's forecast for a period in the past. You might get the average of January, February, and March to forecast April, even though you're into the next December and you've known about April for eight months now. You use those forecasts for prior periods to evaluate how well your forecasting methods work.

This sort of thing is easier to see on a chart. Figure 14-1 has an example.

In Figure 14-1, the first legitimate moving average is in cell C7. It's the average of the actual revenues shown in cells B4:B6. The earlier attempts at calculating moving averages are:

- » **Cell C6:** As you can see from the formula, this is the average of the values in cells B3:B5. But the value in B3 is text. The AVERAGE function ignores text values, so C6 shows the average of B4:B5. The average of two months of revenue isn't *wrong* so much as *different* from the later moving averages, and comparing them directly is misleading.
- » **Cell C5:** This formula attempts to get the average of B2:B4. There are two text values in those cells, which AVERAGE ignores, so it just returns the "average" of the first month of revenue.



**FIGURE 14-1:** You could get an additional legitimate forecast in cell C6 by shifting to two-period averages.

» **Cell C4:** By now I've run out of numbers, and when AVERAGE has no numbers to work with it returns #DIV/0!, which is an error value. It informs you that, with no numbers to average, Excel has tried to divide a sum by zero. This is a mathematical error, of course, not a problem with moving averages as such.

The main point is that you lose as many forecasts from the start of your baseline as you have actual values that contribute to a moving average. The same is not true of the end of the baseline, because you're not usually running out of prior actuals.

You can run out of prior actuals, though, if your baseline is too short for the length of the moving average. A baseline with two actuals in it can't support a moving average of length 3. Wait for more data to come in before you start forecasting.

You can improve things a little if you switch to a smaller number of periods per moving average. In this case, you could choose two periods rather than three, and that would get you one extra forecast at the start. But then the forecasts might start to track the noise as much as the signal. In general, the fewer the periods that comprise a moving average, the more closely the moving average tracks the actuals and fails to smooth out the noise in the baseline.

In sum, moving averages are straightforward to calculate and to understand. They help you distinguish the signal from the noise in a sequential series of observations, such as a year's worth of sales measured on a monthly or weekly basis. But moving averages have drawbacks such as the loss of early observations and the fact that only a (usually, small) subset of the available data contributes to each moving average.

Besides moving averages, the two main approaches to sales forecasting that I discuss in this book are regression and exponential smoothing. Those two techniques can get you good, solid forecasts without all the drawbacks that come with moving averages. (In fact, the main reason that I have discussed moving averages at all is that the topic provides a good foundation for understanding the smoothing techniques.)

But correlation between different segments of your baseline has an effect on how forecasts behave, whether you're using smoothing or regression for a given sales projection. The remainder of this chapter provides a flyover of what those correlations are about, and then I get into the nuts and bolts of exponential smoothing and regression in Chapters 15 and 16.

## Understanding Correlation

Correlation is a fundamental part of forecasting. You can do forecasting without knowing the first thing about correlation, but you handicap yourself if you don't bother. Correlations are key to understanding regression forecasts, and they play an important part in diagnosing how well your smoothing forecasts work. Better yet, correlations aren't really tough to understand.

Do me a favor: At least scan this chapter's material on correlation. If you decide it's not for you, okay, no problem — you can still do your forecasting, even if you don't have all the available tools at hand.

### When did they start going together?

You want to get your virtual hands on two variables. Let's start by assuming those variables are people's height and weight. Now, you know just from general life experience that the taller a person is, the more the person tends to weigh. It's not anything like one for one — people don't automatically weigh 2½ pounds more for every additional inch taller they are. But there is a strong tendency for height and weight to go together. It's not easy to keep them apart.

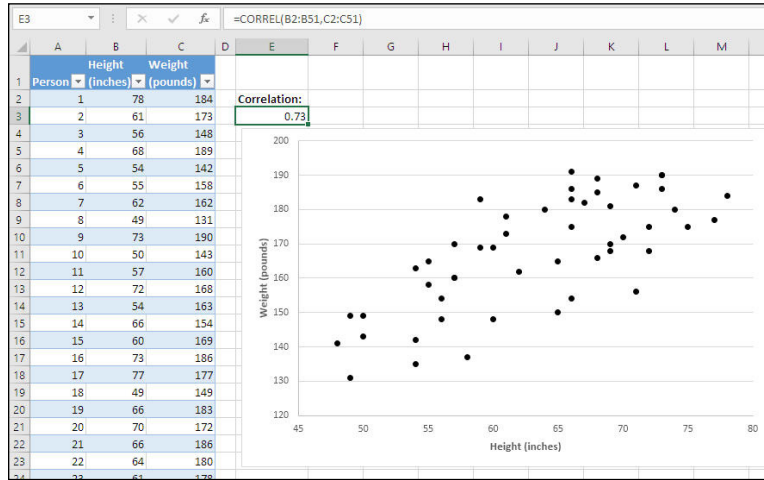
Suppose you decide to do some research. You stand on a street corner in the business district of a large metropolitan area and stop people who are walking by. You engage each person in conversation and while you've got him distracted, you get his height with a tape measure. Then you ask him how much he weighs. Some tell you, and you note down the height and weight. (Some of them just walk away, and that's why you have some missing data.)

After you collect data on about 50 people, you head back to your office and put the data in an Excel worksheet. It looks like the one in Figure 14-2.

	A	B	C	D
		Height	Weight	
1	Person	(inches)	(pounds)	
2		1	78	184
3		2	61	173
4		3	56	148
5		4	68	189
6		5	54	142
7		6	55	158
8		7	62	162
9		8	49	131
10		9	73	190
11		10	50	143
12		11	57	160
13		12	72	168
14		13	54	163
15		14	66	154
16		15	60	169
17		16	73	186
18		17	77	177
19		18	49	149
20		19	66	183
21		20	70	172
22		21	66	186
23		22	64	180

**FIGURE 14-2:** You can't tell just by looking at the numbers that there's a fairly strong relationship between height and weight.

So far, it's just a jumble of numbers. When you're confronted by a numeric jumble, the first thing to do is put it in a chart, like the one in Figure 14-3.



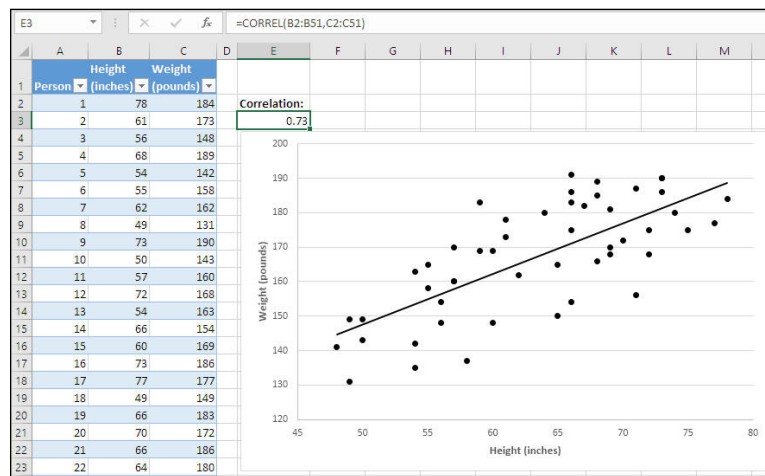
**FIGURE 14-3:** This is an XY (Scatter) chart, the best chart type for showing two numeric variables.

Each point in the chart represents a different person you talked into letting you get a height and weight measurement. If you pick out one point and look over to the vertical axis, you can see what that person's height is. And if you look down from that point to the horizontal axis, you can see what the weight is.

Now the jumble starts to resolve into some patterns:

- » The points on the chart that are higher up the vertical axis also tend to be farther along the horizontal axis.
- » The points describe a sort of cigar shape, running from the lower left to the upper right.
- » The points do *not* lie directly on a straight line, but you can imagine one running through the middle of the cigar, as in Figure 14-4. Or you can draw it: Right-click a point in the charted data series and choose Add Trendline from the shortcut menu.

**FIGURE 14-4:**  
Lower left to  
upper right  
means a positive,  
or *direct*,  
correlation.

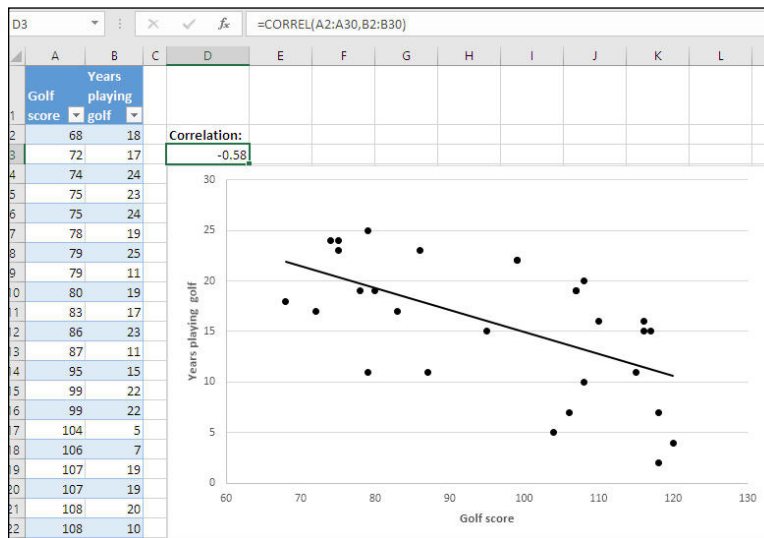


You can make some statements about the relationship between two variables:

- » The closer those points come to lying on the imaginary straight line, the stronger the relationship between the two variables.
- » You can express the strength of the relationship with a number. It turns out that, because of the way it's calculated, the number must be between  $-1$  and  $+1$ . That number is called the *correlation coefficient*.
- » If the correlation coefficient is positive, like  $0.6$ , then the imaginary straight line runs from the lower left to the upper right. If the correlation coefficient is negative, then the line runs from the upper left to the lower right.
- » The closer the correlation coefficient is to  $+1.0$  or  $-1.0$ , the stronger the relationship. The closer it is to zero, the weaker the relationship.

Suppose you analyzed the golf scores of 100 golfers, whose skill levels range from beginner to expert. You could put them on a chart, just like height and weight, with, say, their scores on the vertical axis and their years of golfing experience on the horizontal axis (see Figure 14-5).

**FIGURE 14-5:**  
Upper left to  
lower right  
means a negative,  
or *inverse*,  
correlation.



So you see that the more the years of experience playing golf, the lower the golf score. The fact that the correlation is negative has nothing to do with the *strength* of the relationship — just with its direction.

Using Excel, you can calculate the correlation coefficient between two variables very easily. Just use the CORREL function, with two ranges of data values as its arguments. In Figure 14-5, the formula used in cell D2 is:

```
=CORREL ( A2 : A30 , B2 : B30 )
```

## Charting correlated data

When you use Excel to chart correlated data, it's almost always best to use an XY (Scatter) chart.



For simplicity, I refer to this as an XY chart. Excel uses the term *Scatter* in the name because earlier applications — and this goes back to mainframe days — referred to this kind of chart as a *scatter chart* or *scattergram*.

The main reason that an XY chart is best is that, if you're working with correlations, you're automatically working with variables that are entirely numeric. Height and weight are both numeric variables. Golf score and years of experience are both numeric variables.

When you chart them, you want a chart that has a numeric horizontal (or X) axis, and a numeric vertical (or Y) axis. Other chart types don't offer this arrangement. For example, a Column chart assumes that categories (such as make of car or political affiliation) are on its horizontal axis, and numbers (such as number of cars sold or number of registered voters) are on its vertical axis.

One problem with putting numeric values on a category axis is that Excel doesn't reflect the magnitude of the difference between numbers in their spacing on the axis. So, the numbers 2, 4, and 8 would show up with equal distances between them on a category axis. But the difference between 2 and 4 is 2, and the difference between 4 and 8 is 4. Only a numeric axis can represent those differences accurately, and only an XY chart has two numeric axes to handle your two numeric variables.

There are other reasons to use an XY chart for two numeric variables. Among them:

- » Trendlines are calculated and drawn accurately.
- » R-squared values and regression equations are based on the proper values.

## Understanding Autocorrelation

Autocorrelation is a specific kind of correlation that's comes up frequently in forecasting. You interpret it just as you would standard correlation — as the strength of the relationship between two variables. The difference is that you're correlating one set of values with a different set of values of *the same variable*.

The previous section discusses the relationship between height and weight: how taller people tend to weigh more, and how shorter people tend to weigh less. When you're considering autocorrelation, you think in slightly different terms. You think of how earlier values in the baseline are related to later values in the baseline.

With height and weight, you're looking at how Smith's height relates to Smith's weight and how Anderson's height relates to Anderson's weight, and so on.

With autocorrelation and sales forecasting, you're looking at how January's revenue relates to February's revenue, how February's revenue relates to March's revenue, and so on.





TIP

This sort of effect is not limited to a month-to-month relationship. It can and does apply to daily, weekly, quarterly, and annual measures.

Frequently, an earlier time period's value carries forward into a later period. Suppose your company spifs a particular product line for a month. (*Spif* is short for Special Product Incentive Fund, or Sales Performance Incentive Formula, or something else, depending on whose commission plan you're reading.) The spif might take the form of a bonus, over the normal commission, for sales of that product during May. The effect of that spif sometimes carries forward into June, July, and August, or even longer.

A one-month spif on that vital automobile option, undercoating, can have effects on revenue in subsequent months, for a number of reasons — among them:

- » As the sales reps concentrate on selling undercoating, they gain more experience in drawing customers' attention to undercoating's undisputed effects on the longevity of the vehicle.
- » Satisfied customers refer their friends to the company's dealerships, and these new customers all ask for undercoating.
- » Some, perhaps most, of the transactions may be in the form of leases. If the company recognizes revenue only as it receives monthly payments, then the additional revenue due to the spif will carry forward through the life of each lease.

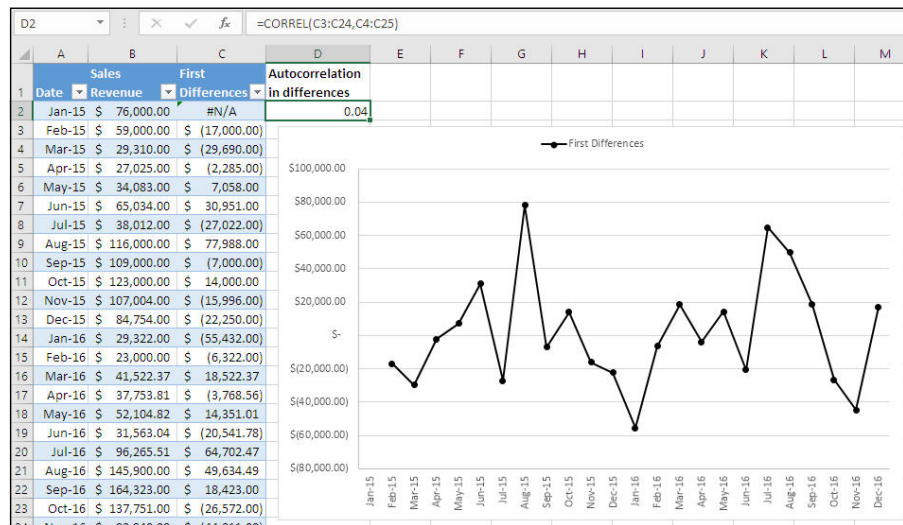
Different situations bring about different reasons that one time period's results can persist through later time periods. If you're selling a political candidate rather than undercoating — a distinction without a difference — there can be a real bandwagon effect. The higher a candidate rises in the polls, the more coverage provided by the media, the more the name recognition, and the higher the candidate rises in the polls.

- » **Autocorrelation usually means that the errors are correlated.** That is, the difference between the forecast value and the actual value for one period is related to the difference for another period. If you're using regression (rather than moving averages or exponential smoothing), this violates one of the assumptions that form the basis of a regression forecast: independence of errors.
- » **If you're using moving averages or exponential smoothing, autocorrelation (whether positive or negative) often causes your forecasts to be too high when the actuals are low, and too low when the actuals are high.**

In either case, the answer is often taking the first differences. You can find more information on taking differences in Chapter 17, but briefly:

- » You take first differences by calculating the difference between one actual value and a previous value, often but not always the immediately prior value.
- » With the autocorrelation removed by using the first differences, you're okay to go ahead and make your forecasts using those differences. You still have to account for the differencing when you make your forecasts in the original metric — Chapter 17 shows you how to do that.

The first differences in Figure 14-6 are calculated by subtracting the previous value from an actual value. For example, you get the difference in cell C3,  $-\$17,000$ , by subtracting B2 from B3.



**FIGURE 14-6:** No indication that the first differences are anything but independent of one another.

The autocorrelation in the differenced series, 0.04, is shown in cell D2. This correlation coefficient is very close to zero. (You see how to calculate an autocorrelation later in this chapter, in the section titled, well, “Calculating autocorrelation.”) That indicates that you’ve removed the autocorrelation in the series by taking first differences. And there’s no special regularity in the chart of the first differences — regularity can still occur when first differencing was not enough to dispose of the autocorrelation, and if so, you might want to do a second differencing.

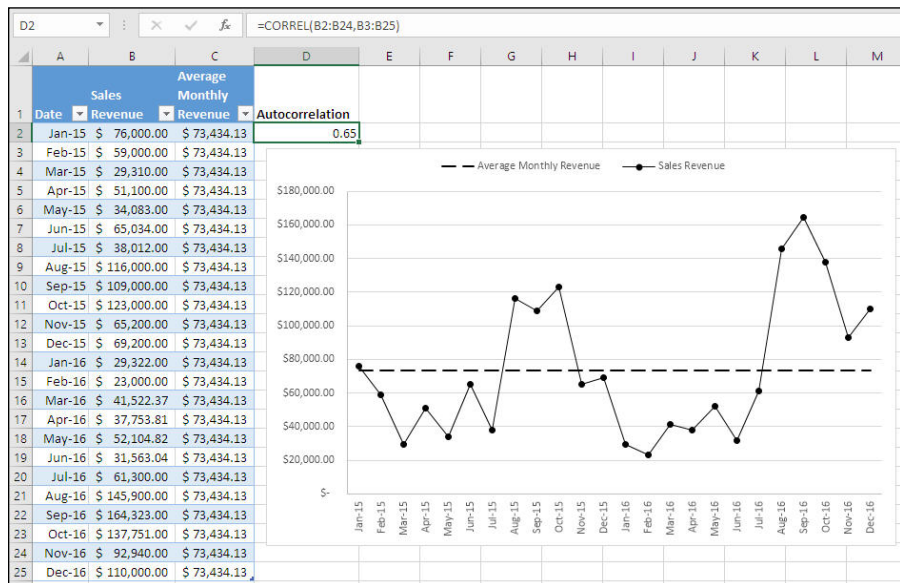
Compared to the original baseline in B2:B25 of Figure 14-6, the first differences in C3:C25 are what you want as the basis for your forecast. But there are two drawbacks to using first differences:

- » You lose the data for one time period. Notice the #N/A value in cell C2 of Figure 14-6. When you take first differences, you always wind up with one time period fewer than in the original baseline. With a good, long baseline — say, 50 time periods, just as a rule of thumb and with no statistical theory to back it up — the loss is trivial.
- » At first it's not intuitively clear what first differences represent, much less the forecasts of first differences. Don't worry about it: It takes some time and experience to get a feel for what's going on. And when you've un-differenced the data (you see how later in this section) you're back to the original scale and back on more comfortable ground.



The autocorrelation you see in cell D2 of Figure 14-7 is not technically an autocorrelation function as the term is used in forecasting — but it's close. In Chapter 16, I show you how to use VBA code to calculate what are formally autocorrelation functions.

**FIGURE 14-7:**  
Positive autocorrelations cause drift, as shown here. Negative autocorrelations make the baseline bounce up and down.



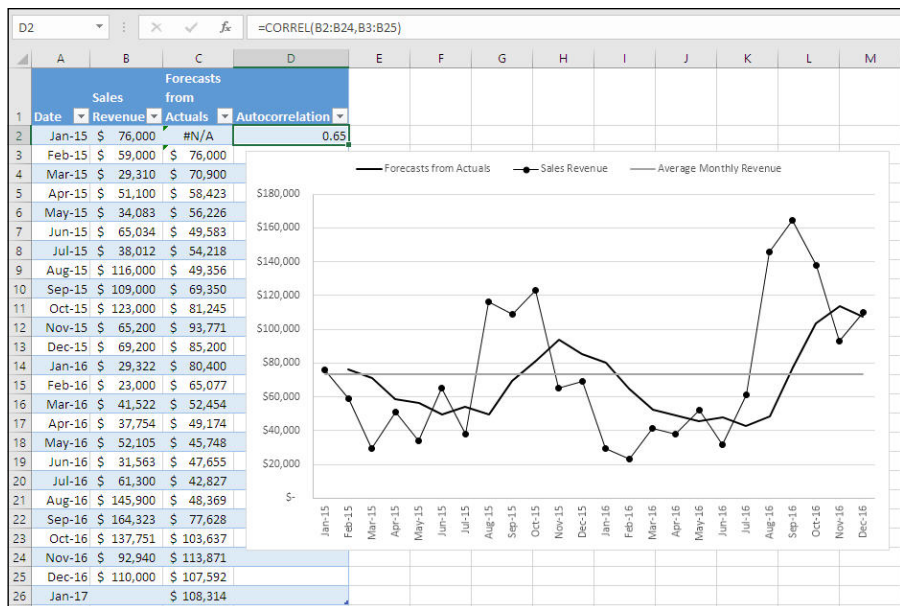
The baseline shown, and charted, in Figure 14-7 is typical of positive autocorrelations. The data drift below the average value for several periods, and then spend some periods above the average, and so on. This pattern can represent a cycle, but not usually a seasonal baseline.



REMEMBER

A cycle is different from a seasonal pattern because it often covers more than one year, and because the high and low points do not follow a regular schedule. A cycle such as the business cycle often spans several years, but you expect a seasonal pattern to repeat each year. A cycle might reach a high point in 2018, and then bottom out in 2019, only to climb gradually to another high point in 2022. But the sales of flip-flops peak every summer and crash every winter.

Figure 14-8 shows the forecasts that you get for the baseline in Figure 14-7, and they're charted along with the actuals. This forecast uses exponential smoothing rather than moving averages, to show autocorrelation in a different context. The chart also shows the average value of the actuals in the baseline.



**FIGURE 14-8:** The forecasts tend to smooth the noise out of the actual observations.



TIP

Using moving averages, each forecast averages a fixed number of preceding observations — typically, anywhere from 3 to 8. Using exponential smoothing, each forecast is a weighted average of all the preceding observations. This book begins a detailed look at exponential smoothing in Chapter 15.

There are 23 time periods that have both an actual and a forecast. In 19 of those 23 periods:

- » The forecasts are higher than the actuals when the actuals are below the average (13 cases).

» The forecasts are lower than the actuals when the actuals are higher than the average (6 cases).

As I mentioned earlier in this chapter, this pattern is typical of baselines with positively autocorrelated values. Because you want your forecasts to be as accurate as you can make them, you'd like to get rid of this source of errors.



TIP

You can see the average of the actuals in the chart in Figure 14-8, but you can't see their values on the worksheet. Nevertheless, they're on the worksheet. To avoid cluttering things up even worse than they are, I make the averages invisible. I select the range of cells that contains the average values (for what it's worth, that's E5:E25) and choose Format → Cells, select the Number tab, and choose Custom from the Category list box. In the Type box, I enter ;; (that's two consecutive semicolons). When I click OK, the average values vanish from view, although they stay put in their cells.

You can make the autocorrelation go away by taking first differences. Figure 14-9 illustrates the process, which is really a pretty simple one.

	A	B	C	D	E
1	Date	Revenue	First Differences	Autocorrelation (Actuals)	Autocorrelation (First Differences)
2	Jan-15	\$ 76,000		0.65	-0.10
3	Feb-15	\$ 59,000	\$ (17,000)		
4	Mar-15	\$ 29,310	\$ (29,690)		
5	Apr-15	\$ 51,100	\$ 21,790		
6	May-15	\$ 34,083	\$ (17,017)		
7	Jun-15	\$ 65,034	\$ 30,951		
8	Jul-15	\$ 38,012	\$ (27,022)		
9	Aug-15	\$ 116,000	\$ 77,988		
10	Sep-15	\$ 109,000	\$ (7,000)		
11	Oct-15	\$ 123,000	\$ 14,000		
12	Nov-15	\$ 65,200	\$ (57,800)		
13	Dec-15	\$ 69,200	\$ 4,000		
14	Jan-16	\$ 29,322	\$ (39,878)		
15	Feb-16	\$ 23,000	\$ (6,322)		
16	Mar-16	\$ 41,522	\$ 18,522		
17	Apr-16	\$ 37,754	\$ (3,769)		
18	May-16	\$ 52,105	\$ 14,351		
19	Jun-16	\$ 31,563	\$ (20,542)		
20	Jul-16	\$ 61,300	\$ 29,737		
21	Aug-16	\$ 145,900	\$ 84,600		
22	Sep-16	\$ 164,323	\$ 18,423		
23	Oct-16	\$ 137,751	\$ (26,572)		
24	Nov-16	\$ 92,940	\$ (44,811)		
25	Dec-16	\$ 110,000	\$ 17,060		
26	Jan-17				

**FIGURE 14-9:** Taking first differences usually makes a baseline stationary.

**1. Clear the forecasts from column C that were shown in Figure 14-8.**

If you're sure you don't have any other data in the column that you need to keep, the fastest way is to click the column header and press the Delete key.

2. Select cell C3, and enter =B3-B2, either by typing the whole thing or by selecting the cells with your mouse.
3. Right-click cell C3.
4. Choose Copy from the shortcut menu.
5. Select the range C4:C25.
6. Right-click cell C4 and choose the Paste icon from the shortcut menu.

You now have the series of first differences in column C, and you can use those first differences to make your forecasts. First, though, you should check to make sure that first differencing really did remove the autocorrelation.

That check is shown in cell E2 of Figure 14-9. The autocorrelation in the first differences is  $-0.10$ , which is small enough to be negligible. (You can see the formula for the autocorrelation in the Formula Bar.)

You're ready to make a forecast based on first differences. That's shown in Figure 14-10.

	A	B	C	D
	Sales		First	Smoothed
1	Date	Revenue	Differences	Differences
2	Jan-15	\$ 76,000	#N/A	#N/A
3	Feb-15	\$ 59,000	\$ (17,000)	#N/A
4	Mar-15	\$ 29,310	\$ (29,690)	\$ (17,000)
5	Apr-15	\$ 51,100	\$ 21,790	\$ (20,807)
6	May-15	\$ 34,083	\$ (17,017)	\$ (8,028)
7	Jun-15	\$ 65,034	\$ 30,951	\$ (10,725)
8	Jul-15	\$ 38,012	\$ (27,022)	\$ 1,778
9	Aug-15	\$ 116,000	\$ 77,988	\$ (6,862)
10	Sep-15	\$ 109,000	\$ (7,000)	\$ 18,593
11	Oct-15	\$ 123,000	\$ 14,000	\$ 10,915
12	Nov-15	\$ 65,200	\$ (57,800)	\$ 11,841
13	Dec-15	\$ 69,200	\$ 4,000	\$ (9,052)
14	Jan-16	\$ 29,322	\$ (39,878)	\$ (5,136)
15	Feb-16	\$ 23,000	\$ (6,322)	\$ (15,559)
16	Mar-16	\$ 41,522	\$ 18,522	\$ (12,788)
17	Apr-16	\$ 37,754	\$ (3,769)	\$ (3,395)
18	May-16	\$ 52,105	\$ 14,351	\$ (3,507)
19	Jun-16	\$ 31,563	\$ (20,542)	\$ 1,851
20	Jul-16	\$ 61,300	\$ 29,737	\$ (4,867)
21	Aug-16	\$ 145,900	\$ 84,600	\$ 5,514
22	Sep-16	\$ 164,323	\$ 18,423	\$ 29,240
23	Oct-16	\$ 137,751	\$ (26,572)	\$ 25,995
24	Nov-16	\$ 92,940	\$ (44,811)	\$ 10,225
25	Dec-16	\$ 110,000	\$ 17,060	\$ (6,286)
26	Jan-17			\$ 718
27				
28	Smoothing Constant	Damping Factor		
29	0.3	0.7		

**FIGURE 14-10:**  
You lose one forecast due to differencing, and another due to smoothing.

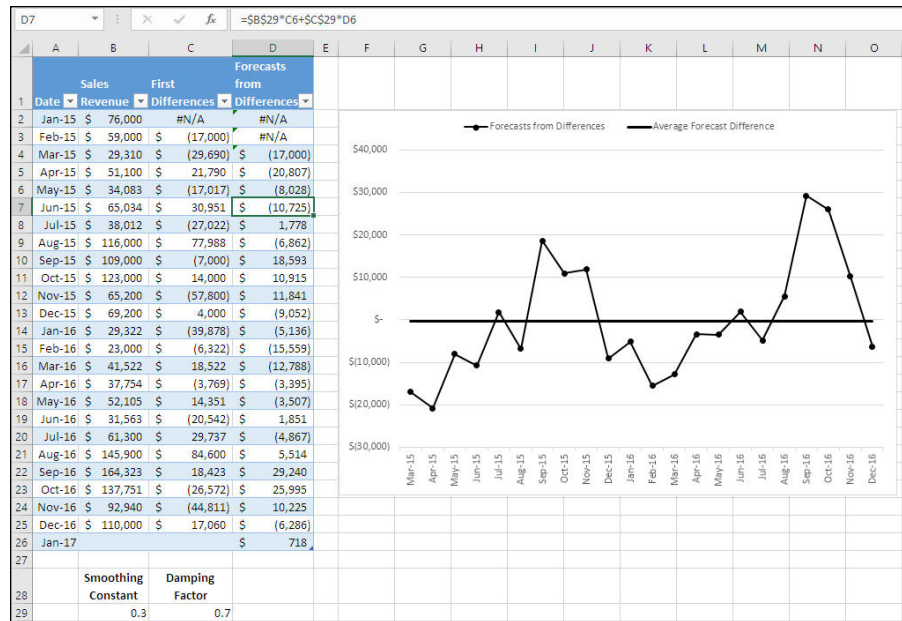
In Figure 14-10, the forecast differences appear in the range D4:D26. They were created by using exponential smoothing. The smoothing constant is in cell A29 and the damping factor is in cell B29 (these two constants are discussed briefly in

Chapter 2 and Chapter 10, and get more extensive coverage in Chapter 15). The formulas in D4:D26 use those addresses rather than constant numbers, so you can easily change the smoothing constant if you want to see the result.



REMEMBER

The forecast differences, those shown in D4:D26 of Figure 14-11, are just that: forecasts of the differences between the actuals. In each cell in the range D5:D26, we forecast what the difference between two specific actuals would be, if the prior forecast had been more accurate — this is the basic idea that underlies exponential smoothing. (As the first forecast, D4 is calculated differently from D5:D26.)



**FIGURE 14-11:** The drifts below and above the mean are much shorter than in the original baseline.

Finally, you're ready to integrate the forecast differences back into the original actuals. (Note: It's called *integration* but we're not talking calculus here, I promise.) This last step, along with a chart of the results, is shown in Figure 14-12.

You do the integration by adding the forecast difference to the associated actual value that preceded it. So, to get the first integrated forecast of \$42,000 in cell E4 of Figure 14-12, you would enter `=B3+D4` and copy and paste it down through E5:E26. Create the chart by following these steps:

1. Select A3:B26, and release the mouse button.
2. Hold down the Ctrl key and select E3:E26.

You should now see two ranges highlighted, one in columns A and B and one in column E.

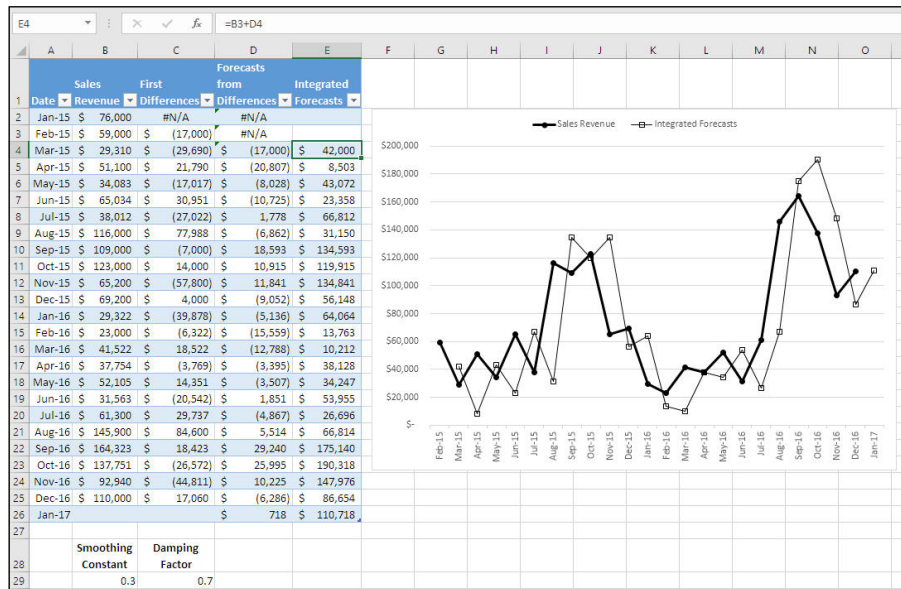


**3. Go to the Ribbon's Insert tab and choose the Line Chart in the charts group. Choose the subtype you prefer to create the embedded chart.**

Because the horizontal axis will represent the values in A3:A26, the time period dates, a Line chart is a good choice.

Notice in Figure 14-12 that the forecasts track the actuals much better than they do in Figure 14-8, where the forecasts are based on the actuals, with all that autocorrelation.

**FIGURE 14-12:**  
The integrated forecasts still show some lag but do not over- and underestimate as much as the forecasts based directly on the actuals.



**TIP**

Steps 1 and 2 advised you to select one range (in columns A and B) and then hold down Ctrl and select a third range (in column E). Whether you're selecting ranges or individual cells, this is the way to get a *multiple selection* in Excel. A multiple selection is two or more cells or ranges that are *not* contiguous — and here, the range E3:E26 is not contiguous to A3:B26, because columns C and D are in between. If you want to select a range of contiguous cells (such as A3:B26), click A3, hold down Shift, and then click B26. Using Shift means that, in this example, all cells between A3 and B26 are selected — so it's not a multiple selection.

One more thing (I feel like I'm wearing a rumpled raincoat, holding a stogie, and raising one hand over my head with my forefinger extended): Although this example used exponential smoothing to create the forecasts in Figures 14-8 and 14-12, you'd get roughly the same outcome if you used moving averages or regression.



I say “roughly the same” because moving averages, exponential smoothing, and regression always produce forecasts that are at least slightly different from one another. The point is that moving averages and regression can react to autocorrelation in ways that can mislead you. In particular, if you’re using regression, you often want to difference the series so that the forecast errors (the differences between the forecasts and the actuals) are independent of one another.

## Calculating autocorrelation

Figure 14-13 shows an example of calculating an autocorrelation coefficient.

	A	B	C	D	E
	Sales				
1	Date	Revenue			
2	Jan-15	\$ 76,000			
3	Feb-15	\$ 59,000		Autocorrelation	
4	Mar-15	\$ 29,310		0.71	
5	Apr-15	\$ 27,025			
6	May-15	\$ 34,083			
7	Jun-15	\$ 65,034			
8	Jul-15	\$ 38,012			
9	Aug-15	\$ 116,000			
10	Sep-15	\$ 109,000			
11	Oct-15	\$ 123,000			
12	Nov-15	\$ 107,004			
13	Dec-15	\$ 84,754			
14	Jan-16	\$ 29,322			
15	Feb-16	\$ 23,000			
16	Mar-16	\$ 41,522			
17	Apr-16	\$ 37,754			
18	May-16	\$ 52,105			
19	Jun-16	\$ 31,563			
20	Jul-16	\$ 96,266			
21	Aug-16	\$ 145,900			
22	Sep-16	\$ 164,323			
23	Oct-16	\$ 137,751			
24	Nov-16	\$ 92,940			
25	Dec-16	\$ 110,000			
26	Jan-17				

**FIGURE 14-13:**  
The CORREL function can handle overlapping ranges.

The autocorrelation of 0.71 shown in Figure 14-13 is between the values in cells B2:B24 and the values in cells B3:B25. The formula in cell D3 is:

```
=CORREL ( B2 : B24 , B3 : B25 )
```

Notice the similarity between this formula and the one used in Figure 14-5:

```
=CORREL ( A2 : A30 , B2 : B30 )
```

In each case, you're using the CORREL function to calculate a correlation between two different worksheet ranges. The only real difference between the two formulas is that in the case of autocorrelation, you're getting the correlation between the first through the 23rd values, and the second through the 24th values of the same variable.

It can be easier to see what's going on by splitting the baseline into two different ranges. This is done in Figure 14-14. The correlation of 0.71 in cell G3 of Figure 14-14 is the same as in cell D3 of Figure 14-13. This isn't surprising, because the values in the ranges in each figure are identical. The only unfamiliar aspect in Figure 14-13 is that the two ranges overlap.

Sales		Sales					
Date	Revenue	Date	Revenue				
Jan-15	\$ 76,000	Feb-15	\$ 59,000			Correlation	
Feb-15	\$ 59,000	Mar-15	\$ 29,310			0.71	
Mar-15	\$ 29,310	Apr-15	\$ 27,025				
Apr-15	\$ 27,025	May-15	\$ 34,083				
May-15	\$ 34,083	Jun-15	\$ 65,034				
Jun-15	\$ 65,034	Jul-15	\$ 38,012				
Jul-15	\$ 38,012	Aug-15	\$ 116,000				
Aug-15	\$ 116,000	Sep-15	\$ 109,000				
Sep-15	\$ 109,000	Oct-15	\$ 123,000				
Oct-15	\$ 123,000	Nov-15	\$ 107,004				
Nov-15	\$ 107,004	Dec-15	\$ 84,754				
Dec-15	\$ 84,754	Jan-16	\$ 29,322				
Jan-16	\$ 29,322	Feb-16	\$ 23,000				
Feb-16	\$ 23,000	Mar-16	\$ 41,522				
Mar-16	\$ 41,522	Apr-16	\$ 37,754				
Apr-16	\$ 37,754	May-16	\$ 52,105				
May-16	\$ 52,105	Jun-16	\$ 31,563				
Jun-16	\$ 31,563	Jul-16	\$ 96,266				
Jul-16	\$ 96,266	Aug-16	\$ 145,900				
Aug-16	\$ 145,900	Sep-16	\$ 164,323				
Sep-16	\$ 164,323	Oct-16	\$ 137,751				
Oct-16	\$ 137,751	Nov-16	\$ 92,940				
Nov-16	\$ 92,940	Dec-16	\$ 110,000				

**FIGURE 14-14:**  
The values in column E start with the value in B3.

## Diagnosing autocorrelation

Earlier in this chapter, I imply that if the autocorrelation in a baseline is 0.65, you should take first differences to remove the autocorrelation. A little later, I said that the autocorrelation of the first differences was negligible if it was as small as  $-0.10$ . So, that begs the question, "How small is negligible?"

Chapters 4 and 17 describe tests that help you decide whether a correlation is real or phantom. Here's another test that you'll probably find quicker to carry out than the others. Take these steps:

1. **Copy the baseline you're going to forecast from (whether it's the actuals or the first differences you've taken from the actuals) and, for convenience, paste it into a new worksheet.**

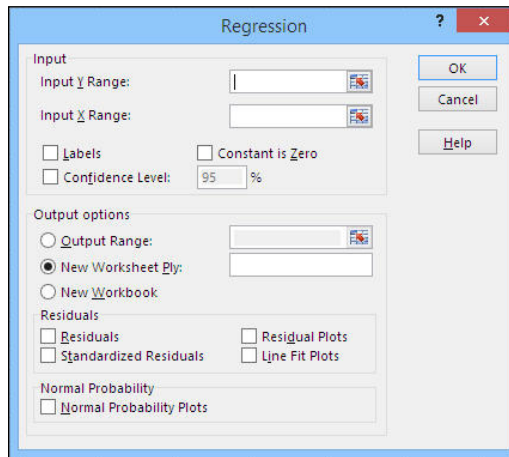
Let the paste start in, say, cell A1. I'm using first differences here to demonstrate how the Regression tool's results show how you can test statistically for its presence or absence.

2. **Select cell B1 and paste again.**

You now have identical values in columns A and B of the new worksheet. Suppose that these values are in A1:A48 and B1:B48.

3. **Choose Tools ⇨ Data Analysis, and select Regression in the Analysis Tools list box.**

The Regression dialog box, shown in Figure 14-15, appears.



**FIGURE 14-15:**  
The Input Y Range box can take the focus when you select the Output Range button, which can cause you to lose the address in the Input Y Range box.

4. **In the Regression dialog box, click in the Input Y Range box and drag through B2:B48.**

5. **Click in the Input X Range box and drag through A1:A47.**

The Regression tool will not allow these two ranges to overlap on the worksheet, which is why this list of steps asks you to create two separate ranges in columns A and B.

6. **Select the Output Range option button.**

7. **Make sure that the Output Range address box has the focus rather than the Input Y Range box.**

8. Click in a worksheet cell that has blank cells below it and to its right.
9. Click OK.

The results look something like those in Figure 14-16.

	A	B	C	D	E	F	G	H	I
	Sales	Sales							
1	Revenue	Revenue2		SUMMARY OUTPUT					
2	\$ (17,000)	\$ (17,000)							
3	\$ (29,690)	\$ (29,690)		Regression Statistics					
4	\$ 21,790	\$ 21,790		Multiple R	0.102				
5	\$ (17,017)	\$ (17,017)		R Square	0.010				
6	\$ 30,951	\$ 30,951		Adjusted R Square	-0.039				
7	\$ (27,022)	\$ (27,022)		Standard Error	36553.520				
8	\$ 77,988	\$ 77,988		Observations	22				
9	\$ (7,000)	\$ (7,000)							
10	\$ 14,000	\$ 14,000		ANOVA					
11	\$ (57,800)	\$ (57,800)			<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
12	\$ 4,000	\$ 4,000		Regression	1	281129179.62	281129179.62	0.210	0.651
13	\$ (39,878)	\$ (39,878)		Residual	20	26723196178.90	1336159808.94		
14	\$ (6,322)	\$ (6,322)		Total	21	27004325358.52			
15	\$ 18,522	\$ 18,522							
16	\$ (3,769)	\$ (3,769)							
17	\$ 14,351	\$ 14,351							
18	\$ (20,542)	\$ (20,542)							
19	\$ 29,737	\$ 29,737							
20	\$ 84,600	\$ 84,600							
21	\$ 18,423	\$ 18,423							
22	\$ (26,572)	\$ (26,572)							
23	\$ (44,811)	\$ (44,811)							
24	\$ 17,060	\$ 17,060							

**FIGURE 14-16:**  
Look to the Significance of F to determine whether you should treat the autocorrelation as real.

The cell to focus on is the one labeled *Significance of F*. As you've set up the regression analysis, that cell tells you the statistical significance of the correlation between cells A1:A47 and B2:B48 — which is, in fact, the autocorrelation between A1:A47 and A2:A48.

The higher the value labeled *Significance of F*, the less likely it is that the autocorrelation is real. (Sorry about that: Statements of statistical significance are usually backwards.) In Figure 14-16, the significance level is a touch over 0.65. This means that you would have gotten an autocorrelation as large as  $-0.10$  about 65 percent of the time, even if there were no real autocorrelation in the data at all. So you can assume that the autocorrelation in the baseline actuals has been removed by the first differencing.

Suppose that the autocorrelation was 0.65 and the resulting significance level was, say, 0.05. This would mean that only 5 percent of the time would you see an autocorrelation as large as 0.65, if the real autocorrelation were zero. So in that case you should assume that the autocorrelation is real.

What's a "real" autocorrelation? The values that you're testing are usually samples from a much longer baseline. Perhaps the baseline *could* go back to, say, January 1920, and you don't have the time, patience, or resources to retrieve all those values. If you did, you could calculate the real autocorrelation, the one based on the full population of values, rather than just a sample from that population. What you're doing with the significance test is checking the likelihood that you'd get an autocorrelation as large as  $-0.10$  in your sample if the autocorrelation in the population were zero. If that likelihood is, say, 0.65 (that is, 65 percent), it's hard to argue that the population's autocorrelation is non-zero.

In sum, the presence of autocorrelation can both hinder and assist forecasts. I give it quite a bit of coverage in this chapter because you can see autocorrelation's effects more clearly in forecasts made by exponential smoothing and regression than in moving average forecasts.

Regardless of your choice of forecasting method, differencing the baseline is often the best way to manage autocorrelation. You see examples in this chapter of how a baseline's first differences show virtually no autocorrelation. You see in Chapter 17 how returning forecast differences to the baseline helps you deal with one of the causes of autocorrelation: the presence of trend in a time series.



## Chapter 15

# Smoothing: How You Profit from Your Mistakes

Smoothing, in this case *exponential* smoothing, is a kind of modified moving average. Don't let the name put you off. You won't have to deal with any exponents (or, for that matter, proponents, deponents, opponents, or components). It's a simple idea, tricked out — as so many simple ideas are — with a fancy name. The idea is to correct a prior forecast and use that correction in making the next forecast. That's pretty straightforward, no?

The Data Analysis add-in does exponential smoothing on your behalf. The Exponential Smoothing tool is one of the tools in the Data Analysis add-in that returns a formula, so if your input data changes, the forecasts will update automatically. But you want a bit more control over the formulas than the Exponential Smoothing tool gives you, and this chapter shows you how to get that control.

Exponential smoothing uses something called — here's another bit of jargon — a *smoothing constant*. It helps determine the amount of the error in an earlier forecast to use in making the next one.

After you've chosen a smoothing constant, even just for the moment, you've automatically chosen a *damping factor* — yet another piece of jargon. I wouldn't even bring it up here except that the Data Analysis add-in's Exponential Smoothing tool uses that term.

The point to take away from this prolog is that, when you get past the snooty terminology, exponential smoothing is no more difficult than the moving averages concepts that it's based on.

## Correcting Errors: The Idea Behind Smoothing

You don't *have* to get a handle on the idea behind exponential smoothing. But it doesn't take much time or effort, and it can help you understand what's going on when you make your forecast. Plus, if you're ever asked to explain a forecast, you don't want to have to say, "We used Excel's Data Analysis add-in. It took care of making the actual forecasts. It's really convenient." That doesn't help your credibility.

Much better is something like, "We used exponential smoothing. It's a standard forecasting technique. We could have used moving averages, but in this case, we would have lost too much of the baseline. We could have used regression, but exponential smoothing gave us more accurate forecasts. Using the error in the last forecast to fine-tune the current forecast — and that's the basic idea behind exponential smoothing — improved the accuracy of our forecasts. Now, if we can return to the chart . . ."

### Adjusting the forecast

Forecasts are usually wrong. Most often, if they're created rationally, the forecasts are not off by too much, but they *are* off.

You'd love to go back and adjust your forecast for, say, January after the actual results came in. You'd take your forecast of 26,000 units sold, compare it to the actual of 25,437 units sold, and adjust your forecast down by 563 units. You'd look like a wizard.

Sadly, we're not wizards. But although you can't pretend that you made that adjustment before you saw January's actuals, you can still use it. You can use it on the next forecast.



Suppose that a baseline of sales revenue has been gradually trending up. Some months the sales are up a good bit, some months they're down, but when you chart the baseline, you can tell that the general trend is up. On the basis of that general trend, you make a forecast for June. Your May actuals were \$510,545, and you forecast \$519,827 for June. When the next set of actuals comes in, you find that the revenue for June was \$516,188.

Of course, you'd like to take some of your rosy forecast back, but you can't. The C-suite has already seen it, they've seen the actuals, and they've noticed the difference. Fortunately, you've prepared them emotionally for this sort of thing and you know they won't jump your case — not as long as they know that you've noticed it too and will take it into account in your next forecast.

And you will, if you're using exponential smoothing. In effect, here's what you do:

**1. Decide how much weight you want to give to the error in the June forecast.**

That will be somewhere between 0 percent and 100 percent. Suppose you choose 30 percent.

**2. Multiply that weight times the error.**

The error for June was \$516,188 – \$519,827, or \$(3,639). Thirty percent of \$(3,639) is \$(1,091.70), which is the weighted error.

**3. Add the weighted error to the prior forecast.**

At this point the error is a negative number. This is because your forecast was too high and you subtract the forecast from the actual to get the error. The effect of adding June's negative weighted error to June's forecast is to pull the forecast for July down.

The result would be:

$$\text{July Forecast} = \text{June Forecast} + 0.3 (\text{June Actual} - \text{June Forecast})$$

$$\text{July Forecast} = \$519,827 + [0.3 \times \$(3,639)]$$

$$\text{July Forecast} = \$518,735.30$$

You wish you could have done that for June, because it would have made your June forecast closer to the June actual. But it's not too late to build it into the July forecast.

Keep in mind a few things about the exponential smoothing approach:

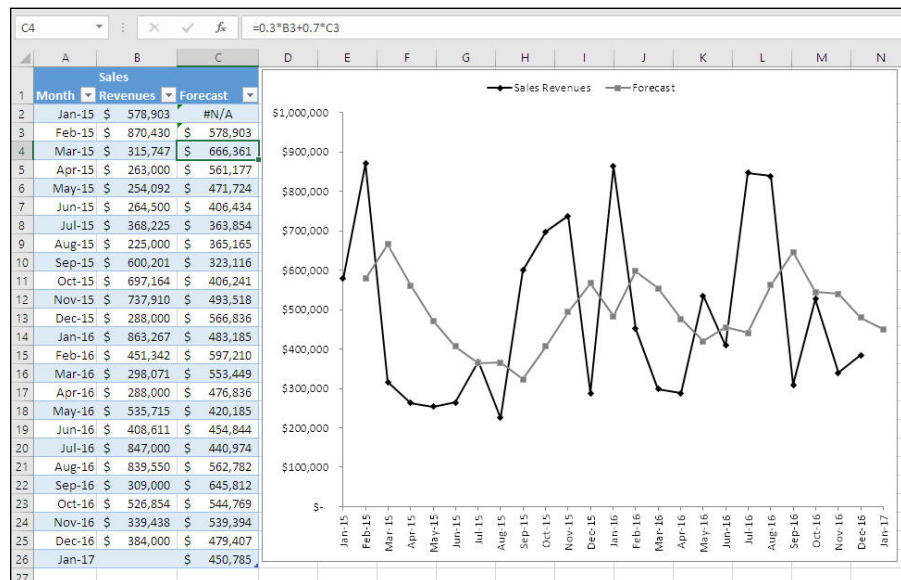
- » **It helps your forecasts start reacting quickly when a baseline is starting to turn up or to turn down.** In fact, it starts as quickly as the very next period. If things start ticking down in June, your July forecast will reflect that, same as it would reflect an uptick.
- » **Your choice in the matter, after you've decided to use exponential smoothing, is limited to *how much* you'll react to the previous error.** In this example, you made that choice when you decided to give 30 percent weight to the error in the prior forecast.
- » **That 30 percent (equivalently, 0.3) weight is the smoothing constant.** The smoothing constant is required to be between 0.0 and 1.0, inclusive.

Figure 15-1 shows an example of forecasting with exponential smoothing.

There's a more convenient form of the equation than we've looked at so far, and you'll see it more frequently than any other form. In cell C4 of Figure 15-1, it's

$$=0.3*B3+0.7*C3$$

That is, the smoothing constant times the prior actual, plus the damping factor times the prior forecast. This form is convenient to copy and paste down the worksheet, but it doesn't tell you what's really going on — not, at least, in an intuitive sense. For that, see the following section.



**FIGURE 15-1:**  
The forecasts in column C were created by Excel's Exponential Smoothing tool.

# Why they call it “exponential smoothing”

You can safely skip this section and still use exponential smoothing as an approach to forecasting. But if you’re interested, read on.

Make these assumptions:

- » You’re working with a baseline built with monthly data.
- » The baseline starts in April.
- » The smoothing constant you’re using is 0.3.

Then the forecast for July would be:

$$\text{July forecast} = 0.3 \times \text{June actual} + 0.7 \times \text{June forecast}$$

What’s the June forecast? It’s

$$\text{June forecast} = 0.3 \times \text{May actual} + 0.7 \times \text{May forecast}$$

Here, the smoothing constant is 0.3 and the damping factor is, therefore, 0.7.



REMEMBER

The smoothing constant and the damping factor always add up to 1.0. If you know one of them, you know the other.

## Expanding the equation

In the July forecast, if you remove the June forecast figure and replace it with the *formula* for the June forecast, you get:

$$\text{July forecast} = 0.3 \times \text{June actual} + 0.7 \times (0.3 \times \text{May actual} + 0.7 \times \text{May forecast})$$

The May forecast is:

$$0.3 \times \text{April actual} + 0.7 \times \text{April forecast}$$

Now, if the baseline begins in April, you can’t go back to March and forecast a value for April: You have neither an actual nor a forecast result for March, and both are needed to make a forecast for April. Further, without a forecast for April, you can’t make a forecast for May. So, to stop this series of toppling dominoes, you take the April actual to be the May forecast. This is standard operating procedure for forecasting. When you get back near the start of a baseline looking for a forecast of the baseline’s second value, you use the actual value of the first period instead. (There’s no earlier value available to forecast from.)



TIP

There are other ways to get the first forecast. One is called *backcasting*. The idea is to turn the baseline around and forecast into the past rather than into the future. (Sort of like Spock and Kirk jumping through that portal into the 20th century where they meet Joan Collins. Right.) There are a few other ways to get around that first forecast problem. I don't get into them in this book, but you should be aware that they exist.

So, the July forecast is:

$$=0.3 \times \text{June Actual} + 0.7 \times \text{June forecast}$$

Substituting the formula for the June forecast:

$$\text{July forecast} = 0.3 \times \text{June actual} + 0.7 \times (0.3 \times \text{May actual} + 0.7 \times \text{May forecast})$$

But the forecast for May is taken to be the actual observation for April, because April is the first period in the baseline. So the forecast for July becomes

$$\text{July forecast} = 0.3 \times \text{June actual} + 0.7 \times (0.3 \times \text{May actual} + 0.7 \times \text{April actual})$$

Compare those equations with those in the range C3:C5 of Figure 15-1.

## Understanding the exponents

Suppose that you're at the end of the third month of tracking a product's sales, and you're ready to forecast where the sales will head on the fourth month. (Of course, normally you would wait for a much longer baseline, but it clarifies the discussion to work with a smaller sample.) You use this formula:

$$\hat{y}_4 = \alpha y_3 + (1 - \alpha) \hat{y}_3$$

That is, the forecast for Month 4 ( $\hat{y}_4$ ) equals:

Alpha ( $\alpha$ , also termed the smoothing constant)

Times the actual sales during Month 3 ( $y_3$ )

Plus  $(1 - \alpha)$

Times the forecast for Month 3 made at the end of Month 2 ( $\hat{y}_3$ )

Now, in the forecast for Month 4, replace the forecast for Month 3,  $\hat{y}_3$ , with its own calculation:

$$\hat{y}_3 = \alpha y_2 + (1 - \alpha) \hat{y}_2$$

So:

$$\hat{y}_4 = \alpha y_3 + (1 - \alpha)(\alpha y_2 + (1 - \alpha)\hat{y}_2)$$

Now, the forecast for Month 2 is the actual value observed for Month 1:

$$\hat{y}_2 = y_1$$

And therefore:

$$\hat{y}_4 = \alpha y_3 + (1 - \alpha)(\alpha y_2 + (1 - \alpha)y_1)$$

Expand the second term by multiplying through by  $(1 - \alpha)$ :

$$\hat{y}_4 = \alpha y_3 + (1 - \alpha)(\alpha y_2) + (1 - \alpha)(1 - \alpha)y_1$$

$$\hat{y}_4 = \alpha y_3 + (1 - \alpha)(\alpha y_2) + (1 - \alpha)^2 y_1$$

Now, if you raise a number to the power of 0, the result is 1. And if you raise any number to the power of 1, you get the number itself. So you can rewrite the prior equation in this way:

$$\hat{y}_4 = (1 - \alpha)^0 \alpha y_3 + (1 - \alpha)^1 \alpha y_2 + (1 - \alpha)^2 y_1$$

Finally, just rearrange the order of the terms in the prior equation to show the older observations on the left and the newer observations on the right. This makes it easier to see what's going on:

$$\hat{y}_4 = (1 - \alpha)^2 y_1 + (1 - \alpha)^1 \alpha y_2 + (1 - \alpha)^0 \alpha y_3$$

Suppose you set  $\alpha$  to the value of 0.3. Then, using numbers instead of  $(1 - \alpha)$  raised to some power:

$$\hat{y}_4 = .49y_1 + .7\alpha y_2 + 1\alpha y_3$$

I think I can actually see your eyes glazing over. But bear with me just a little further. You can see that each of the first three months' observations —  $y_1$ ,  $y_2$  and  $y_3$  — appears in the forecast for Month 4. The oldest observation,  $y_1$ , is multiplied by  $.7^2$  or  $.49$ . The next oldest observation,  $y_2$ , is multiplied by  $.7^1$ , or  $.7$ . And the most recent observation,  $y_3$ , is multiplied by  $.7^0$ , or 1. The older the observation, the smaller its multiple, and therefore the smaller the observation's contribution to the current forecast.

## Making sense of the equations

And that's why it's called exponential smoothing. The farther back you go toward the start of the baseline, the larger the exponent for  $(1 - \alpha)$ , or, if you prefer, one minus the smoothing constant. And because you're raising a fraction (in this

example, that's 0.7) to higher and higher powers, the contribution made by older actuals gets smaller and smaller ( $0.7^2 = 0.49$ ,  $0.7^3 = 0.34$ ,  $0.7^4 = 0.24$ , and so on).

Intuitively, this is how things should be. The farther away a baseline value is from the present day, the smaller you expect its lingering influence to be.

## Fooling around with the smoothing constant

Perhaps the idea of fooling around with a smoothing constant has never occurred to you. But it's a useful exercise anyway.

Chapter 13 talks some about the effect of longer and shorter lengths in a baseline that contribute to a moving average. Other things being equal, these statements are true:

- » The more baseline periods that go into a moving average, the more slowly the moving average reacts to changes in the baseline, and the smoother the series of forecasts. Also, there are fewer forecasts you can make because you lose more periods at the start of the baseline.
- » The fewer baseline periods that go into a moving average, the more quickly the moving average reacts to — or *tracks* — changes in the baseline, and the more closely the forecasts come to the baseline values themselves. Also, you lose fewer periods at the start of the baseline.

You can see the same effect in exponential smoothing, but because of the smoothing constant and the damping factor, the effect is somewhat different from moving averages. Apropos, here are another couple of truths about exponential smoothing:

- » The higher the smoothing constant, the more quickly the forecast tracks the baseline.
- » The lower the smoothing constant, the slower the tracking — and therefore the smoother the forecast series tends to appear.

Notice what happens to the equation at the smoothing constant's extremes of 1.0 and 0.0:

$$\text{July forecast} = \text{Smoothing Constant} \times \text{June actual} + \text{Damping Factor} \times \text{June forecast}$$

$$\text{July forecast} = 1.0 \times \text{June actual} + 0.0 \times \text{June forecast}$$

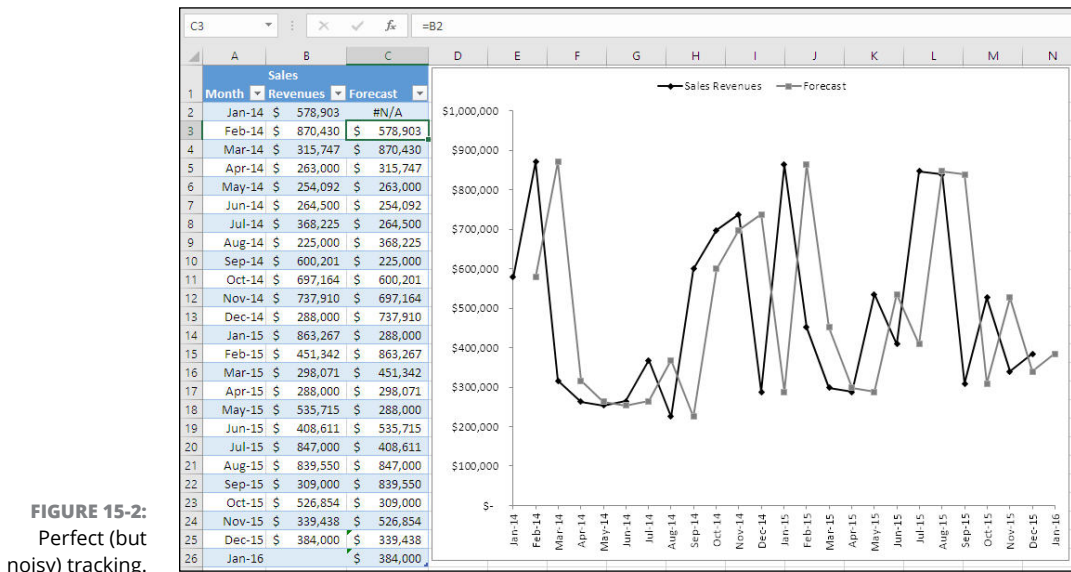
So with a smoothing constant of 1.0, the forecasts track the actuals precisely, just one period late. The closer the smoothing constant is to 1.0, the closer the tracking.

On the other hand:

$$\text{July forecast} = 0.0 \times \text{June actual} + 1.0 \times \text{June forecast}$$

In this case, the only actual that comes into play is the first one in the baseline. Because the baseline's first value is also taken to be the second period's forecast, all forecasts are equal to the first value in the baseline.

Figures 15-2 and 15-3 show how this works. As a practical matter, the forecasts in Figure 15-2 are useless because they're nothing more than the actuals, delayed a month. Figure 15-3's forecasts are useless because they're nothing more than January 2014's actual, projected forward 23 months.

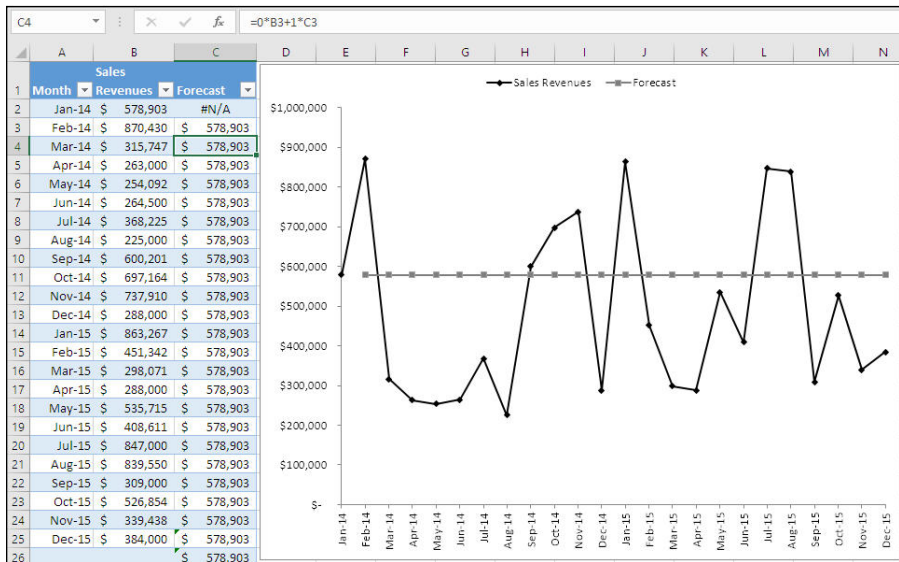


**FIGURE 15-2:**  
Perfect (but noisy) tracking.

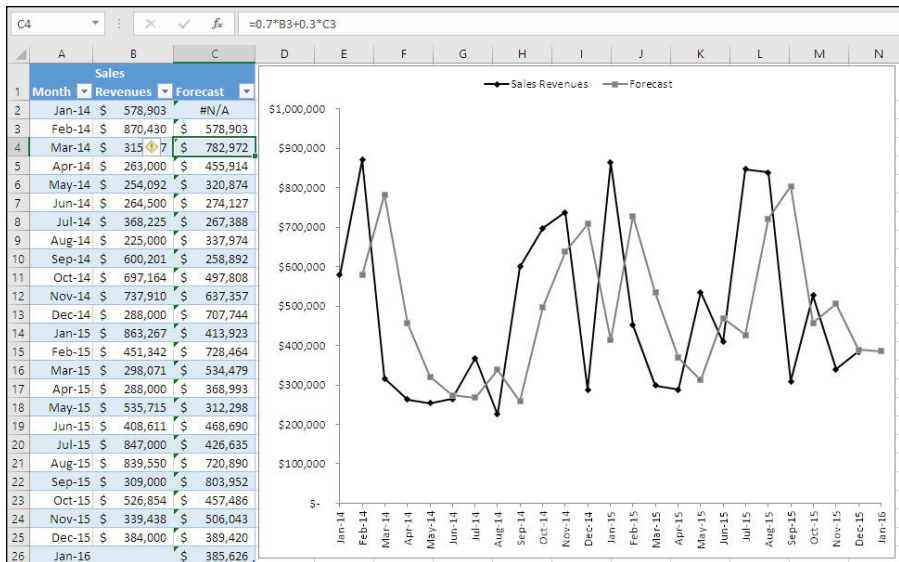
So, in practice, you should never have use for either 0.0 or 1.0 as a smoothing constant. But seeing what happens when you use those values helps you remember what happens when you use a smaller or a larger smoothing constant.

For a more realistic view, take a look at Figure 15-4. It's another version of Figure 15-1, but there the smoothing constant was 0.3 (and, therefore, the damping factor was 0.7). Figure 15-4 reverses them: The smoothing constant is 0.7 and the damping factor is 0.3.

**FIGURE 15-3:**  
Perfect (but information-free) smoothing.



**FIGURE 15-4:**  
Notice that the forecasts are more volatile than in Figure 15-1.



In Figure 15-4, where the smoothing constant is 0.7, the forecasts track the actuals much more closely than they do in Figure 15-1, where the smoothing constant is 0.3. The issue once again is tracking versus smoothing. The higher the smoothing constant, the closer the forecasts track the actuals, the more that noise enters the picture and the more that noise obscures the signal. The more smoothing



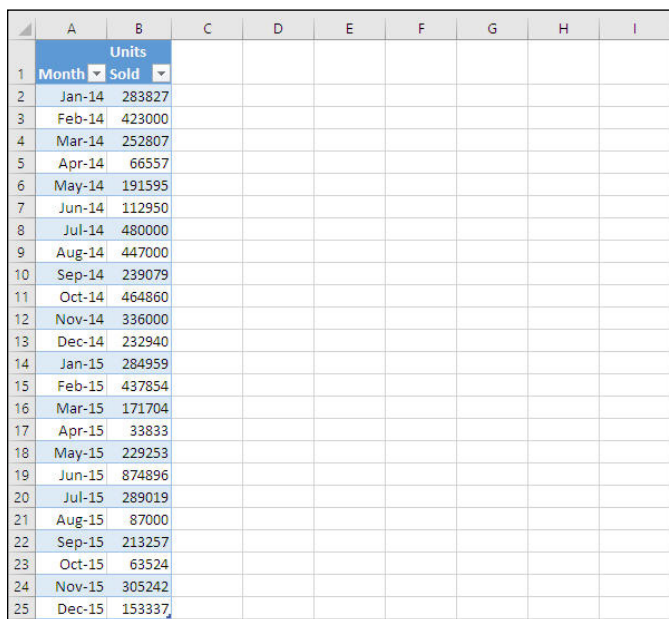
there is in the forecasts, the more slowly the forecasts react to fundamental, lingering changes in the baseline — therefore, the longer it may take you to recognize that something important has occurred.

## Using the Smoothing Tool's Formula

Chapter 10 gives a brief overview of using the Data Analysis add-in to make a forecast by means of its Exponential Smoothing tool. This section recaps that process and shows you how to tinker with the forecasts after the basic spadework has been done.

### Getting a forecast from the Exponential Smoothing tool

Figure 15-5 shows a baseline you might start with.



	A	B	C	D	E	F	G	H	I
1	Month	Units Sold							
2	Jan-14	283827							
3	Feb-14	423000							
4	Mar-14	252807							
5	Apr-14	66557							
6	May-14	191595							
7	Jun-14	112950							
8	Jul-14	480000							
9	Aug-14	447000							
10	Sep-14	239079							
11	Oct-14	464860							
12	Nov-14	336000							
13	Dec-14	232940							
14	Jan-15	284959							
15	Feb-15	437854							
16	Mar-15	171704							
17	Apr-15	33833							
18	May-15	229253							
19	Jun-15	874896							
20	Jul-15	289019							
21	Aug-15	87000							
22	Sep-15	213257							
23	Oct-15	63524							
24	Nov-15	305242							
25	Dec-15	153337							

**FIGURE 15-5:** You don't need the date information, but having it there lets you verify the sales figures.

Here are the steps to add to this baseline a series of forecasts using the Data Analysis add-in's Exponential Smoothing tool:

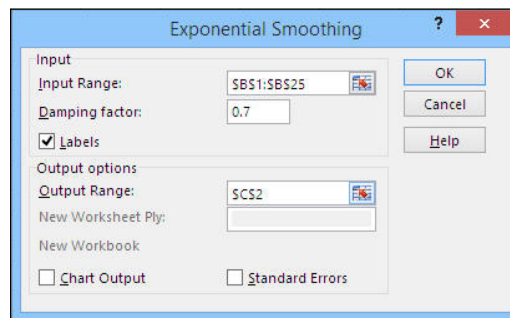
- 1. Make sure the Data Analysis add-in is installed in Excel.**  
See Chapter 7 for installation information.
- 2. If necessary, activate the worksheet that contains the baseline.**
- 3. Go to the Ribbon's Data tab and click Data Analysis in the Analyze group.**  
The Data Analysis dialog box opens.
- 4. In the Data Analysis dialog box, click on Exponential Smoothing in the list box, and click OK.**
- 5. Click in the Input Range box, and drag through B1:B25 on the worksheet.**
- 6. Click in the Damping Factor box and enter 0.7.**  
This corresponds to a smoothing constant of 0.3.
- 7. Because your input range included cell B1, which names the baseline in B2:B25, select the Labels check box.**
- 8. Click in the Output Range box, and then click in the cell where you want the output to begin.**

On the worksheet shown in Figure 15-5, that would be C2 — it would not be C1. The Exponential Smoothing tool does not provide a label for its forecasts, so clicking C1 would put the first output value in C1, one row higher than the first value in the baseline. That first output value is always #N/A because the Exponential Smoothing tool has no way to forecast a value for the first period. Starting the output at C2 puts the first observed value in the correct location, cell C3, as the forecast for the second period.

Figure 15-6 shows how the dialog box appears just before you click OK.

- 9. Click OK.**

**FIGURE 15-6:** This example continues by putting the forecasts into a new chart, so don't bother to select the Chart Output check box.



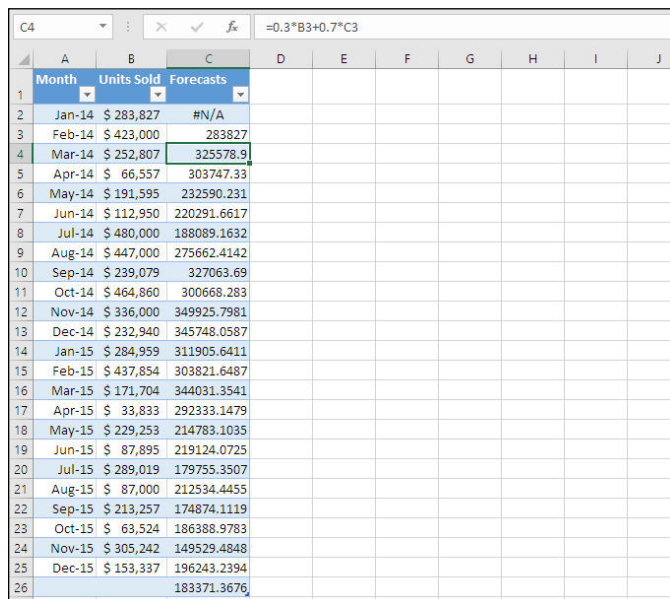
## 10. Type the label Forecasts in cell C1.

You'll find the results of using the Exponential Smoothing tool in column C of Figure 15-7.

One problem with the Exponential Smoothing tool is that it doesn't automatically give you the forecast for the period following the baseline's last period — and that's what you're most interested in. So, to get that value, take these steps:

### 1. Using the layout in Figure 15-7, select cell C25.

More generally, select the bottommost cell that the Exponential Smoothing tool returned.



Month	Units Sold	Forecasts
Jan-14	\$ 283,827	#N/A
Feb-14	\$ 423,000	283827
Mar-14	\$ 252,807	325578.9
Apr-14	\$ 66,557	303747.33
May-14	\$ 191,595	232590.231
Jun-14	\$ 112,950	220291.6617
Jul-14	\$ 480,000	188089.1632
Aug-14	\$ 447,000	275662.4142
Sep-14	\$ 239,079	327063.69
Oct-14	\$ 464,860	300668.283
Nov-14	\$ 336,000	349925.7981
Dec-14	\$ 232,940	345748.0587
Jan-15	\$ 284,959	311905.6411
Feb-15	\$ 437,854	303821.6487
Mar-15	\$ 171,704	344031.3541
Apr-15	\$ 33,833	292333.1479
May-15	\$ 229,253	214783.1035
Jun-15	\$ 87,895	219124.0725
Jul-15	\$ 289,019	179755.3507
Aug-15	\$ 87,000	212534.4455
Sep-15	\$ 213,257	174874.1119
Oct-15	\$ 63,524	186388.9783
Nov-15	\$ 305,242	149529.4848
Dec-15	\$ 153,337	196243.2394
		183371.3676

**FIGURE 15-7:** The Exponential Smoothing tool does not apply formats (here, currency) in the input range to its output range.

### 2. Move your mouse pointer over the bottom-right corner of that cell until the pointer changes to a crosshair.

That corner is called the *fill handle*.

### 3. Press and hold down the left mouse button, drag down one row, and release the mouse button.

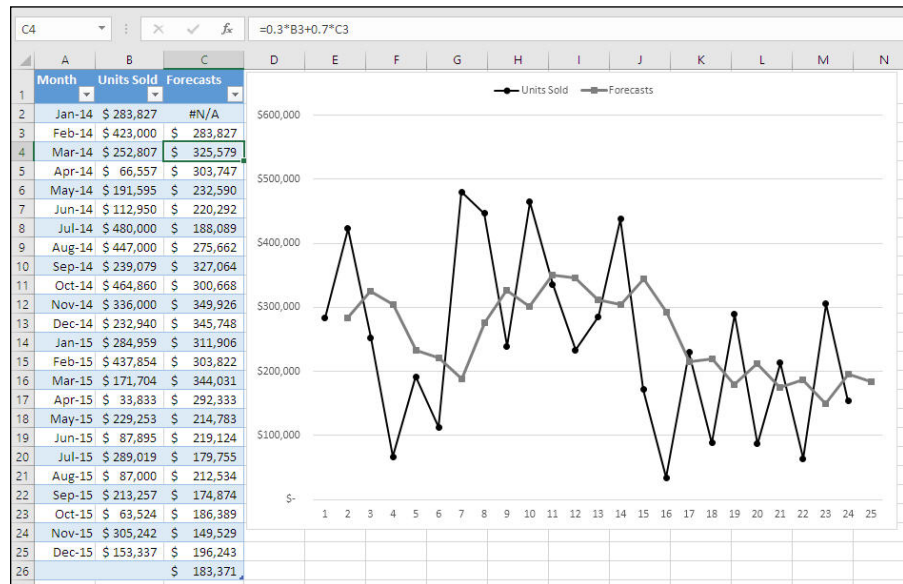
Steps 1 through 3 get you a forecast for the next time period. All that's left is to put the forecasts into the chart.

**4. Using the layout in Figure 15-7, select the range B1:C26.**

More generally, select the cell that contains the label for your baseline field; in Figure 15-7, that's B1.

**5. Go to the Ribbon's Insert tab and click the icon for Line charts in the Charts group. Select a line chart subtype.**

Change the chart formatting however you want. In Figure 15-8, I made the chart larger by dragging its corner handles, set the font size to 11 for all text in the chart, and called for a legend at the top.



**FIGURE 15-8:** A cell that contains either a blank value or an #N/A value will not appear in the chart.



**REMEMBER**

When you're using the Data Analysis add-in's Exponential Smoothing tool, you need to think in terms of a damping factor, because the tool insists on it. Specifying a damping factor and getting the smoothing constant by subtracting the damping factor from 1.0 isn't unreasonable, but it is unusual.



**TIP**

Don't try putting the damping factor in a worksheet cell and then referencing the cell in the Exponential Smoothing dialog box's Damping Factor box. That box won't accept a cell address. If you want to take that tack — and I recommend you do so (see the next section) — you have to do it after the tool has created the forecast.

## Modifying the smoothing constant

After the Exponential Smoothing tool has given you a series of forecasts, you're in a position to change the forecasts by tinkering with the value of the smoothing constant. The formulas for the forecasts follow the form mentioned at the end of the "Adjusting the forecast" section, early in this chapter. The formulas for the forecasts in C4:C6 of Figure 15-8 are

```
=0.3*B3+0.7*C3  
=0.3*B4+0.7*C4  
=0.3*B5+0.7*C5
```

This is good news and bad news. The good news is that you have formulas to work with, not static values like these:

```
325578  
303747  
232590
```

which are the results of the formulas in C4:C6.

The bad news is that the formulas contain constants, 0.3 and 0.7, which you can't easily change. You can turn that bad news into good news by taking these steps, using the worksheet in Figure 15-8 as a basis:

- 1. In cell B28, enter =1 - A28.**  
It doesn't matter where you enter it, as long as it's in some blank cell so that you're not overwriting anything important.
- 2. In cell A28, enter 0.3.**  
Again, it doesn't matter where you enter it.
- 3. If necessary, format A28:B28 to the Number format, with two decimals: Select A28:B28, go to the Ribbon's Home tab, and choose Number from the drop-down at the top of the Number group.**
- 4. Select cell C4.**

In the Formula Bar, you'll see its formula:

```
=0.3*B3+0.7*C3
```

- 5. Click in the Formula Bar, change the 0.3 in the formula to \$A\$28, change the 0.7 in the formula to \$B\$28, and press Enter or click the Enter button.**

The formula should now be:

```
=$A$28*B3+$B$28*C3
```

- 6. If necessary, select cell C4 again. Then move your mouse pointer over its lower-right corner until the pointer changes to a crosshair. Press and hold the left mouse button, and drag down through C26, and release the mouse button. Instead of dragging down, you could simply double-click the fill handle.**



WARNING

Be sure to include the dollar signs in Step 5. If you don't, the references will be relative. Then, when you do the autofill in Step 6, the subsequent references to the smoothing constant and damping factor point to different and empty cells and your forecasts go to zero. An alternative, if you're an Excel names maven, is to name cells A28 and B28, using absolute references.

The effects of Steps 1 through 6 are to

- » Put the values for the smoothing constant and the damping factor explicitly on the worksheet.
- » Make the forecast formulas in column C refer to the smoothing constant and damping factor in A28:B28.

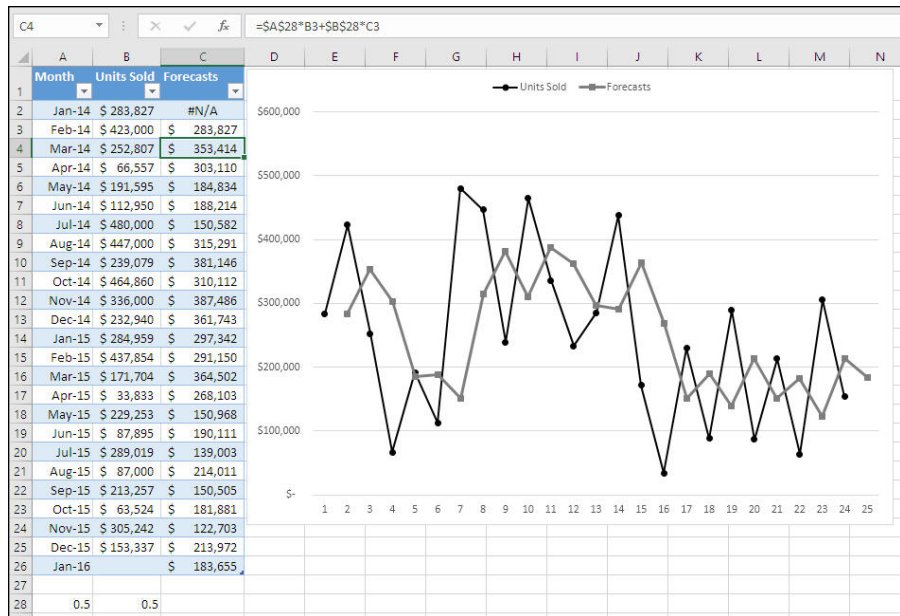
Now you can change the smoothing constant in A28. When you do, the damping factor in B28 changes accordingly. And when they've changed, so do the forecasts in column C that depend on those two values. Furthermore, the chart updates. That's a lot more efficient than the alternative that the Exponential Smoothing tool leaves you with, which is to do a Find and Replace in the formulas every time you want to try the effect of a different set of constants.

Figure 15-9 shows the result of changing the smoothing constant in A28 from 0.3 to 0.5.

Comparing Figures 15-8 and 15-9, you can see that Figure 15-8's forecasts chart a little smoother than do Figure 15-9's. The other side of that coin is that Figure 15-9's forecasts respond a little faster to changes in the baseline than do Figure 15-8's. You can experiment with different values of the smoothing constant to see what effect they have on the charted forecasts.

Experimenting with different smoothing constants is a good way to get familiar with the relationship between the constants, the forecast, and the appearance of the forecasts when they're charted with the baseline actuals. But when it comes time to settle on a smoothing constant for your forecasts, there's nothing like running the numbers.

**FIGURE 15-9:**  
A smoothing constant of 0.5 tracks the baseline more closely than a smoothing constant of 0.3.



## Finding the Smoothing Constant

Perhaps the best way to decide on a value for the smoothing constant is to use a yardstick of some sort to compare the results you get using different constants. Probably the standard method is to minimize the forecast errors you get with different constants.



REMEMBER

The forecast error, or just *error*, is the difference between the forecast for a given period and the actual for the same period.

### Developing the yardstick

Figure 15-10 shows a baseline, a smoothing constant, and a damping factor, the forecasts created by the Data Analysis add-in's Exponential Smoothing tool, and the forecast errors obtained by subtracting the forecasts from the actuals.

Here's an overview of what's coming next: You're going to enter a formula that tells you how much error there is in all your forecasts taken together. After you have that formula, you'll use either trial and error (or a very short VBA procedure) to tell you the best value of the smoothing constant to *minimize* that error. That's your yardstick.

1	Month	Units Sold	Forecasts	Forecast Errors				
2	1/1/2016	3029	#N/A			Smoothing Constant	Damping Factor	Total
3	2/1/2016	2825	3029.0	-204.0		0.1	0.9	1
4	3/1/2016	3746	3008.6	737.4				
5	4/1/2016	3945	3082.3	862.7				
6	5/1/2016	3007	3168.6	-161.6				
7	6/1/2016	3309	3152.4	156.6				
8	7/1/2016	3404	3168.1	235.9				
9	8/1/2016	3933	3191.7	741.3				
10	9/1/2016	3883	3265.8	617.2				
11	10/1/2016	4042	3327.5	714.5				
12	11/1/2016	3934	3399.0	535.0				
13	12/1/2016	3231	3452.5	-221.5				
14	1/1/2017	3434	3430.3	3.7				
15	2/1/2017	3564	3430.7	133.3				
16	3/1/2017	3396	3444.0	-48.0				
17	4/1/2017	3471	3439.2	31.8				
18	5/1/2017	3940	3442.4	497.6				
19	6/1/2017	3488	3492.2	-4.2				
20	7/1/2017	3745	3491.8	253.2				
21	8/1/2017	3088	3517.1	-429.1				
22	9/1/2017	2862	3474.2	-612.2				
23	10/1/2017	3609	3413.0	196.0				
24			3432.6					

**FIGURE 15-10:**  
At this point, you have no idea whether a smoothing constant of 0.1 is good, bad, or indifferent.

The formula starts with the Excel worksheet function SUMXMY2, an unfriendly looking function that means this:

SUM: Sum the results

X: A range of values on the worksheet

M: Minus

Y: Another range of values on the worksheet

2: Squared

So, SUMXMY2(B3:B23, C3:C23) means this:

- 1. Subtract the values in C3:C23 from those in B3:B23 (that is, get B3 – C3, B4 – C4, and so on).**

The results are the errors in your forecasts, the differences between the forecast for a period and its actual result. This is the XMY part of the function.

- 2. Square the results.**

This is the XMY2 part of the function.

- 3. Get the sum of the squares.**

This is the SUM part of the function.





TECHNICAL  
STUFF

The idea behind squaring the differences is to get them to all be positive values. If you used just the simple differences, some would be positive and some would be negative, and these tend to cancel one another out. You could use the sum of the absolute value of the differences, and some forecasters do use something called the Mean Absolute Deviation (MAD). But the technique I describe here has broader applicability than the MAD.

Now, divide  $\text{SUMXMY2}(B3:B23, C3:C23)$  by the number of forecasts:

```
SUMXMY2(B3:B23, C3:C23)/COUNT(B3:B23)
```

The COUNT function gives you the number of numeric values in a range. In effect, you're getting the average squared error in your forecasts.

Finally, take the square root of the average squared error:

```
=SQRT(SUMXMY2(B3:B23, C3:C23)/COUNT(B3:B23))
```

In Figure 15-11, the result appears in cell G6. It's the value on the yardstick that tells you how good your forecasts are. The larger the value, the greater the errors, the worse the forecasts.

The result of taking the square root is called the *square root of the mean squared error* (most forecasters call it the *root mean square error*, or RMSE for short).

Month	Units Sold	Forecasts	Forecast Errors				
1/1/2016	3029	#N/A					
2/1/2016	2825	3029.0	-204.0				
3/1/2016	3746	2845.4	900.6				
4/1/2016	3945	3655.9	289.1				
5/1/2016	3007	3916.1	-909.1				
6/1/2016	3309	3097.9	211.1				
7/1/2016	3404	3287.9	116.1				
8/1/2016	3933	3392.4	540.6				
9/1/2016	3883	3878.9	4.1				
10/1/2016	4042	3882.6	159.4				
11/1/2016	3934	4026.1	-92.1				
12/1/2016	3231	3943.2	-712.2				
1/1/2017	3434	3302.2	131.8				
2/1/2017	3564	3420.8	143.2				
3/1/2017	3396	3549.7	-153.7				
4/1/2017	3471	3411.4	59.6				
5/1/2017	3940	3465.0	475.0				
6/1/2017	3488	3892.5	-404.5				
7/1/2017	3745	3528.5	216.5				
8/1/2017	3088	3723.3	-635.3				
9/1/2017	2862	3151.5	-289.5				
10/1/2017	3609	2891.0	718.0				
		3537.2					

**FIGURE 15-11:** Start and end your formula with rows that have both a baseline actual and a forecast — here, those are row 3 and row 23.

Next, the idea is to find the smoothing constant that *minimizes* the root mean square error (that is, the constant that will make each forecast value in your baseline — and, you hope, for the next and as-yet-unknown actual value — as accurate as possible).

## Minimizing the square root of the mean square error

Suppose you put your smoothing constant and damping factor on the worksheet, as in F3 and G3 in Figure 15-10, and point your forecast formulas at those cells. For example, in cell C4 of Figure 15-10:

```
=F$3*B3+G$3*C3
```

Now you can change the smoothing constant; the damping factor recalculates accordingly, and so do your forecasts. With the formula for the root mean square error on the worksheet (cell G6 in Figure 15-11), you can try different values for the smoothing constant and find the one that minimizes the root mean square error value on the yardstick.

Your task isn't too difficult. It's usually enough to try smoothing constants between 0.1 and 0.9, noting the resulting root mean square error each time.

Here's a short Visual Basic for Applications (VBA) procedure that will help you zero in on the best smoothing constant for this baseline, the one that results in the smallest root mean square error.

```
Sub Minimize_Smoothing_Constant()  
Dim i As Single, RowNumber As Integer  
RowNumber = 12  
With ActiveSheet  
    For i = 0.1 To 1 Step 0.1  
        .Cells(3, 6) = i  
        .Cells(RowNumber, 6) = i  
        .Cells(RowNumber, 7) = .Cells(6, 7)  
        RowNumber = RowNumber + 1  
    Next i  
End With  
End Sub
```

Here's a walk through this code. If you're not familiar with VBA, the following will give you an idea of just how easy it is to use. Plus, this will show in a very concrete way how you can find the best smoothing constant, given your baseline data:



TIP

- » The Sub statement just names the procedure. You can refer to its name, `Minimize_Smoothing_Constant`, in several different ways. One of them is shown in the list of numbered steps near the end of this section.
- » The Dim statement names the variables that the procedure uses. It says that `i` can take on decimal values (`single` means *single precision* and defines a decimal variable), and `RowNumber` is an integer.
- » `RowNumber` is set to begin at the number 12. This means that the first row of results is written in row 12.
- » A With block starts. In this case, anything in the block that begins with a dot (such as `.Cells`) is taken to belong to the `ActiveSheet`.
- » A loop starts. The keyword For says, “Here begins a loop.” A loop is just a sequence of statements that repeats some number of times. Here, the statements repeat as `i` goes from 0.1 to 1, in increments, or *steps*, of 0.1. So `i` will equal 0.1, and then 0.2, and then 0.3, and so on until it gets to 1.  
  
You can get even more precision in this sort of analysis by making the increments even smaller, such as 0.05 rather than 0.1.
- » Each time through the loop, the four statements following the For are executed. The Next statement says, “Here ends the loop.” Behind the scenes, VBA checks to see whether it should run the loop again — that is, if `i` has gotten beyond the final value of 1. When it has, the loop has finished and the End Sub statement says the procedure is over.

The statements inside the loop do the following:

- » `.Cells(3, 6) = i`: The current value of `i` is written to the active worksheet, in the third row and the sixth column (that is, cell F3). This is the key statement. It sets the smoothing constant to the current value of `i` (0.1, 0.2, 0.3 . . . 0.9). When it does, all the forecasts in C3:C24 recalculate, as does the value of the root mean square error in G6. Note that the `ActiveSheet` used in the With statement is taken to “own” cell F3.
- » `.Cells(RowNumber, 6) = i`: The value of `i` is written to the sixth column of the active worksheet, in the row identified by `RowNumber`. `RowNumber` starts out with a value of 12, so the first time through the loop, the current value of `i` is written in cell F12. `RowNumber` will be incremented by 1 each time through the loop, so later loops will write the information further down the worksheet.
- » `.Cells(RowNumber, 7) = .Cells(6, 7)`: The current value of the root mean square error is in column G, row 6 — which corresponds to `Cells(6, 7)`. This statement picks up that value and writes it to the row identified by `RowNumber`, in column 7 (or G) — right next to the current value of `i`.
- » `RowNumber = RowNumber + 1`: `RowNumber` is incremented by 1. The next time through the loop, the value of `i` and the current value of the root mean

square error are written the next row down. After that, the With block is ended and so is the subroutine.

To run this code, follow these steps:

**1. Look to see if the Developer tab is visible on the Ribbon.**

If not, click the File tab and choose Options from the navbar. Choose Customize Ribbon from the Excel Options navbar. Fill the Developer check box in the Main Tabs list box and click OK. You should now have a new Developer tab on Excel's Ribbon.

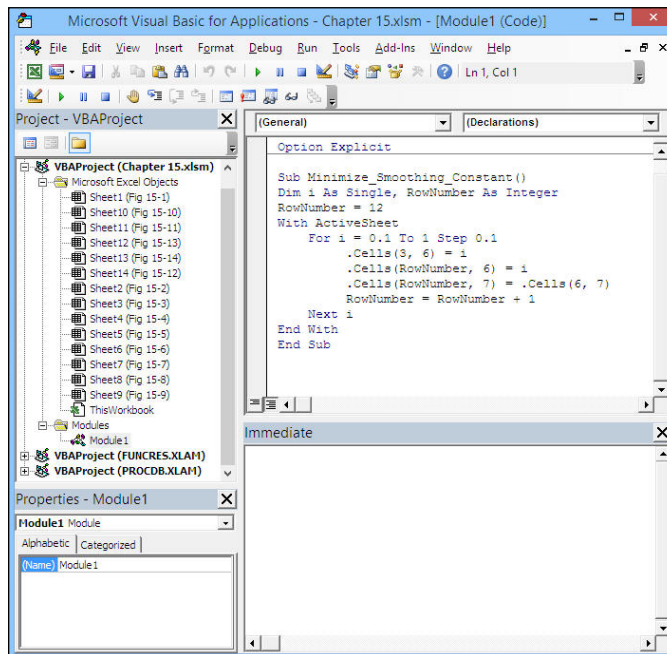
**2. With the workbook for this chapter open and the worksheet named Fig 15-11 active, go to the Developer tab.**

Actually, you can have open any workbook that contains a subroutine named Minimize\_Smoothing\_Constant.

**3. Click the Macros icon in the Code group. The Macro dialog box opens.**

**4. Select Minimize\_Smoothing\_Constant from the Macros list box and click Run.**

The macro shown in Figure 15-12 will execute and put the different values for the smoothing constant in F12:F20, and the corresponding values of the root mean square error in G12:G20.

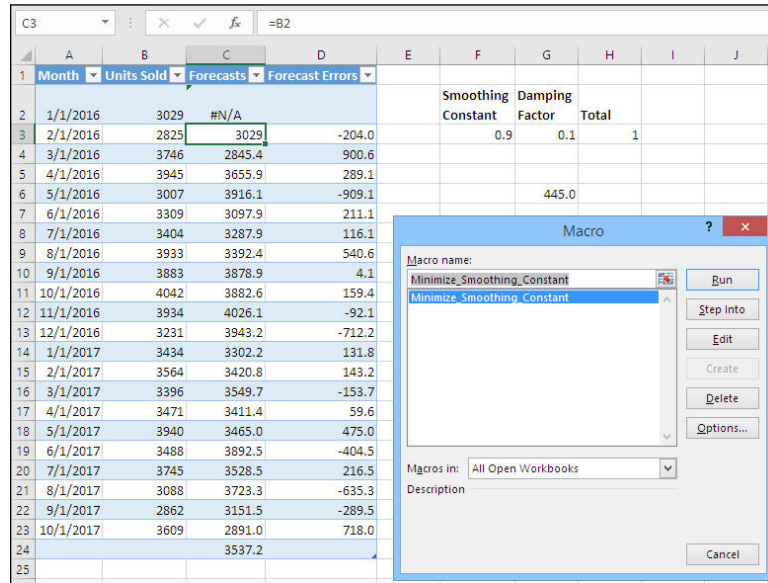


**FIGURE 15-12:**  
The Visual Basic Editor.

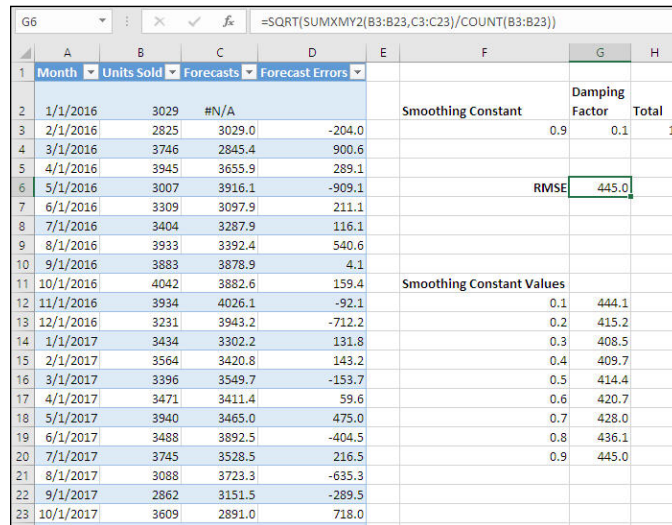
The results appear in Figure 15-13.

By looking at the root mean square errors in G12:G20, you can see that it's minimized when the smoothing is 0.3. So you can enter 0.3 into cell F3 and know that you've chosen the constant that minimizes the amount of error in your forecasts (see Figure 15-14).

**FIGURE 15-13:** You can walk through your code, statement by statement, if you click Step Into rather than Run.



**FIGURE 15-14:** You can use the information in cells F12:G20 to select the best value for the smoothing constant.



If you're familiar with the Solver, another add-in that comes with Excel, it might occur to you to use the Solver to find the optimum value of the smoothing constant — that strategy often works well. You would tell the Solver to minimize the value of the root mean square error, by changing the smoothing constant, and specifying criteria that the smoothing constant must be between 0 and 1. Using Solver in this way is just about mandatory in more complex smoothing situations, such as with trended and seasonal baselines, when you're using more than just the one smoothing constant discussed here.



WARNING

Whether you use a VBA routine or Solver to minimize the root mean square error, be aware that a *trended* baseline often results in a smoothing constant of 1.0. This in turn brings about what's called *naïve forecasting*, in which each forecast is exactly equal to the prior actual. If your baseline shows trend, whether up or down, you're better off using methods described in Chapters 17 and 18. One of these methods involves differencing the series to get a stationary baseline of differences, and forecasting the differences before integrating them back into the original baseline.

## Problems with Exponential Smoothing

Nothing's perfect, least of all a forecaster. Two of the problems with exponential smoothing aren't really too terrible. I give you the ugly little details in the following sections.

### Losing an observation at the start

You've probably noticed that the first value that the Data Analysis add-in's Exponential Smoothing tool returns is #N/A, or Not Available. That's because there's no observation prior to the first one on which you can base a forecast of the first time period.

And the first forecast you make is the actual value of the first time period. That's because subsequent forecasts are a combination of the prior actual and the prior forecast. But there's no forecast for the first time period, and the best you can do is to use the first period's actual.

This is better than the periods you lose with moving averages, where you lose one, two, or more time periods at the start, depending on how many actuals go into your moving averages.



Actually, a technique called *backcasting* can take you back prior to the first time period — kind of like asking what preceded the Big Bang. I don't cover it here, but if you really get into forecasting you should know that it exists so you can look into it.

## The Regression tool's standard errors: They're wrong

The Exponential Smoothing tool in the Data Analysis add-in has an option to output standard errors. Here's an example of the formulas it returns:

```
=SQRT(SUMXMY2(A2:A4,C2:C4)/3)
```

Look familiar? It's very similar to the formula developed earlier in this chapter for the root mean square error. The only real difference is that it restricts itself to three baseline values and their corresponding forecast values.

There's nothing in statistical or forecasting theory that holds that standard errors should be based on three values. Furthermore, the notion that different parts of the baseline have different standard errors is one on which there is no good consensus.

I don't know what happened to the Exponential Smoothing tool here. My best guess is that a developer was trying to add a Standard Errors capability to the tool, got it to work satisfactorily in the three-value case, and forgot to generalize it later.

Regardless of the reason that the Standard Errors tool is there, or that it's formulated as it is, my recommendation is that you not bother with it.





## Chapter 16

# Fine-Tuning a Regression Forecast

Other chapters, particularly Chapter 11, look at using regression to forecast one variable, such as sales revenue, from a predictor variable, such as time period or number of sales reps. This sort of analysis is sometimes termed *simple regression*. Forecasting one variable from more than one predictor is possible and sometimes useful. You may try forecasting sales revenue from *both* time period and number of sales reps. This approach is termed *multiple regression* and this chapter shows you how to do it in Excel.

Perhaps the most valuable aspect of Excel charts to come along since the charts themselves were introduced is the trendline. Using the trendline, you can, in one step, display the relationship between your predictor variables and your forecast variable. The trendline can be linear or nonlinear, or it can represent a moving average. The trendline can visually inform you about the direction and strength of the relationship of the predictors to the forecast. You can also choose whether to show underlying information such as the R-squared value and the regression equation itself.

Taking a number that's produced by a computer with no salt at all is tempting. Take it with a grain or even two. Apart from the usual warnings such as the venerable "garbage in, garbage out," there are other pitfalls in your path. You'll save yourself some grief if you evaluate your forecasts before you buy into them, so it's just as well that there are some tools available in Excel to help you do your due diligence.

# Doing Multiple Regression

Multiple regression is a way of using more than one predictor variable to forecast another variable, such as sales revenues. It can make your forecasts more accurate, but you have to know how to get Excel to do multiple regression. Not only do you have to know how to do it, you need to know how to interpret its results. The rest of this chapter shows you how.

## Using more than one predictor

How can you use more than one predictor at a time to forecast some other variable? Here's where multiple regression comes in. It combines, say, two predictor variables to form a new composite variable.

Suppose you were interested in forecasting the weight of men between their infancy and their 18th birthday. You may have, in an Excel worksheet, information on 30 men, conveniently including their weight, height, and age. If you direct Excel's attention to those three variables in just the right way, Excel's regression functions will do the following:

- » Combine height and age into a new variable.
- » Calculate that new variable so that it's the best possible predictor of young men's weight, given the information you made available to Excel.

For example, multiple regression might calculate that new variable like this:

$$\text{New Variable} = -20.34 + (3.92 \times \text{Age}) + (2.21 \times \text{Height})$$

Regression would choose the factors — usually termed the *coefficients* — for age and height (3.92 and 2.21, respectively) so that the new variable has the highest correlation with weight that is possible, given the data in your worksheet. The Data Analysis add-in's Regression tool itself doesn't show you this new variable, unless you choose one of the residual output options. But you could create it easily enough with your data on age and height, in combination with the factors that regression figures out for you. Or, with less effort, you could use Excel's TREND function to display the new variable. You'll find it in column E of Figure 16-1.



TIP

The language of forecasting in general (including that used in this book) tends to mean a baseline that's headed either up or down when it uses the term *trend*. Don't be misled by Microsoft's choice of the word *trend* to name the worksheet function TREND. The function works equally well for any linear baseline, whether its general direction is up, down, or sideways.

	A	B	C	D	E	F	G	H
1	Height (inches)	Age (years)	Weight (pounds)		Weight forecast			
2		31	3	44	59.88	=CORREL(C2:C25,E2:E25)		0.91
3		33	3	48	64.27			
4		35	3	60	68.67			
5		40	5	120	87.53			
6		45	6	134	102.46			
7		48	7	122	112.99			
8		49	6	142	111.24			
9		52	7	130	121.78			
10		52	9	132	129.65			
11		56	7	115	130.56			
12		58	8	137	138.90			
13		59	8	107	141.09			
14		59	9	118	145.03			
15		59	7	145	137.15			
16		60	9	125	147.23			
17		64	17	173	187.53			
18		64	10	168	159.96			
19		65	13	146	173.97			
20		68	17	184	196.32			
21		69	12	190	178.82			
22		70	14	210	188.90			
23		71	16	215	198.97			
24		71	16	220	198.97			
25		72	17	202	205.11			

**FIGURE 16-1:**  
You have to array-enter the TREND formula just as you do LINEST, with Ctrl+Shift+Enter.

The new variable is actually the forecast values for weight: Forecast by multiplying a person’s age by its coefficient, the person’s height by its coefficient, and adding the results to what’s termed a *constant* or *intercept*. Then it’s on to the next person to forecast *his* weight. Figure 16-1 shows this process in some detail.

Figure 16-1 shows you the situation that this section discusses: forecasting weight, given knowledge of age and height. (I know you’re not interested in forecasting weight, but using these variables makes the discussion easier to follow. I move on to sales forecasts with regression in “Interpreting the coefficients and their standard errors.”)

The basic data is in Figure 16-1 in columns A, B, and C. The TREND function is in column E: It shows the forecasts of weight, given age and height. Another way of saying that is that column E contains the combination of age and height that has the highest possible correlation with actual weight in this data set.

If all you were interested in were the forecasts, you might stop there. And that would be understandable, because it’s the forecasts you’re after. But you haven’t gone far enough yet. Looking at some of the other information from a regression analysis is important. This other information will tell you such things as

- » Whether forecasting weight from age and height is even worth doing
- » Whether you should use both age and height, or whether one or the other would be enough
- » How useful the regression equation is likely to be if you use it with a different set of inputs

This is important stuff, because it helps you decide whether and how to use the regression equation, whether you should look for some predictors other than age and height, how stable the equation will be as your baseline gets longer, and so on. A more complete analysis — one that shows more than just the forecast values — is shown in Figure 16-2.

	A	B	C	D	E	F	G	H	I
12				=IFS\$2*A2+\$ES\$2*B2+\$GS\$2					
	Height (inches)	Age (years)	Weight (pounds)	=LINEST(C2:C25,A2:B25,,TRUE)				Forecasts with LINEST	
2	31	3	44	3.939	2.197	-20.044		59.88	
3	33	3	48	2.090	0.776	27.562		64.27	
4	35	3	60	0.836	20.616	#N/A		68.67	
5	40	5	120	53.672	21	#N/A		87.53	
6	45	6	134	45623.186	8925.439	#N/A		102.46	
7	48	7	122					112.99	
8	49	6	142					111.24	
9	52	7	130					121.78	
10	52	9	132					129.65	
11	56	7	115					130.56	
12	58	8	137					138.90	
13	59	8	107					141.09	
14	59	9	118					145.03	
15	59	7	145					137.15	
16	60	9	125					147.23	
17	64	17	173					187.53	
18	64	10	168					159.96	
19	65	13	146					173.97	
20	68	17	184					196.32	
21	69	12	190					178.82	
22	70	14	210					188.90	
23	71	16	215					198.97	
24	71	16	220					198.97	
25	72	17	202					205.11	

**FIGURE 16-2:** The coefficients returned by LINEST in E2:F2 appear in reverse of the order in which the associated variables appear on the worksheet.

Column I in Figure 16-2 shows the forecasts of the Weight variable, given knowledge of the Age and Height variables. Notice that the forecast amounts are identical to those shown in Figure 16-1. Also notice that the forecasts are calculated by multiplying the regression coefficients returned by LINEST in E2:F2 and the observed Age and Height values, and adding the constant in cell G2. Evidently, the TREND function used in Figure 16-1 is simply a quicker and easier way to get the forecasts than using the coefficients and constant returned by LINEST in Figure 16-2.

Figure 16-2 also shows how the LINEST function gives you some of the information you need to interpret a multiple regression. Here are the steps you take to get the LINEST function on the worksheet. The steps assume that you have the data laid out as in Figure 16-2. If your data is in different columns or rows, just change the addresses used in Steps 1 and 2 accordingly.

**1. Count the number of predictor variables, and add 1.**

In Figure 16-2, there are two predictor variables (height and age), so you get 3.

2. **Select a range of cells — blank ones, unless you have some cells with data you don't care about — that has five rows and as many columns as the number you got in Step 1.**

In this example, then, you'd select a range of cells five rows high and three columns wide; Figure 16-2 uses E2:G6.

3. **Type =LINEST(C2:C25,A2:B25,,TRUE), but *don't* press Enter yet.**
4. **Press Ctrl+Shift+Enter.**

This *array-enters* the formula. LINEST is one of the Excel functions that you must array-enter so as to get the right results.



REMEMBER

If, as here, you're array-entering a formula that will occupy a range of cells, begin by selecting the full range. Excel doesn't figure out the dimensions of the range and fill them in on your behalf. As terrific an application as Excel is, there are a few areas in which it can be remarkably obtuse, and this is one of them.

Here's what LINEST shows you:

- » The first row always has what are called the *coefficients* and the *intercept* (or *constant*). These are the numbers you use along with your actuals to create your forecasts. In Figure 16-2, they're in cells E2:G2.
- » The second row always has the *standard errors* of the coefficients and the intercept. These help you decide whether to pay attention to a variable when you're creating a forecast. In Figure 16-2, they're in cells E3:G3.
- » The third through the fifth rows have useful information in only the first and the second columns. The remaining columns always show the error value #N/A in the third, fourth, and fifth rows. In Figure 16-2, that useful information is in cells E4:F6.

What is that useful information? Here's an overview. And I do mean overview. Even intermediate statistics texts have entire chapters devoted to each one of these topics.

## Squaring R

In the third row, first column of the LINEST results you'll find the *R-squared* value. This is the square of the correlation coefficient between the actuals and the forecasts. In Figure 16-2, it's the square of the correlation between the values in C2:C25 (the actuals) and the values in I2:I25 (the forecasts). In this case, the R-squared value is 0.836. That's the square of the *multiple R*, calculated explicitly in cell H2 of Figure 16-1.

The R-squared value is actually a percentage. It tells you what percent of the variation — the *spread* — in the forecast variable that you can attribute to variation in the predictor variables. In this example, you can attribute 83.6 percent of the variation in actual weight to the combination of age and height. So, differences in weights are associated with differences in age and height.

If that sounds to you suspiciously like the meaning of a correlation, you're right. R-squared is called that because it's the square of the correlation coefficient (the correlation coefficient is often called  $r$ , in lower case, for short while the multiple correlation is always  $R$ , in upper case).

The higher the R-squared, the better the job your predictor variables are doing as forecasters. Because it's a squared number, R-squared can never be negative. And because the maximum value of a correlation coefficient is 1.0, R-squared itself can never be greater than 1.0.

So, in the LINEST results, look for the value of R-squared in LINEST's third row, first column. The closer the number you see there is to 1.0, the better your forecast. The closer to 0.0, the worse.

## Nobody's perfect . . .

And neither is LINEST. The value in LINEST's third row, second column is the *standard error of estimate*, and it helps you understand how much error is involved in the forecasts you make using LINEST.

In Figure 16-2, the standard error of estimate is found in cell F4, and it's about 20.6. If you go up and down by two standard errors from any forecast, you'll bracket that forecast within two numbers, and you can be just about 95 percent confident that an actual observation will be between those two numbers.

For example, suppose you wanted to forecast the weight of a young male whose age was 13 years and whose height was 64 inches. The equation that LINEST returns in Figure 16-2 would forecast this:

$$-20.044 + (3.939 \times \text{Age}) + (2.197 \times \text{Height}) = \text{Forecast Weight}$$

$$-20.044 + (3.939 \times 13) + (2.197 \times 64) = 171.78$$



REMEMBER

Chapter 12 discusses the fact that LINEST displays the coefficients in the *reverse* order that they appear on the worksheet. That's why you use 3.939 for Age and 2.197 for Height, even though on the worksheet the values for Height come first and Age second, and the coefficients' order is the other way around.

Now, the standard error of estimate in this case is 20.6. So if you were to make this forecast, you could be 95 percent confident that the person's actual weight would be between 130.6 and 213.0:

$$171.8 - 2 \times 20.6 = 130.6$$

$$171.8 + 2 \times 20.6 = 213.0$$



People often misinterpret the phrase “95 percent confident.” It doesn't mean that the probability is 95 percent that the actual value lies between the lower limit of 130.6 and the upper limit of 213.0. That probability is either 0.0 (it doesn't lie within those limits) or 1.0 (it does). But if you got these measurements on thousands of young males and ran LINEST on them, 95 percent of those people would have an actual weight within two standard errors of their forecast weight.

As you may guess, the larger the R-squared value (a measure of the accuracy of the forecast), the smaller the standard error of estimate (a measure of the inaccuracy of the forecast).

## The other LINEST statistics

Rows 4 and 5, columns 1 and 2 of the LINEST results are fairly esoteric, and unless you feel comfortable with intermediate-level statistical analysis, I suggest that you ignore them. They're the building blocks for further analyses that tell you whether you can regard the results as “statistically significant.”

There are probably three categories of users' interest in this sort of thing:

- » **Low interest:** You should just skip this stuff and use the TREND function to get your forecasts.
- » **Moderate interest:** You should use the Data Analysis add-in to get the information about your regression equation. It does those further analyses for you. See Chapter 11 for information on interpreting the results you get from the Data Analysis add-in's Regression tool.
- » **High interest:** This is the Glutton for Punishment category. If you're deranged enough to want to know the details, check out the following list:
  - **The F ratio:** The F ratio divides the mean square (regression) by the mean square (residual). The F ratio is in the fourth row, first column of LINEST's results. If you use Excel's F.DIST function along with the F ratio, the number of predictor variables, and the residual degrees of freedom (see the next item), you can determine the statistical significance of the regression. This is the same as testing whether R-squared is significantly different from zero.



- **Residual degrees of freedom:** The degrees of freedom (or *df*) is found in LINEST'S fourth row, second column. It tells you what to divide by to convert the residual sum of squares to a residual mean square. It also tells you the third argument to F.DIST.
- **Regression sum of squares:** The sum of squares (or *SS*) regression is in LINEST's fifth row, first column. Divide it by the number of predictor variables to get the mean square regression.
- **Residual sum of squares:** The *SS* residual is in LINEST's fifth row, second column. Divide it by the residual degrees of freedom to get the mean square residual. Divide the mean square regression by the mean square residual to get the F ratio.

## The thinking person's approach to multiple regression

The previous section mentions the Regression tool in the Data Analysis add-in as a middle-of-the-road approach to getting a regression forecast. It provides you with most of the LINEST results that you'd want, both to generate forecasts and to diagnose how well your regression equation does at forecasting. So, if you decide to use the Data Analysis add-in, you haven't decided to ignore LINEST's results completely.

At the same time, you've decided that you don't want to gin up all these diagnostics yourself, basing them on the LINEST results. You could do so, that's true, but it's the sort of thing only a purist would do, and the result is something only a mother could love.

Chapter 11 gives you an overview of the parts of the Regression tool's output that are the really critical ones to examine before you go public with a forecast. Here's a review, along with a *really* brief overview of each:

- » **Multiple R:** This is, redundantly, the square root of R-squared. You can look at either to judge the accuracy of the regression equation. If you're more comfortable thinking in terms of correlation coefficients, look at the Multiple R. If you're more comfortable thinking in terms of proportions of shared variance, look at R-squared.
- » **Intercept and coefficients:** These are the numbers you apply to the values of the predictors that you have in hand, in order to get the best forecast that's available to you.



- » **Confidence levels:** These bracket the intercept and coefficients, giving you a sense of how much they might jump around if you got other samples of the predictor variables and the forecast variable.
- » **Residual plot:** Look at this chart to see whether you've got something funny going on with the error values.
- » **Line fit plot:** This chart plots the actual and the forecast values against the predictor's values. It's less useful in a multiple regression situation than with a single predictor, because as the chart is designed it can't handle more than one predictor at once.

## The incredible shrinking R-squared

Figure 16–3 shows an example of the output, based on the same Age, Height, and Weight data that I've been showing you.

**FIGURE 16-3:**  
You get a lot more bang for your buck by using the Regression tool than from the LINEST function, but it has some drawbacks too.

	A	B	C	D	E	F	G	H	I	J	K
1	Height (inches)	Age (years)	Weight (pounds)		SUMMARY OUTPUT						
2	31	3	44								
3	33	3	48		Regression Statistics						
4	35	3	60		Multiple R	0.915					
5	40	5	120		R Square	0.836					
6	45	6	134		Adjusted R Square	0.821					
7	48	7	122		Standard Error	20.616					
8	49	6	142		Observations	24					
9	52	7	130								
10	52	9	132		ANOVA						
11	56	7	115			df	SS	MS	F	Significance F	
12	58	8	137		Regression	2	45623.186	22811.593	53.672	0.000	
13	59	8	107		Residual	21	8925.439	425.021			
14	59	9	118		Total	23	54548.625				
15	59	7	145								
16	60	9	125			Coefficient	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
17	64	17	173		Intercept	-20.044	27.562	-0.727	0.475	-77.361	37.273
18	64	10	168		Height (inches)	2.197	0.776	2.832	0.010	0.584	3.810
19	65	13	146		Age (years)	3.939	2.090	1.885	0.073	-0.407	8.286
20	68	17	184								
21	69	12	190								
22	70	14	210								
23	71	16	215								
24	71	16	220								
25	72	17	202								

The Adjusted R-squared appears in cell F6 of Figure 16–3. Another term for Adjusted R-squared is the Shrinkage Estimator, and it has nothing to do with *Seinfeld*. Nevertheless, if you're forecasting sales, you're going to want to pay attention to it.

Suppose you get data on another sample of men and use the intercept and coefficients that you got in *this* sample with the new sample. In other words, you could

take the values in cells F17 through F19 of Figure 16-3 and use them along with the age and height of the men in your new sample, to forecast their weight.

Then you could find the correlation of these new forecasts with the new actuals. (This is what the Multiple R in Figure 16-3 is giving you: the correlation between the forecasts and the composite of the actuals in your original sample.)

The new Multiple R will almost always be smaller than the original Multiple R — that is, it *shrinks*. The reasons for this are a little arcane, but they have to do with some capitalization on chance in the original sample. Why should you be concerned with this? There are a couple of good reasons.

In sales forecasting, you constantly have new data coming in. Shortly after you make your July forecast, or your second-quarter forecast, the next set of actuals comes in and it's time to make a forecast for August, or for the third quarter. Of course, you want to compare those new actuals with the prior period's forecasts. You have to do that in order to tell whether your forecasts are spot on, useless, or something in between.

If you pay attention to the Adjusted R-squared, you have a sort of worst-case estimate of how much the accuracy of the regression forecast could drop if you applied the regression equation to a batch of entirely new observations. Of course, in a forecasting situation, only one new period is coming in, not the replacement of the entire set of data. If the Adjusted R-squared is still acceptable to you, you can feel reasonably comfortable that your forecasts will continue to be reasonably accurate.

The other important reason to pay attention to the Adjusted R-squared is that it's sensitive to the relationship between the number of observations in the baseline and the number of predictor variables in the equation. Here's the formula:

$$\text{Adjusted R-squared} = 1 - [(1 - \text{R-squared}) \times ((N - 1) \div [N - K - 1])]$$

where N is the number of observations and K is the number of predictor variables in the analysis. A good way to conceptualize this is to keep in mind that the larger the number of observations relative to the number of predictors, the more accurate the forecasting equation.

To see how the formula works out with sales data, see Figure 16-4. In Figure 16-4, I've deployed the Data Analysis add-in's Regression tool on a data set that shows revenues, dollars spent on advertising, size of sales force, and the discount offered to customers (the discounts were based on the need to reduce inventory). The Regression tool's output appears in the range F1:L20.

**FIGURE 16-4:** The SUMMARY OUTPUT table contains static values, not formulas, so you have to rerun the Regression tool if your inputs change.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Revenues (\$000)	Advertising Dollars	Sales Force	Discount %		SUMMARY OUTPUT						
2	\$ 1,519	\$ 344	5	17%		Regression Statistics						
3	\$ 3,955	\$ 534	11	27%		Multiple R	0.791					
4	\$ 4,340	\$ 3,834	14	26%		R Square	0.626					
5	\$ 8,413	\$ 2,798	54	37%		Adjusted R Square	0.555					
6	\$ 371	\$ 795	11	5%		Standard Error	1811.845					
7	\$ 706	\$ 3,534	13	1%		Observations	20					
8	\$ 255	\$ 2,143	21	12%		ANOVA						
9	\$ 762	\$ 1,518	38	1%			df	SS	MS	F	Significance F	
10	\$ 6,758	\$ 207	25	33%		Regression	3	87732076.672	29244025.557	8.908	0.001	
11	\$ 352	\$ 289	5	38%		Residual	16	52524496.328	3282781.021			
12	\$ 429	\$ 1,022	30	17%		Total	19	140256573.000				
13	\$ 3,068	\$ 5,195	56	4%								
14	\$ 5,775	\$ 5,489	27	36%								
15	\$ 3,664	\$ 7,093	21	1%								
16	\$ 497	\$ 405	10	1%								
17	\$ 592	\$ 627	4	14%								
18	\$ 5,937	\$ 3,736	36	33%		Intercept	-1211.183	962.353	-1.259	0.226	-3251.280	828.914
19	\$ 3,007	\$ 560	32	24%		Advertising Doll	0.413	0.218	1.893	0.077	-0.050	0.876
20	\$ 4,974	\$ 924	56	26%		Sales Force	34.543	25.437	1.358	0.193	-19.381	88.468
21	\$ 7,616	\$ 1,282	3	31%		Discount %	13908.522	3134.294	4.438	0.000	7264.115	20552.929
22												
23						Number of Predictors	Adjusted R-square					
24						3	0.56	=1-(1-(\$G\$5)*(((\$G\$8-1)/(\$G\$8-F23-1)))				
25						4	0.53	=1-(1-(\$G\$5)*(((\$G\$8-1)/(\$G\$8-F24-1)))				
26						5	0.49	=1-(1-(\$G\$5)*(((\$G\$8-1)/(\$G\$8-F25-1)))				
						6	0.45	=1-(1-(\$G\$5)*(((\$G\$8-1)/(\$G\$8-F26-1)))				

Applying the formula for the Adjusted R-squared in cell G23, you get 0.56, just as reported by the Regression tool in cell G6. Adding more predictors to the equation causes the Adjusted R-squared to drop quickly, all the way to 0.45 with six predictors, as shown in cell G26.

Now, bear in mind that this is an “other things being equal” analysis. For example, if one of the variables you add to the regression analysis has a perfect 1.0 correlation with sales revenues, then the R-squared value would jump to 1.0, making a liar of both the Adjusted R-squared and me. But first go find a variable that predicts sales revenues perfectly, and *then* write me.

## Counting noses

The Regression tool’s output also shows the number of observations — in a forecasting context, that’s the number of records in the baseline — that went into the regression analysis. In Figure 16-4, you’ll find that in cell G8. It can come in handy if you’re figuring Adjusted R-squared values. Notice that the equations for Adjusted R-squared in the range G23:G26 all rely on G8 to provide the count of observations.

## Degrees of freedom

Two values are of interest: the degrees of freedom for the regression and the degrees of freedom for the residual. Getting a good grasp on the *why* of degrees of freedom can take quite a bit of study and head-scratching. For now, let it go at knowing that you divide the sums of squares in H12:H13 by their respective degrees of freedom to get the mean square.

## Mean squares

A *mean square* (MS) is just another term for a variance. (One definition of a variance is the average squared deviation of each observation from the mean of the data set. Hence, *mean square*.) You divide the MS for the regression by the MS for the residual. The result is the F ratio.

## The F ratio

The F ratio is what you actually test to help you decide whether the regression is significant: more precisely, whether the R-squared value is significantly different from zero. This chapter touches on the issue in “The other LINEST statistics.”

## Significance of F

The discussion of LINEST earlier in the chapter mentions how to use the F.DIST function, along with the F ratio and the degrees of freedom, to determine how confident you can be that the true R-squared is greater than zero. The Data Analysis add-in’s Regression tool does that for you. In Figure 16-4, you can find the significance level in cell K12. The smaller the significance level (and 0.001 is quite small), the more confident you can be that you have a statistically significant regression.

That may sound important, and perhaps it is. But all it really means is that the true R-squared, the one you would calculate if you had access to all possible baseline observations in the population of your periodic sales revenues, is very unlikely to be zero.



REMEMBER

All these esoteric statistics, the SS and the df and the MS, for the regression and for the residual, and the F ratio, are shown by the Data Analysis add-in’s Regression tool as a matter of convention. Since Sir Ronald Fisher invented the Analysis of Variance (or, as it’s labeled in cell F10 of Figure 16-4, ANOVA), it’s been traditional to show all these values in an ANOVA table. The Data Analysis add-in is just following convention. It’s nice to know these numbers, but only the Significance of F (if even that) is necessary for you to decide whether the relationship between the predictors and the sales revenues is reliably different from 0.0.

## Interpreting the coefficients and their standard errors

The final section of the Regression tool’s output concerns the intercept and the coefficients that you use to make your forecast. In Figure 16-4, you’ll find the intercept in cell G17 and the coefficients in cells G18:G20.



TIP

Even though it's found in a column labeled *Coefficients*, the Intercept is not a coefficient. It's just a number that you add into the regression equation as a scaling adjustment. The coefficients are the numbers by which you multiply your predictor variables.

The standard errors that are associated with the coefficients and the intercept help you gauge whether they really belong in the equation. These standard errors are used much like the standard error of estimate, which I discuss in the “Nobody’s perfect” section, earlier in the chapter. Adding two standard errors to the coefficient and subtracting two standard errors from the coefficient, then seeing if the resulting range spans zero, is typical. Plus and minus two standard errors is a bracket — a span of values. You could also construct a bracket defined by plus and minus *three* standard errors. Figure 16–5 shows how this works with the sales data in Figure 16–4.

	A	B	C	D	E	F	G	H	I
1									
2				<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
3			Intercept	-1211.183	962.353	-1.259	0.226	-3251.280	828.914
4			Ad	0.413	0.218	1.893	0.077	-0.050	0.876
5			Sales Force	34.543	25.437	1.358	0.193	-19.381	88.468
6			Discount %	13908.522	3134.294	4.438	0.000	7264.115	20552.929
7									
8									
9			Predictor	Coefficient	Coefficient minus 2 Standard Errors	Coefficient plus 2 Standard Errors	Spans zero?		
10			Intercept	-1211.183	-3135.889	713.522	Yes		
11			Ad	0.413	-0.023	0.850	Yes		
12			Sales Force	34.543	-16.331	85.418	Yes		
13			Discount %	13908.522	7639.933	20177.111	No		

**FIGURE 16-5:** If a coefficient's bracket spans zero, you need to deal with the possibility that it really *is* zero.

All those columns in the range C2:I6 are telling you the same thing: Only the Discount variable is reliably related to sales revenues — at least, in the data set that the regression was calculated on.

In particular, if you add and subtract two standard errors from the intercept and the coefficients, each of the resulting brackets spans zero, except for Discount. That means you can't depend on it that the true population coefficients for Advertising and Sales Force *aren't* zero. And if you use a zero coefficient in the equation, along with an intercept of zero, then the equation could change from this:

$$\text{Revenue} = -1211.18 + (0.41 \times \text{Ad } \$) + (34.54 \times \text{Sales Force}) + (13908.52 \times \text{Discount})$$

to this:

$$\text{Revenue} = 0 + (0 \times \text{Ad } \$) + (0 \times \text{Sales Force}) + (13908.52 \times \text{Discount})$$

That is, the intercept, Advertising Dollars, and Sales Force drop out of the equation because you don't have reason to believe that the true coefficients and the true intercept are nonzero.

Perhaps a more familiar term for the brackets this section discusses is *confidence intervals*. The Regression tool calculates them for you. You'll find the lower and upper limits of the regression coefficients' 95 percent confidence intervals in the range H3:I6 of Figure 16-5. They will not precisely match the limits you'll get from going up and down two standard errors from the calculated coefficients, but they'll be close. "Two standard errors" is just a close approximation. (The discrepancy comes about because the number of standard errors that define a confidence interval's limits varies slightly with the number of observations in the regression analysis.)

Chapter 4 mentions that one of the goals of the regression approach to forecasting is *parsimony*: The fewer the predictor variables, the better. On that principle, if you can get rid of Advertising Dollars and the size of the Sales Force and still have a good forecast, that's useful. At the very least, you save the time and cost of collecting the information on those predictor variables each month or quarter. And thinking back to what this chapter has to say about the Adjusted R-squared and shrinkage, restricting the number of predictor variables may be a good idea anyway.

Figure 16-6 shows you another way to look at things. In fact, the remainder of this section shows you how you can use the information in the Regression tool's output to demonstrate whether a coefficient is to be trusted (that is, if it's significantly different from zero) or not (if it's not significantly different from zero). So far, I've given you a brief look at brackets by using standard errors. Next I take you further into t statistics and confidence intervals. In your own forecasts, they should all reach the same conclusions. If they don't, you should take a closer look at what they're telling you.

	A	B	C	D	E	F	G	H	I
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									

Predictor	Coefficient	Standard Error	Coefficient over Standard Error	T.DIST
Intercept	-1211.183	962.353	-1.259	0.226 =T.DIST.2T(ABS(F10),16)
Ad	0.413	0.218	1.893	0.077 =T.DIST.2T(ABS(F11),16)
Sales Force	34.543	25.437	1.358	0.193 =T.DIST.2T(ABS(F12),16)
Discount %	13908.522	3134.294	4.438	0.000 =T.DIST.2T(ABS(F13),16)

**FIGURE 16-6:** Compare the values labeled *P-value* (in G2:G6) with the values under T.DIST (in G10:G13).

A statistic called Student's *t* statistic (or just *t* statistic or *t* ratio) helps you assess the statistical significance of the difference of a number from zero. The term *statistical significance* as used here simply means the probability of getting a regression coefficient as large as the one you've observed in your sample if the coefficient in the full population is zero.

Suppose you use regression to evaluate the relationship between a person's height and his zip code. You take a sample of people and find that, in the sample, the *t* statistic for zip code is 3.5. What's the probability of getting that large a *t* statistic in your sample when the same statistic, if calculated on the entire population, is zero? Put differently, what's the probability of getting a really large *t* statistic based on a bad sample from a population where there's no relationship in that population between a person's height and his or her zip code? That's what statistical significance assesses.

In this situation, you calculate the *t* statistic by dividing the number (the intercept or the coefficient) by its standard error. This has been done in F10:F13 of Figure 16-6, and you can tell that the results are identical to the *t* statistics that the Regression tool returns (F3:F6 in Figure 16-6).

The Regression tool also tests the statistical significance of the *t* statistic, and the results are shown in G3:G6. You interpret these results in the same way as you do the significance of *F*. (See the "Significance of *F*" section earlier.)

In the case of the intercept and the coefficients for Advertising Dollars and Sales Force, the *p*-values reported by the Regression tool are not below the level of 0.05 that has traditionally been taken as the criterion for statistical significance. The Discount predictor variable, however, has a *p*-value smaller than 0.05, and so by that criterion this analysis says that its coefficient is significantly greater than zero. (But it's up to you to decide whether one chance in 20, or 0.05, is sufficiently unusual to decide that the Discount regression coefficient is really nonzero. You really should decide on that criterion before you see the results of your analysis.)



REMEMBER

The *smaller* the *p*-value (or, in the ANOVA table, the Significance of *F*), the *more* significant the number being tested.

These *p*-values have also been calculated in the range G10:G13 by using Excel's T.DIST function. Notice that the *p*-values are identical to those returned by the Regression tool.



WARNING

Excel's T.DIST function cannot cope with a negative *t* statistic. If you use the T.DIST function, you should consider using it in conjunction with Excel's ABS function, which returns the absolute value of a number. (The absolute value is always positive: The absolute value of  $-2.6$  is  $+2.6$ .) If you don't use the ABS function, as shown in cells G10:G13 of Figure 16-6, T.DIST will return the #NUM! error



value if the first argument is negative. This issue is a little tricky, and if you're not a statistician, find one (and you might mention that I sent you). You can get caught up in the toils of things called nondirectional hypotheses.

So, this t statistic analysis has the same outcome as the brackets-spanning-zero analysis: You're in good shape, using Discount to predict these sales revenues, but there's no argument for using the other variables when their 95% confidence intervals span zero, or equivalently when the t statistic's p-value is greater than 0.05.

Finally, consider the information you get from the Regression tool's confidence interval analysis (see Figure 16-7).

	A	B	C	D	E	F	G	H	I	J
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										

	A	B	C	D	E	F	G	H	I	J
2										
3										
4										
5										
6										
9										
10										
11										
12										
13										

**FIGURE 16-7:** Using Student's t statistic to create 95 percent intervals.

As you may expect from the earlier analyses, the intercept, the Advertising Dollars and the Sales Force predictors have brackets that span zero: The upper 95 percent is in each case greater than zero, and the lower 95 percent is below zero (see cells H3:I5 in Figure 16-7). So you can't reject the possibility that the true population coefficients (and the intercept) are in fact zero. This once again leads you to believe that the best forecast is based on the Discount predictor alone.

The difference between the upper/lower 95 percent brackets and the brackets that were created with +/- two standard errors is that the former are based on a t value of 2.1, rather than an even 2.0. As I noted earlier, the reasons for this are a little complicated, but they have to do with the fact that the value of the t statistic is partly a function of sample size, and is not a constant that's independent of the number of observations.



The T.INV function that I've used here is very different from Excel's older TINV function, as to both arguments and results. If you're a longtime user of Excel, be sure that your arguments and expectations for T.INV are up to date.



So, you can calculate the upper and lower 95 percent values using the value returned by T.INV, as shown in F10:F13 of Figure 16-7. With 0.025 as the first argument (which is  $(1 - 0.95)/2$ ; if you wanted a 90 percent value, you'd use 0.05 rather than 0.025) and the number of degrees of freedom as the second argument, you get a value of 2.12. Multiplying 2.12 times the standard error and adding the result to the coefficient or intercept, you get the upper 95 percent values. You get the lower 95 percent values by subtracting rather than adding.

In this example, all three methods point to the same conclusion: Use Discount as your predictor variable; regression will calculate a new intercept for you. There are some close cases in which the methods do not agree (although the t statistic method and the 95 percent bracket method will nearly always point to the same conclusion). When they don't agree, your best bet is to get more data: Either dig farther back into the past, or wait for some more actuals to come in. More data often brings a fuzzy outcome into focus.

And bear in mind that the formulas shown or implied in Figures 16-5 through 16-7 give you a way to let the analyses update if you change your input values. This is the main reason that, in the long run, I prefer to use my own formulas instead of relying on the static values provided by the Data Analysis add-in.

## Getting a Regression Trendline into a Chart

Looking at your regression forecast on a chart is almost always helpful so that you can visually compare your baseline, what's already happened, with what regression has forecast is going to happen next. That's the role of trendlines, and this section shows you how to get them.

It's never, ever a good idea to just accept a computer-generated forecast (or any sort of analysis, for that matter) at face value. You need to look at some of the underlying statistics to judge whether the forecast is sense or nonsense. In the previous section, I show you how to use statistics that Excel reports back to you so as to make that kind of judgment.

But numbers don't tell the whole story. Figure 16-8 shows a situation in which the number alone would lead you astray, but looking at the data in a chart would put you straight again.

**FIGURE 16-8:** Eta squared is a generalized version of R-squared. It returns the same value for a linear relationship and is a more accurate index of the strength of a nonlinear relationship.

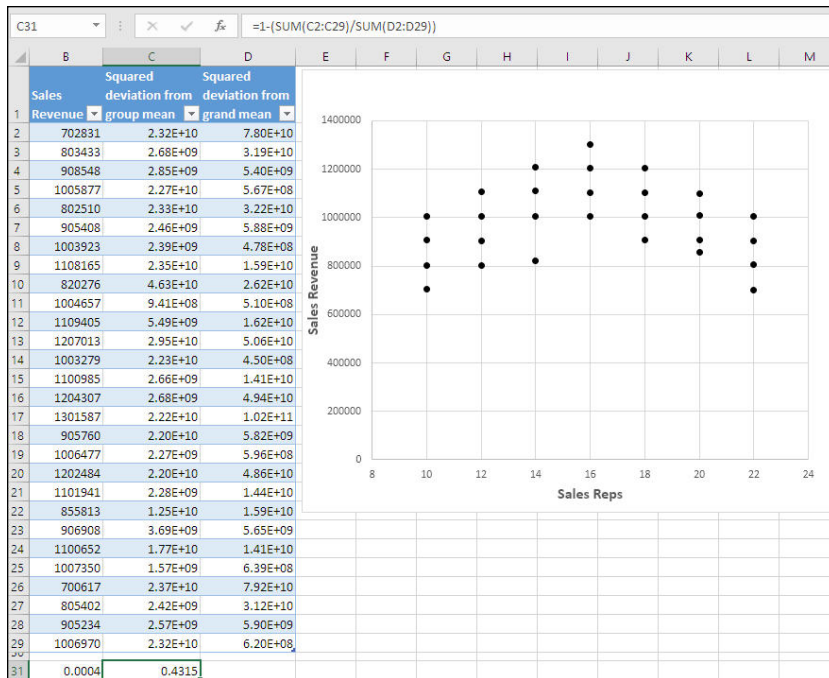


Figure 16-8 depicts a clear relationship between the number of sales reps and the sales revenue. Up to a point, the more reps you have, the greater the revenues. Beyond 16 reps, though, the revenues actually start to fall. This can happen for several reasons, among them:

- » Territories are defined both geographically and by national accounts. In these cases, Jamie's national account might well have a presence in Jim's geographical territory. Without double commissions, this situation leads to competition *within* the sales office, and not of the good sort.
- » The sales territory can support up to, but not more than, a given number of sales reps.
- » A third variable, such as a product line that is losing market share, prompts management to throw more reps at the territory, instead of studying the marketability of the product itself.

The correlation coefficient, also discussed in Chapter 14, is a *linear* statistic. In other words, it assumes that the two variables you hand off to it have a linear, straight-line relationship with each other — something such as a person's height and weight. The relationship need not be perfect, but in general the higher one variable, the higher the other variable.



REMEMBER

The relationship can work the other way around, such as that between frequency of car accidents and driver's age in years (up to age 30, say). The older the driver, the fewer the accidents. This would result in a negative correlation but can nevertheless be a strong one.

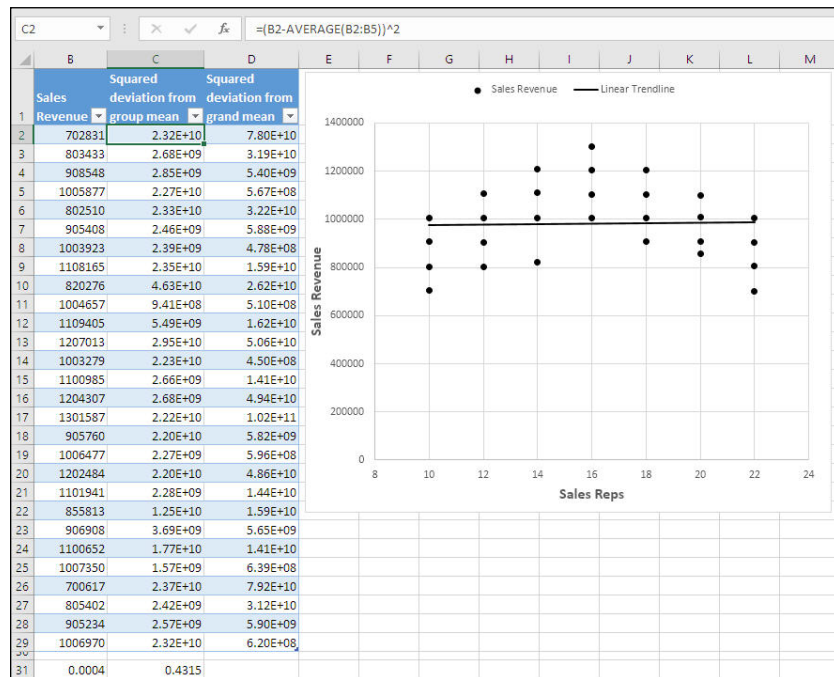
In Figure 16-8, the R-squared between sales reps and sales revenues appears in cell A31. You can see that it is, in effect, zero. But the nonlinear relationship, measured by a statistic called *Eta Squared* or the *correlation ratio*, in cell B31, agrees with the chart: Both the chart and cell B31 suggest strongly that there is a dependable relationship between number of sales reps and the amount of sales revenue — just not a linear one.



TIP

Cell A31 in Figure 16-8 calculates R-squared as the square of CORREL. If you find yourself calculating R-squared frequently (as I do) you may find it more convenient to use Excel's RSQ function, which calculates R-squared directly. In this case, you would use `=RSQ(A2:A29, B2:B29)`.

Another way of looking at things is to dispense with the numeric analysis and just look at the chart. The linear trendline puts you in a position to evaluate how well a linear analysis fits the data (see Figure 16-9).



**FIGURE 16-9:** This is an extreme example, used to make a point. You could tell you've got a nonlinear situation even without the trendline.

You get the trendline by following these steps:

**1. Click the chart to activate it.**

Notice that the Design tab appears under Chart Tools in the Ribbon.

**2. Right-click the charted data series and choose Add Trendline from the shortcut menu.**

The Format Trendline (*sic*) dialog box, shown in Figure 16-10, appears.

**3. Choose Linear under Trendline Options. If you want to stop at this point, click a cell in the worksheet.**

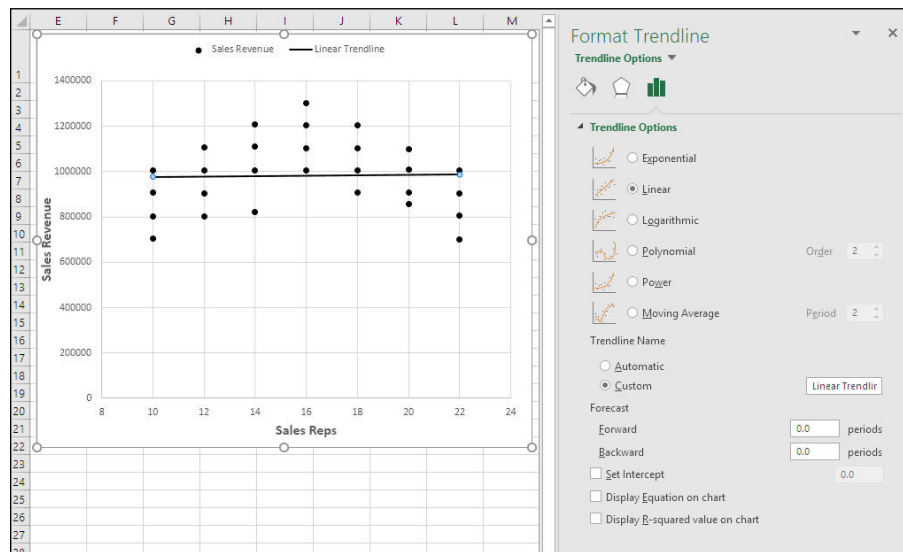
Notice in Figure 16-9 that the trendline is virtually horizontal.



TIP

If you have more than one data series in a chart, first select the data series for which you want the trendline, in Step 1 in the preceding list.

**FIGURE 16-10:** You can easily over-model a baseline by using nonlinear trendlines. You need a really good reason to do so — usually a statistical test called “goodness of fit.”



TIP

When there's a strong correlation between two variables, the charted data points lie close to the trendline. The very mathematics of correlation bring this about.

If you know when you first call for the trendline that you'll want additional information, you can continue from here after you've chosen the type of trendline in the preceding list. If necessary, scroll down to the bottom of the Format Trendline pane.

The options available to you are as follows:

- » You can give the trendline a more descriptive name than its default. This is useful primarily if you're including a legend on the chart, where you can display the name you've chosen. Select the Custom radio button and type the name into the text box.
- » If you've chosen any type of trendline other than Moving Average, you can extend the forecast forward, into the future, or backward, into the past. Click the up arrow on one of the spinners to increase the number of periods to forecast, and the down arrow to reduce the number of periods.
- » You can set the intercept to a particular value. Select the Set Intercept check box and type a value in the box. I don't recommend using this option.
- » You can display the equation on the chart. Select the associated check box.
- » You can display the R-squared value on the chart. Again, select the associated check box. (This is the same thing that the Data Analysis add-in terms R squared: the proportion of variability in the forecast variable that's attributable to the predictor variables.)

And here's some additional information about the intercept, the trendline equation, and the R-squared value:

- » Setting the intercept manually is not recommended. Beginning analysts find that they can increase the R-squared by setting the intercept to, say, zero. But they're looking at a quirk in the mathematics of regression. If your data set really does have a zero intercept, the analysis will put the intercept either at zero or close enough.
- » You can move the equation and the R-squared labels around on the chart, to get them out of the way of other elements such as gridlines. Click on the label and drag it wherever you want.
- » You can adjust the number of decimals and the font size of both the equation and the R-squared value. Right-click the label and choose Format Trendline Label. Choose Number from the Category drop-down and adjust the Decimal Places check box.



WARNING

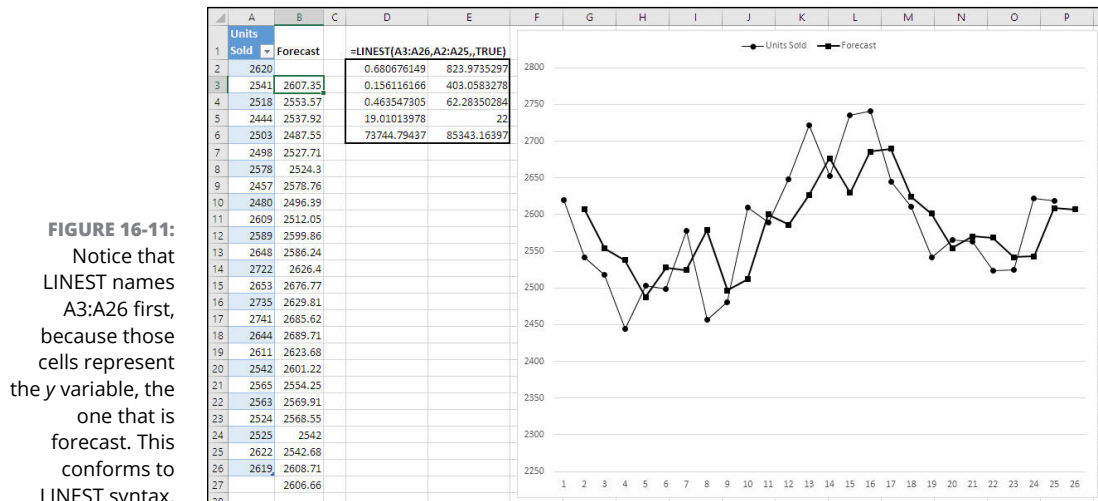
Some people try to forecast values by using the equation that optionally accompanies the trendline. That's a mistake. The intercept and the coefficients almost certainly do not display enough decimals, it's too easy to make a typing error when you transcribe the numbers to the worksheet, and it's a waste of your time. Use `LINEST` instead to get the intercept and coefficients directly on the worksheet, or use `TREND` to bypass the equation completely and get the forecast values in one step. See Chapter 12 for information on using the `TREND` function.

# Evaluating Regression Forecasts

When you make a regression forecast, you should look out for some problems beyond those mentioned in other chapters, such as independence of errors. Two of the more important problems that can arise concern autoregression and using two trended series.

## Using autoregression

The topic of autoregression has come up in several other chapters in this book. Briefly, when you use autoregression, you use one set of values in your baseline to predict another set of values in the same baseline. It's a lot easier to see autoregression than to read about it, so have a look at Figure 16-11.



In Figure 16-11, the LINEST formula (in cells D2:E6) uses the values in A2:A25 as the predictor variable, and the values one row down, A3:A26, as the predicted variable. In effect, what you're asking LINEST to do is to forecast each value in the baseline from the prior value. In words, you're saying, "Please forecast the second value in the series from the first value, given what I know about the relationship between A2:A25 and A3:A26. Then forecast the third value from the second, and so on."

In one way at least, autoregression is similar to simple exponential smoothing as performed by the Data Analysis add-in: It uses a prior period's value to help forecast the next period's value.

One of the differences, though, between autoregression and simple exponential smoothing is that simple exponential smoothing always uses the prior actual in the baseline to help forecast the next period. In autoregression, you may want to forecast using not the prior time period's value, but the value that's two or even three periods back.

Using the LINEST equation in Figure 16-11, you can get the forecasts shown in column B from the coefficient and the constant (also known as the intercept). For example, the value in cell B3 is obtained by this formula:

```
=E$2+$D$2*A2
```

That is, add the intercept to the product of the coefficient and the prior period's actual. Copy and paste this formula down through B27 to get the remaining forecasts.

But two issues remain. One is that you want to work with a stationary baseline; otherwise, if the baseline is trended, you could easily get some spurious results. The other is that you don't know how far back to look for your predictor value: One period? Two? Three or more? In Figure 16-11, you look one period back. But to answer those two questions for other data sets, you need to look at a couple of charts.

You can download from the publisher's website an Excel workbook named *Correlograms.xlsm* that contains VBA code that will analyze your baseline and tell you first whether it's stationary, and second how far back you should go in your baseline to create your LINEST formula (or, equivalently, your TREND formula).



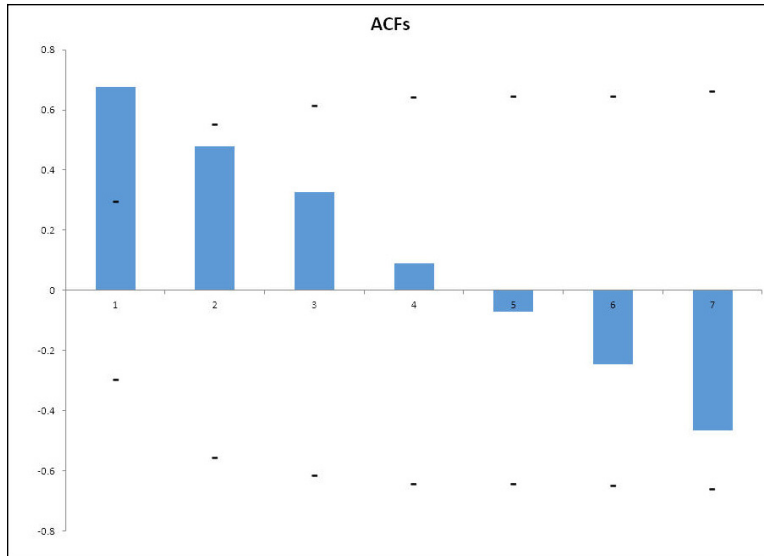
TIP

I have left the VBA code in *Correlograms.xlsm* unprotected, so if you want you can see how the autocorrelation functions and partial autocorrelation functions are calculated.

When you open the workbook, named *Correlograms.xlsm*, you get a new menu item, *Correlograms*, in the Add-ins tab. Add a worksheet to the *Correlograms.xlsm* workbook and put your baseline there. Then click *Correlograms* on the Add-ins tab. A dialog box named *Box-Jenkins Model Identification* appears, with a reference edit box labeled *Input Range for Time Series* where you can enter the range that your baseline occupies. When you click OK in the dialog box, a new workbook opens with two charts: an ACF chart and a PACF chart.

ACF stands for autocorrelation function, and PACF stands for partial autocorrelation function. Don't worry about the terms; you don't need them. What you do need is to look at the charts. Figure 16-12 shows the ACF chart for the data in Figure 16-11.

ACF charts — one type of *correlogram* — show the autocorrelations between a series of observations and other observations from the same data set that are one period back, two periods back, three periods back, and so on. In a stationary data series, you expect to see all these correlations at or near zero. A series with trend, whether up or down, has autocorrelations that are large at the shortest lags and drop gradually to zero and below, as they do in Figure 16-12. They may even head back up, depending both on the length of your time series and on the number of lags you request Correlograms.xlsm to display.



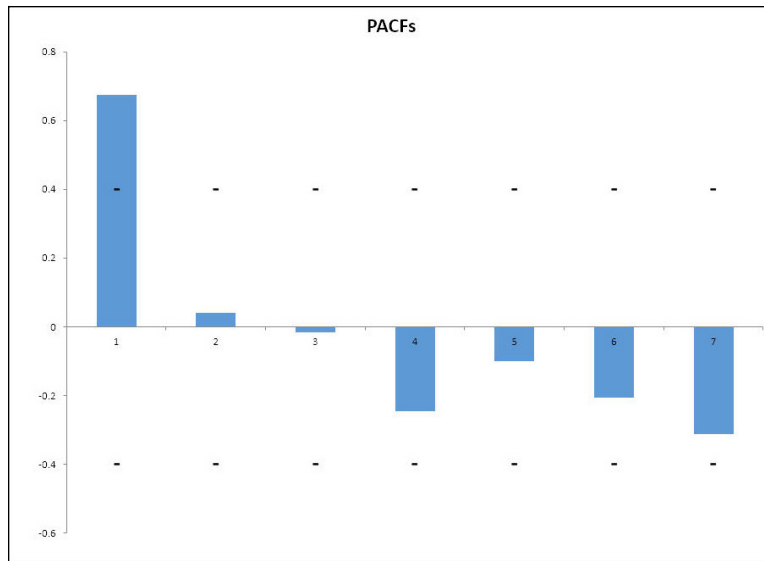
**FIGURE 16-12:**  
The height of the column represents the size of the autocorrelation.

The dashed lines in Figures 16-12 and 16-13 (which are curved in an ACF chart and straight in a PACF chart) show the statistical significance limit: You can consider an ACF or a PACF that extends beyond that limit to be statistically significant — that is, a reliable finding.

If you create an ACF chart that shows this pattern, in which the size of the correlations declines gradually, then you do not have a stationary baseline and you should detrend it by taking first differences and, if necessary, second differences (it's rare that you need to go beyond first differences; see Chapter 14 for more information).

Figure 16-13 shows the PACF chart. This is a good guide to how *far* back you need to go in creating your autoregression analysis. In this case, the first PACF spikes above the significance limit (the lines at  $\pm 0.4$ ) but none of the others do. This indicates that you need go back only one time period to develop your forecast equation: That is, the LINEST formula would involve A2:A25 forecasting A3:A26, just as in Figure 16-11.





**FIGURE 16-13:**  
A stationary series usually has only a single spike in the PACFs.

However, if the second PACF were to spike, you would go back two periods and use  $A_2:A_{24}$  to forecast  $A_4:A_{26}$  (two periods back) as well as (perhaps)  $A_3:A_{25}$  to forecast  $A_4:A_{26}$  (one period back).

## Regressing one trend onto another

If the phrase “causation and correlation” sounds familiar, it’s probably because at some point you’ve read about the difference between the two. Suppose that some sociologist studies the relationship in 100 communities between the number of books in the public libraries and the number of residents of those communities who go on to careers in string theory.

The sociologist finds a strong and positive relationship between those two variables — number of books and number of string theorists. Does this mean that if a community increases the number of books in a community’s libraries, the number of the community’s high school graduates who go on to college and graduate school will also increase?

Of course not. There’s no causation here, just correlation. Probably there’s another variable (or cluster of variables) behind the sociologist’s finding. Communities with a higher per-capita income spend more on community resources such as libraries. And families who live in those communities have more personal resources available to support their children’s higher education.

But if you had some way of increasing the per-capita income of a community, you might well see an increase in both the number of books in the libraries and the number of students going on to higher education. That's the gold standard for deciding that causation exists: Apply a treatment to one randomly constituted group, withhold it from another, and look for a difference between the two groups.

The same effect can occur when you use regression to investigate the relationship between a variable that you want to forecast and some other variable that might be related to it. Suppose you have sales results that have been trending up for the last ten years. You notice that, over the same period of time, your company has been offering more and more product lines.

It's possible to conclude that you can increase sales by increasing the number of product lines, particularly if there's some real perceived differentiation among the products. But it ain't necessarily so. For one thing, both a company's breadth of product line and its sales tend to grow over time — the product line, because companies have to keep up with changing technologies; the sales, because, in the long run, companies either grow or die.

The issue here is that both sales and breadth of product line change in response to time, not necessarily to one another. A third variable, the passage of time, is at work here, just as per-capita income works on both a community's number of books and its number of post-graduate students.

One solution is to detrend both series first, and then see if you can get a useful regression equation from the detrended series. There are various ways of detrending (such as changing a variable to a rate like per-capita income), but a convenient one is to take first differences (see Chapter 17).

## IN THIS CHAPTER

Understanding when to remove the trend from a baseline

Keeping your baseline from jumping around

Putting the pieces of a baseline back together

# Chapter 17

# Managing Trends

Sometimes you'll decide that you'd prefer to use a single-variable forecast method — for example, one of the two single-variable methods that this book discusses: moving averages and simple exponential smoothing. If you're going to do that — and there are some pretty good reasons to go that direction, including improved accuracy and the presence of seasonality in the baseline — you should first check to see if the baseline has a trend. As several other chapters discuss, a trend is the tendency in a baseline to move up or down (not usually both) over time. Both moving averages and simple exponential smoothing behave better in baselines that don't have a trend.

One good way to remove a trend from a baseline is called *differencing*. If you use differencing, you subtract one value in the baseline from a subsequent value. Doing that subtraction has some consequences for the values you use to forecast from. The decision to use differencing isn't a slam-dunk, though — some trade-offs are involved.

If you use differencing, you apply your forecast method to the differences. After you've got your forecast, you still have to put it back into the original baseline's scale. This is called *integrating*, and this chapter shows you how to do it.

# Knowing Why You May Want to Remove the Trend from a Baseline

You can remove trend from a baseline in several ways, and this chapter takes a look at them in the “Getting a Baseline to Stand Still” section. First, though, you need to know more about why you may want to remove a trend from a baseline, and something about ways to diagnose whether a trend is really present in a baseline.

## Understanding why trend is a problem

Some forecasters have argued that removing the trend from a baseline before doing any forecasting was almost always best. They cited three reasons:

- » The regression approach often *takes advantage* of trend in a baseline. But because forecast equations using regression were more difficult to calculate than other methods, you needed a pretty good reason to use regression instead of another, simpler approach.
- » Regression makes a number of assumptions about the data that other approaches don't make. So you have to test those assumptions on your data to decide whether you can even *use* regression with a clear conscience.
- » There's always autocorrelation, often substantial, in a trended series. For various reasons, discussed in Chapters 2, 4, 14, and 16, autocorrelation can present a problem. There's some autocorrelation in a baseline from which the trend has been removed, but usually less — often much less — than in a baseline that has trend.

The first objection — mathematical complexity — has gone away in the intervening years. If you have a PC and Excel, you can perform a regression analysis in a minute or so — less if you type fast, more if you don't.

The second objection is thornier. There are several assumptions that should be tested before using the results of a regression analysis. I won't go into them in detail here, but just so you have a general idea:

- » The average of the errors (also known as *residuals*) in the predicted variable at each value of the predictor is zero. For example, suppose you're predicting sales revenues using sales date as the predictor. If you took the difference between each actual revenue value on any given sales date and its associated forecast value (that is, the error in the forecast) and averaged those residuals, you'd get an average value of zero, or very close to zero.

- » The errors are normally distributed. That is, if you graphed the frequency of ranges of error values, you'd see a normal distribution or "bell curve."
- » The variability of the errors is the same for each value of the predictor variable. If your predictor variable is time period, this assumption is not an issue: You have one observation for each value of the predictor, so there's no variability. But suppose you're using a predictor, such as unit price, that might have multiple forecast values such as unit sales for each value of unit price. Then if you charted the forecast errors against unit price, the variance (*not* the range) of those errors on the vertical axis should be about the same for each unit price.
- » The residuals are not autocorrelated. That is, if you calculated the correlation between, say, the residuals for periods 1 through 49 and the residuals for periods 2 through 50, the correlation would not be significantly different from zero.

Formal descriptions of how to test each of these assumptions is beyond the scope of this book. You can get an idea of whether some assumption has been violated by graphing errors against time periods. Excel's Data Analysis add-in does this for you if you ask it to (it's just a matter of filling a check box — see Chapter 11 for more information).

As to autocorrelation, if you're concerned about its magnitude, keep in mind that a baseline from which the trend has been removed will tend to have less autocorrelation than the original baseline.

What does all of this have to do with getting a baseline to stand still? Well, suppose you can't get your data to behave so you're confident that you've met all the assumptions made by the regression method. In that case, you're probably better off using one of the other methods, among them moving averages and exponential smoothing, that don't make such restrictive assumptions about the baseline. And if you're going to do that, you should remove the trend from the forecast variable's baseline. That means you'll transform it from one that's headed up or down over time to one that's roughly horizontal over time. And that's easier than you may suppose. A good place to start is in an upcoming section, "Subtracting one value from the next value."

## Diagnosing a trend

Chapter 4 goes into some detail about diagnosing whether a trend exists in a baseline. Sometimes you can tell easily by eyeballing a chart of the baseline, but sometimes it's close, whether a trend is there or not. In that case, you want to use a numeric test to help you decide.

Chapter 4 describes a test of the correlation between the sales period and the sales revenues. The test helps you decide whether or not the data is trended. I describe another, quicker, simpler test for trend here, called the Sign Test. Although it's not as sensitive a test as the one described in Chapter 4, it's still a good guide to whether a baseline has a trend.

Figure 17-1 shows an example of the Sign Test. In Figure 17-1, the figures shown in cells D24:D30 are calculated with these formulas:

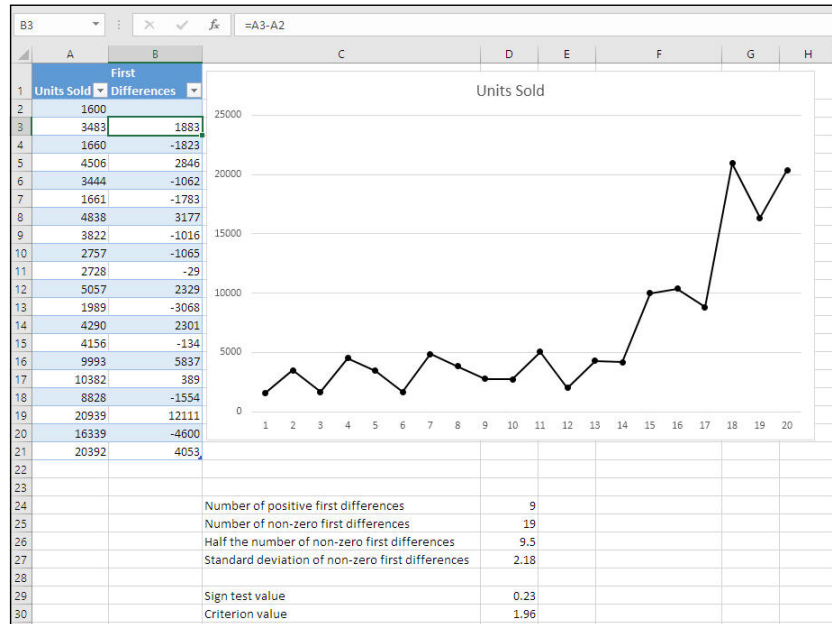
» D24 contains the number of positive first differences in column B. It's calculated with this array formula:

```
=SUM( IF ( B3 : B21 > 0 , 1 , 0 ) )
```

You could also use something like this formula, which does not require that you array-enter it:

```
=COUNTIF ( B3 : B21 , ">0" )
```

In words, the formula looks at B3:B21, and if a value in that range is positive, then the function counts that value — otherwise, it ignores the value. Then it takes the values in the range all together to give you a count of the number of positive figures in B3:B21.



**FIGURE 17-1:**  
An example  
of a Sign Test  
for trend  
in a baseline.

- » D25 contains the number of nonzero first differences. First differences that equal zero can come about if your baseline contains two consecutive values that equal one another. You can get that figure with this array formula:

```
=SUM( IF ( B3 : B21 <> 0 , 1 , 0 ) )
```

This is another instance of an array formula. This time the formula is counting the number of nonzero first differences in B3:B21, not simply the number of positive values. An alternative is:

```
=COUNTIF ( B3 : B21 , " <> 0 " )
```

- » D26 contains half the number of nonzero first differences:

```
=D25/2
```

- » D27 contains the standard deviation of the number of nonzero first differences. For the purposes of the Sign Test, this is:

```
=SQRT(D25/4)
```

- » D29 contains the Sign Test value itself, using Excel's ABS function to return the absolute value:

```
=ABS(D24-D26)/D27
```

- » D30 contains the critical value that is compared to the Sign Test value. If the absolute value of the Sign Test is greater than the value in D30, you decide that the baseline has trend. Otherwise, you decide it doesn't. The absolute value of a number is always positive. The absolute value of 31.2 is 31.2; the absolute value of -31.2 is 31.2. Cell D30 contains:

```
=NORM.S.INV(0.975)
```

where 0.975 is  $1 - 0.025$ . I decide I can live with the probability of making a bad decision about the presence of trend in the baseline 5 percent of the time. I divide 5 percent (0.05) by 2 and get 0.025, and then subtract that from 1.0.

In this case, the Sign Test's absolute value, 0.23, is less than the critical value of 1.96. So, despite the fact that the baseline is rising during four of the final six points, the Sign Test regards the baseline as a stationary one.

## Getting a Baseline to Stand Still

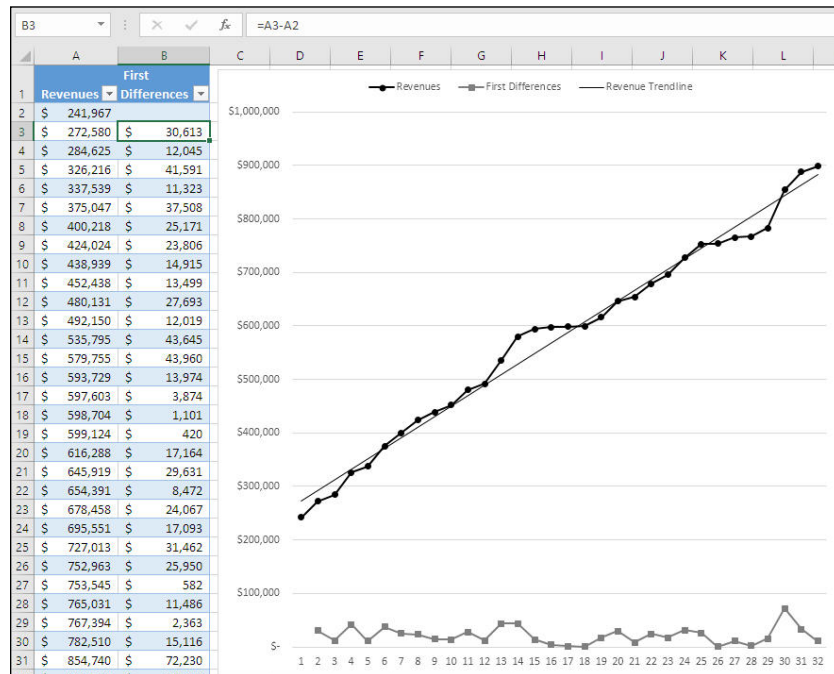
When you have a baseline that has a trend, whether up or down, you can decide to leave the trend in place. If you make that decision, you've in effect decided to use regression as your forecast method.

And that may very well be the right choice. It's often, but certainly not always, the case that a regression-based forecast is more accurate than alternatives such as exponential smoothing or moving averages.

## Subtracting one value from the next value

There is a method called *first differencing* that usually takes the trend out of a baseline. To use it, you just subtract one value from the next value in the baseline — that's why it's called *differencing*. The results of the subtractions are called *differences* or *first differences*. Occasionally, you may need to difference the differences — that's called *second differencing*. You may find you have to do that if your original series grew exponentially — that is, it looks not like a straight-line trend, growing at a roughly fixed amount from period to period. Instead, it's a curve that gets steeper the farther you get into the baseline.

Figure 17-2 shows an example of a baseline with a linear, straight-line trend and the baseline's first differences. The first differences form what's often called a *stationary series* of values.



**FIGURE 17-2:**  
Just by subtracting one value from the next, you can take the trend out of the baseline.

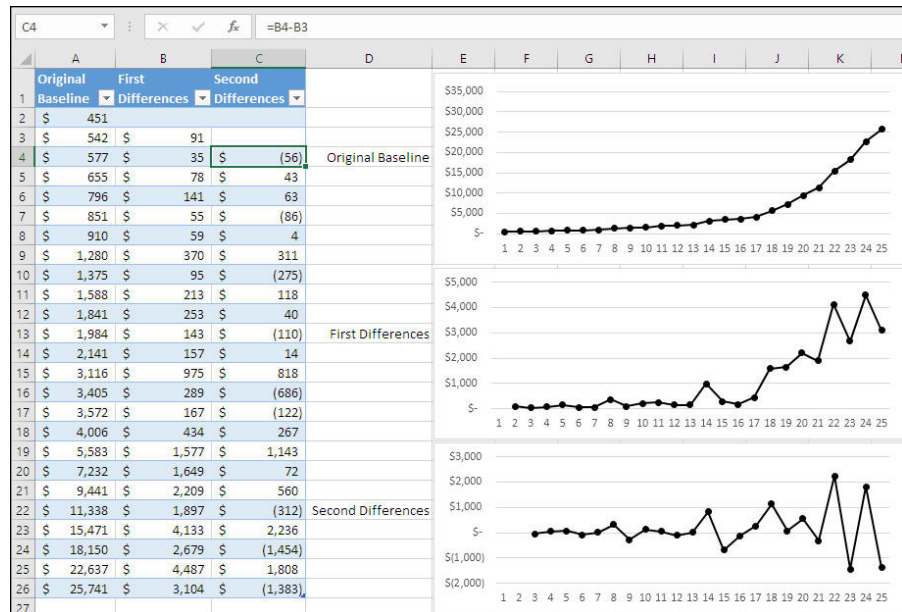


Notice in Figure 17-2 that the original baseline trends up very strongly. But the first differences form a horizontal, stationary series.

I know what you're thinking. "If I'm supposed to make a forecast from the transformed, stationary baseline, that forecast is going to be under \$100,000, but the next actual value will be close to \$1,000,000. What's the deal?"

Good question. Toward the end of this chapter, there's a section called "And All the King's Men: Putting a Baseline Together Again." That section shows you how to undo the first differencing — often called *integrating* the differenced values.

You don't see it all the time, but you do occasionally, so have a look at Figure 17-3 for an example of a baseline that requires second differencing.



**FIGURE 17-3:** When the original baseline isn't linear, you may need to take second differences.

In Figure 17-3, the curve in the baseline is so pronounced that the first differences also have a curve to them. (A gentler original curve would usually produce first differences that looked more like a straight-line trend.) But the second differences are finally horizontal. These are the values you'd use to make your forecast. Then you'd reintegrate that forecast twice so you could compare it to the original baseline's scale. (See the section later in this chapter, "And All the King's Men: Putting a Baseline Together Again.")

## Dividing one value by another

Another approach to stabilizing a baseline that has trend is to divide, rather than subtract, one value by the previous value. The result forms a series of what are called *link relatives*. Figure 17-4 shows the link relatives for the baseline used in Figures 17-1 and 17-2.



	A	B	C
1	Original Baseline	Link Relatives	
2	\$ 451		
3	\$ 542	\$ 1.20	
4	\$ 577	\$ 1.06	
5	\$ 655	\$ 1.14	
6	\$ 796	\$ 1.22	
7	\$ 851	\$ 1.07	
8	\$ 910	\$ 1.07	
9	\$ 1,280	\$ 1.41	
10	\$ 1,375	\$ 1.07	
11	\$ 1,588	\$ 1.15	
12	\$ 1,841	\$ 1.16	
13	\$ 1,984	\$ 1.08	
14	\$ 2,141	\$ 1.08	
15	\$ 3,116	\$ 1.46	
16	\$ 3,405	\$ 1.09	
17	\$ 3,572	\$ 1.05	
18	\$ 4,006	\$ 1.12	
19	\$ 5,583	\$ 1.39	
20	\$ 7,232	\$ 1.30	
21	\$ 9,441	\$ 1.31	
22	\$ 11,338	\$ 1.20	
23	\$ 15,471	\$ 1.36	
24	\$ 18,150	\$ 1.17	

**FIGURE 17-4:**  
Link relatives usually form a stationary series.

To get Excel to calculate link relatives, follow these steps:

1. Using the layout in Figure 17-4, select cell B3.
2. Type `=A3/A2` and press Enter.
3. Select cell B3 and choose Edit ⇨ Copy.
4. Select the range B4:B26 and choose Edit ⇨ Paste.



TIP

There are quicker ways to extend the formula down the column. In Step 3, after you have selected cell B3, move your mouse pointer over the black square in B3's lower-right corner (that's called the *fill handle*). Click it, but don't release the mouse button, and drag down into cell B26. Or, just double-click the fill handle.

Now you have the link relatives. You should have already charted the original baseline to determine whether it's already stationary (if it is, you don't need to be doing this — the only point to calculating either first differences or link relatives is if the original baseline had an up or a down trend). To get the link relatives into that chart, do the following:

1. **Select the entire range of link relatives, including its column header.**

In Figure 17-4, that's the range B1:B26.

2. **Go to the Ribbon's Home tab and click Copy in the Clipboard group.**



TIP

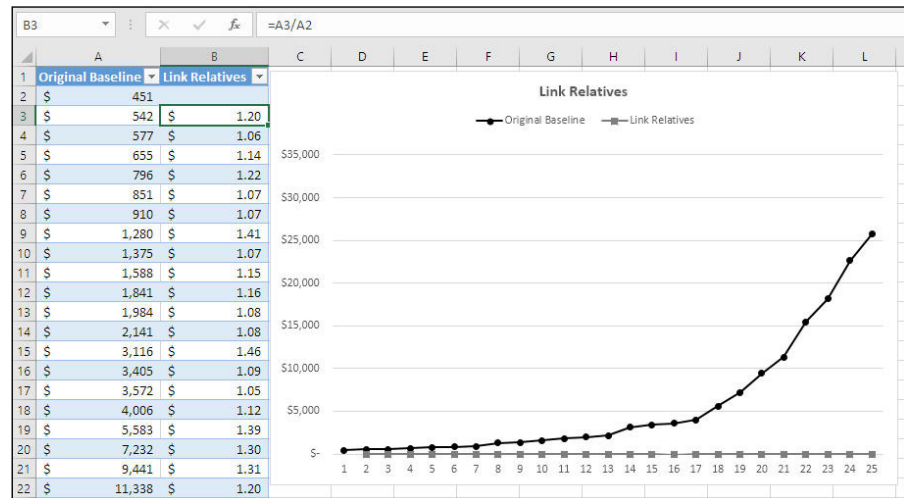
Yes, that means that you've included a blank cell, B2, in the copied range. By selecting the blank cell, you arrange to line up the two data series properly on the chart. By including the Link Relatives column header, you give the series that name, which will appear in the chart legend.

3. **Select the chart by clicking in it.**

4. **Choose Paste from the Home tab's Clipboard group.**

The chart now appears as in Figure 17-5.

**FIGURE 17-5:**  
The chart's primary vertical axis obscures the link relatives at the bottom of the plot area.



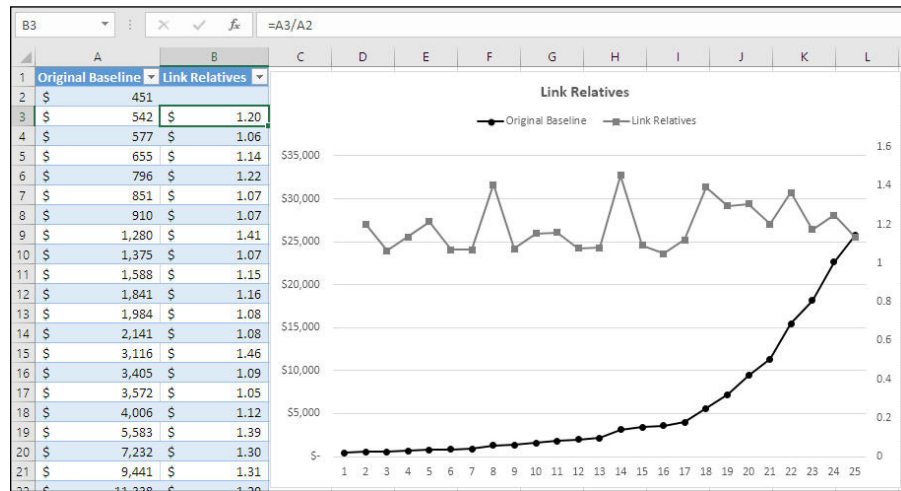
Because they're ratios of one baseline value to the preceding baseline value, link relatives usually range from 0 to 2. So if you want to chart them along with the original baseline values, as in Figure 17-5, you may need to use two different vertical axes. If you use one axis only, the values of the link relatives often get mashed down into the horizontal axis.

To fix the problem, take these steps:

1. **Right-click the Link Relatives series on the chart (Figure 17-6) and choose Format Data Series from the shortcut menu.**

You might find it easier to go to the chart's Format group on the Ribbon and select the Link Relatives series from the drop-down in the Current Selection group. Then click Format Selection.

2. **Under Series Options in the Format Data Series pane, select the Secondary Axis option button.**



**FIGURE 17-6:** Now you can see the link relative much more clearly.

## Getting rates

Another way to make a trended baseline stationary is to change it to a rate of some sort. For example, outside the area of sales forecasting, you may want to forecast the number of traffic accidents that will occur during the next quarter. If you were to divide the number of accidents in your baseline by the number of licensed drivers during that quarter, it's quite likely that the ratio would form a stationary baseline.

This is a sort of per-capita ratio, and people tend to understand per-capita ratios more easily than first differences or link ratios.

But if you adopt this approach to detrending a baseline, you'll need to forecast both the accident rate and the number of licensed drivers, and then apply the forecast rate to forecast of drivers. Forecast errors in each baseline will then tend to compound one another.

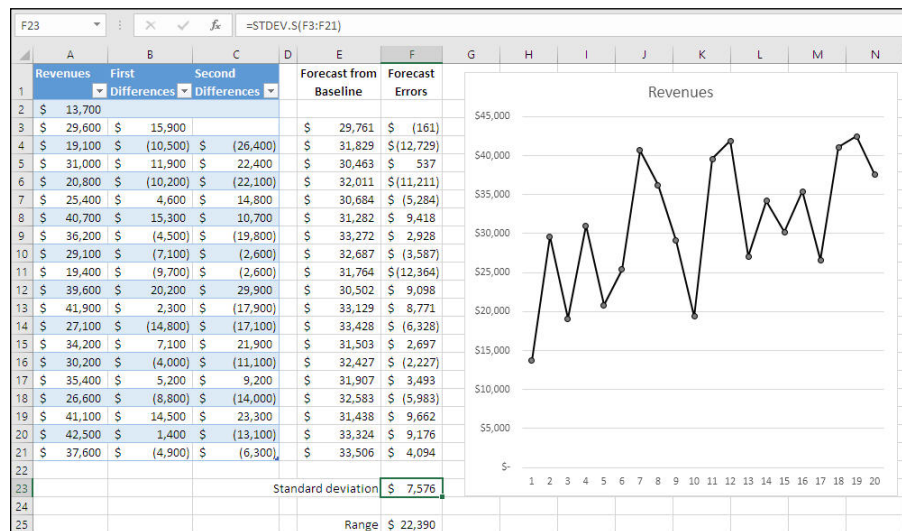
## The downside of differencing

Take another look at Figures 17-1 and 17-3. You'll notice that you lose one value from the beginning of the baseline of first differences, as well as that of link relatives. The reason is that, in each case, you're converting two values into one.

This is yet another reason to put some effort into getting as lengthy a baseline as you can. When your baseline contains only ten values, first differencing (or the use of link relatives) causes you to lose 10 percent of your data. If you've managed to collect 50 values without spending half a career doing so, first differencing loses only 2 percent — not usually a major loss.

It's not just the raw number of values that are lost, it's also the power of the statistical tests that you may apply to determine whether you've really removed the trend (see Chapter 4). The smaller the number of data points involved in the test, the less sensitive the test.

But wait, there's more. Whenever you difference a baseline, the amount of forecast error increases. Have a look at Figure 17-7.



**FIGURE 17-7:** Note that the baseline of revenues has a clear upward trend.

In Figure 17-7:

- » Column A contains the original baseline. The chart shows that the baseline is trended, so you would consider detrending it.
- » Column B contains the first differences of the original baseline. For example, cell B4 contains the difference between A4 and A3, B5 contains the difference between A5 and A4, and so on.
- » Column C contains the second differences of the original baseline. For example, cell C5 contains the difference between B5 and B4, C6 contains the difference between B6 and B5, and so on.

With that data in columns A, B, and C, you can do some forecasting. Excel's TREND worksheet function comes in handy here. TREND does the following:

- » It takes some actual values that you already know, such as a baseline of sales revenues.
- » It takes some associated predictor values that you also already know.
- » It calculates a regression equation between the predictor values and the actual values.
- » It applies that equation to the predictors and returns the forecasts calculated by the regression equation.

In Figure 17-7, column E contains the TREND function. Its predicted variable consists of the revenues in cells A3:A21. Its predictor variable consists of the revenues in cells A2:A20. In words, the TREND function forecasts the 2nd through 20th values in the baseline, using the 1st through 19th values as predictors.



WARNING

This is classic autoregression, where you forecast based on the variable itself — you use a prior value to forecast a later value. Formally, you should detrend the series first. (This first example is not detrended, to make a point about increasing the errors when you do so.)

You enter the TREND function as an array formula. In Figure 17-7, you select E3:E21 and type the following equation:

```
=TREND(A3:A21, A2:A20)
```



REMEMBER

You array-enter a formula not by pressing Enter, but by pressing Ctrl+Shift+Enter.

Several functions in Excel are concerned with regression and require that you array-enter them:

- » **LINEST:** This function returns the regression coefficients as well as several statistics that help you evaluate how accurate the regression equation is likely to prove.
- » **TREND:** See the discussion earlier in this section.
- » **MINVERSE:** This function returns the inverse of a matrix. It's the multiple-value version of a regular inverse — for example, 2/7 is the inverse of 7/2.
- » **MMULT:** This function returns the product of two matrixes.

You also use array formulas outside the context of regression. Array formulas, properly designed, can return results that are both elegant and impossible to achieve any other way.

For example, in cell E3 of Figure 17-7, you see \$29,761. This is the result of plugging the first actual value, \$13,700, into the regression equation. Cell E4, \$31,829, is the result of plugging the second actual, \$29,600, into the equation.

Column F in Figure 17-7 is the difference between the forecasts in column E and the actuals in column A. These are the forecast errors, or *residuals*. The less the variability — the spread — in the forecast errors, the more precise your forecast.

Cell F23 contains the standard deviation of the forecast errors in the range F3:F21. The standard deviation is one measure of the amount of spread in a set of numbers. If you're not familiar with standard deviations, you can instead look at the range in cell F25: the difference between the maximum and minimum values in a set of numbers.

Now do the same thing with the first differences of the baseline, shown in column B of Figure 17-7. The result appears in Figure 17-8.

Figure 17-8 adds two columns to Figure 17-7. Column H uses the TREND function just as column E does, except that it forecasts using the first differences. It develops an equation that uses a prior first difference to forecast a subsequent first difference.

Don't get lost in the funhouse here. Right now, the main point is not the meaning of a forecast first difference; the point is the amount of extra error you induce when you difference a baseline.

1	Revenues	First Differences	Second Differences	Forecast from Baseline	Forecast Errors	Forecast from First Differences	Forecast Errors
2	\$ 13,700						
3	\$ 29,600	\$ 15,900		\$ 29,761	\$ (161)		
4	\$ 19,100	\$ (10,500)	\$ (26,400)	\$ 31,829	\$ (12,729)	\$ (5,770)	\$ (4,730)
5	\$ 31,000	\$ 11,900	\$ 22,400	\$ 30,463	\$ 537	\$ 5,703	\$ 6,197
6	\$ 20,800	\$ (10,200)	\$ (22,100)	\$ 32,011	\$ (11,211)	\$ (4,032)	\$ (6,168)
7	\$ 25,400	\$ 4,600	\$ 14,800	\$ 30,684	\$ (5,284)	\$ 5,573	\$ (973)
8	\$ 40,700	\$ 15,300	\$ 10,700	\$ 31,282	\$ 9,418	\$ (859)	\$ 16,159
9	\$ 36,200	\$ (4,500)	\$ (19,800)	\$ 33,272	\$ 2,928	\$ (5,510)	\$ 1,010
10	\$ 29,100	\$ (7,100)	\$ (2,600)	\$ 32,687	\$ (3,587)	\$ 3,095	\$ (10,195)
11	\$ 19,400	\$ (9,700)	\$ (2,600)	\$ 31,764	\$ (12,364)	\$ 4,225	\$ (13,925)
12	\$ 39,600	\$ 20,200	\$ 29,900	\$ 30,502	\$ 9,098	\$ 5,355	\$ 14,845
13	\$ 41,900	\$ 2,300	\$ (17,900)	\$ 33,129	\$ 8,771	\$ (7,639)	\$ 9,939
14	\$ 27,100	\$ (14,800)	\$ (17,100)	\$ 33,428	\$ (6,328)	\$ 140	\$ (14,940)
15	\$ 34,200	\$ 7,100	\$ 21,900	\$ 31,503	\$ 2,697	\$ 7,572	\$ (472)
16	\$ 30,200	\$ (4,000)	\$ (11,100)	\$ 32,427	\$ (2,227)	\$ (1,946)	\$ (2,054)
17	\$ 35,400	\$ 5,200	\$ 9,200	\$ 31,907	\$ 3,493	\$ 2,878	\$ 2,322
18	\$ 26,600	\$ (8,800)	\$ (14,000)	\$ 32,583	\$ (5,983)	\$ (1,120)	\$ (7,680)
19	\$ 41,100	\$ 14,500	\$ 23,300	\$ 31,438	\$ 9,662	\$ 4,964	\$ 9,536
20	\$ 42,500	\$ 1,400	\$ (13,100)	\$ 33,324	\$ 9,176	\$ (5,162)	\$ 6,562
21	\$ 37,600	\$ (4,900)	\$ (6,300)	\$ 33,506	\$ 4,094	\$ 531	\$ (5,431)
22							
23				Standard deviation	\$ 7,576		\$ 9,186
24							
25				Range	\$ 22,390		\$ 31,100

**FIGURE 17-8:** The variability or spread in the error values of the forecast first differences increases.

Column I shows the difference between the forecasted and the actual first differences. Look at cells I23 (the standard deviation of the forecast errors) and I25 (their range). Both are larger than for the forecast errors based on the original baseline, shown in F23 and F25. The act of differencing the baseline has increased the variability in the forecast errors.

Finally, Figure 17-9 shows that the same effect continues with second differencing. The forecasts, using the TREND function on the values in column C, are in column K, and the forecast errors — the differences between the second differences and the forecast second differences — are in column L. Notice that in cells L23 and L25, both the standard deviation and the range are larger than those for the first differences.

Why is this important? Because the greater the error variation, the less precise your forecast. And that's the trade-off involved in detrending a series:

- » You can use one of the simpler approaches to forecasting — in particular, simple exponential smoothing — by detrending the series. But then you're inducing more variability into the forecast errors, making your forecast less precise.
- » You can leave the trend in place and use a regression approach, typically using time period as the predictor variable. But then you're making more-restrictive assumptions about your data (particularly the independence and randomness of the residuals), which ideally you should test, and those tests can be time consuming and not necessarily easy to interpret.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Revenues	First Differences	Second Differences		Forecast from Baseline	Forecast Errors		Forecast from First Differences	Forecast Errors		Forecast from Second Differences	Forecast Errors
2	\$ 13,700											
3	\$ 29,600	\$ 15,900			\$ 29,761	\$ (161)						
4	\$ 19,100	\$ (10,500)	\$ (26,400)		\$ 31,829	\$ (12,729)		\$ (5,770)	\$ (4,730)			
5	\$ 31,000	\$ 11,900	\$ 22,400		\$ 30,463	\$ 537		\$ 5,703	\$ 6,197	\$ 14,718	\$ 7,682	
6	\$ 20,800	\$ (10,200)	\$ (22,100)		\$ 32,011	\$ (11,211)		\$ (4,032)	\$ (6,168)	\$ (12,767)	\$ (9,333)	
7	\$ 25,400	\$ 4,600	\$ 14,800		\$ 30,684	\$ (5,284)		\$ 5,573	\$ (973)	\$ 12,296	\$ 2,504	
8	\$ 40,700	\$ 15,300	\$ 10,700		\$ 31,282	\$ 9,418		\$ (859)	\$ 16,159	\$ (8,487)	\$ 19,187	
9	\$ 36,200	\$ (4,500)	\$ (19,800)		\$ 33,272	\$ 2,928		\$ (5,510)	\$ 1,010	\$ (6,177)	\$ (13,623)	
10	\$ 29,100	\$ (7,100)	\$ (2,600)		\$ 32,687	\$ (3,587)		\$ 3,095	\$ (10,195)	\$ 11,001	\$ (13,601)	
11	\$ 19,400	\$ (9,700)	\$ (2,600)		\$ 31,764	\$ (12,364)		\$ 4,225	\$ (13,925)	\$ 1,313	\$ (3,913)	
12	\$ 39,600	\$ 20,200	\$ 29,900		\$ 30,502	\$ 9,098		\$ 5,355	\$ 14,845	\$ 1,313	\$ 28,587	
13	\$ 41,900	\$ 2,300	\$ (17,900)		\$ 33,129	\$ 8,771		\$ (7,639)	\$ 9,939	\$ (16,991)	\$ (909)	
14	\$ 27,100	\$ (14,800)	\$ (17,100)		\$ 33,428	\$ (6,328)		\$ 140	\$ (14,940)	\$ 9,931	\$ (27,031)	
15	\$ 34,200	\$ 7,100	\$ 21,900		\$ 31,503	\$ 2,697		\$ 7,572	\$ (472)	\$ 9,480	\$ 12,420	
16	\$ 30,200	\$ (4,000)	\$ (11,100)		\$ 32,427	\$ (2,227)		\$ (1,946)	\$ (2,054)	\$ (12,486)	\$ 1,386	
17	\$ 35,400	\$ 5,200	\$ 9,200		\$ 31,907	\$ 3,493		\$ 2,878	\$ 2,322	\$ 6,101	\$ 3,099	
18	\$ 26,600	\$ (8,800)	\$ (14,000)		\$ 32,583	\$ (5,983)		\$ (1,120)	\$ (7,680)	\$ (5,333)	\$ (8,667)	
19	\$ 41,100	\$ 14,500	\$ 23,300		\$ 31,438	\$ 9,662		\$ 4,964	\$ 9,536	\$ 7,734	\$ 15,566	
20	\$ 42,500	\$ 1,400	\$ (13,100)		\$ 33,324	\$ 9,176		\$ (5,162)	\$ 6,562	\$ (13,274)	\$ 174	
21	\$ 37,600	\$ (4,900)	\$ (6,300)		\$ 33,506	\$ 4,094		\$ 531	\$ (5,431)	\$ 7,227	\$ (13,527)	
22												
23				Standard deviation	\$ 7,576			\$ 9,186				\$ 13,967
24												
25				Range	\$ 22,390			\$ 31,100				\$ 55,617

**FIGURE 17-9:**  
The farther you get from the original baseline, the greater the error variation.

My own tendency is to do it both ways and see which one seems to provide the more accurate forecasts. (Also, the diagnostic tests involved in checking the regression assumptions go pretty quickly, but that's because I've been doing this for so many years.)

## And All the King's Men: Putting a Baseline Together Again

If you decide to difference a baseline, to make it stationary before you use it to make a forecast, you have a little more work to do before you're ready to announce that forecast.

As you've seen, when you difference a baseline you usually wind up with much smaller values in the first differences than are in the original baseline. In Figure 17-9, for example, the first differences in column B are much smaller than the original baseline values in column A.

And, as this chapter explains, the idea behind differencing is to remove the trend from a baseline prior to forecasting with one of the single-variable methods like moving averages and exponential smoothing. They're more effective when you deploy them on a stationary baseline.



This is so primarily because it's a good idea to find the best smoothing constant using an optimization utility such as Excel's Solver. (Solver is an add-in that comes with Excel.) But if you do so with a baseline that is still trended, Solver often tells you that the best smoothing constant is 1.0, which results in naïve forecasting — each forecast is precisely equal to its preceding actual.

So, when you make your forecast, it's in first-difference units, not the original baseline's units (see Figure 17-10).

1	Time Period	Revenue	First Differences	Forecast Differences
2	1	\$ 674		
3	2	\$ 625	\$ (49)	#N/A
4	3	\$ 586	\$ (39)	\$ (49)
5	4	\$ 514	\$ (72)	\$ (46)
6	5	\$ 473	\$ (41)	\$ (54)
7	6	\$ 428	\$ (45)	\$ (50)
8	7	\$ 343	\$ (85)	\$ (48)
9	8	\$ 236	\$ (107)	\$ (59)
10	9	\$ 278	\$ 42	\$ (74)
11	10	\$ 231	\$ (47)	\$ (39)
12	11	\$ 215	\$ (16)	\$ (41)
13	12	\$ 209	\$ (6)	\$ (34)
14	13	\$ 269	\$ 60	\$ (25)
15	14	\$ 299	\$ 30	\$ 0
16	15	\$ 382	\$ 83	\$ 9
17	16	\$ 461	\$ 79	\$ 31
18	17	\$ 442	\$ (19)	\$ 46
19	18	\$ 375	\$ (67)	\$ 26
20	19	\$ 370	\$ (5)	\$ (2)
21	20	\$ 396	\$ 26	\$ (3)
22	21	\$ 439	\$ 43	\$ 6
23	22	\$ 419	\$ (20)	\$ 17
24	23 forecast			\$ 6
25				
26	Smoothing constant		0.3	
27	Damping factor		0.7	

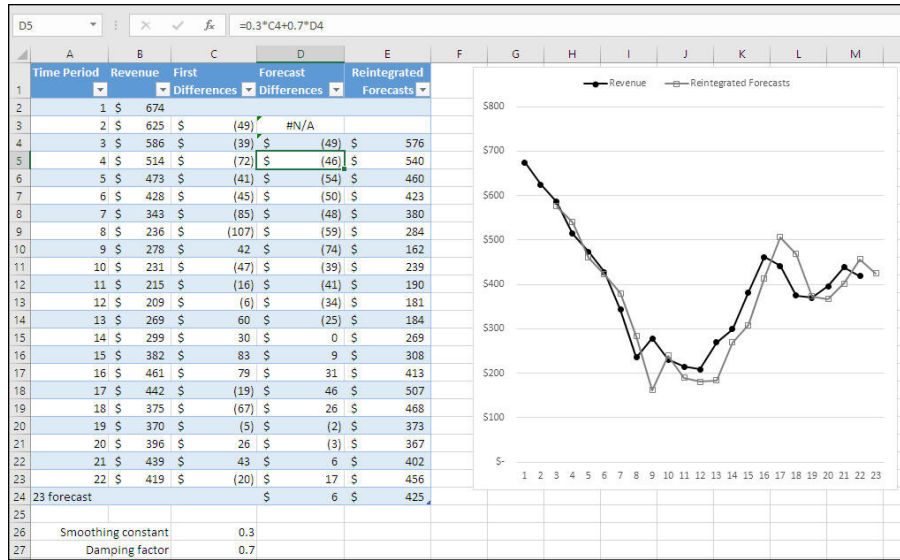
**FIGURE 17-10:** Exponential smoothing forecasts a value of \$6 in cell D24. You still need to reintegrate the forecasts with the original baseline.

You can't report a forecast for period 23 of \$6. People won't understand when you say you're forecasting a first difference. They'll go back to the guy with the pointy hat with the stars and moons on it.

You need to integrate your forecast first difference with the baseline first. This is shown, along with a chart of the baseline and the reintegrated forecast, in Figure 17-11.

You got the first differences by subtracting a baseline value from the value that follows it. Now that you've made your forecast of the first differences (including the forecast of \$6 for the 23rd period), you need to add them back to the baseline. This is how you get back to the original scale. In this example, the original scale in the baseline runs from \$209 to \$674, and the scale of the forecast differences runs from a negative \$107 to a positive \$83.

**FIGURE 17-11:**  
Differencing  
and then  
integrating the  
predicted  
differences  
can lead to a  
good forecast  
of a trended  
data series.



Combining the baseline and the forecast differences is easy. Using the example in Figure 17-11, enter the following formula in cell E5:

=B4+D5

Here you're taking the value in the baseline, which is the basis for the differencing, and adding back into it the difference that you've forecast for the current period. The result is that you return to the original scale of the baseline.

This chapter has already mentioned the two principal downsides to differencing: loss of at least one value (first differencing a 20 value baseline, for example, results in 19 first differences), and greater variability in the forecast errors. When you then use exponential smoothing to make your forecast, you lose yet another value. As Chapter 15 shows, you can't make a forecast from the value in the baseline that precedes the first one, because it doesn't exist. So your first forecast value is always not available — or in Excel worksheet terms, #N/A.

The upshot is that you're going to wind up with two fewer forecast values than values in your baseline: one lost to first differencing, one lost to exponential smoothing. The loss can be even greater with moving averages, depending on how many baseline values go into each moving average.



Recognizing seasonal patterns in your exponential smoothing

Calculating your first forecast

Modifying the formulas to finish your forecast

## Chapter 18

# Same Time Last Year: Forecasting Seasonal Sales

**Y**ears have their seasons, and seasons make their mark on sales — particularly in the retail sector. If you're going to forecast sales in a business segment that has seasonal peaks and valleys, you're going to need a topo map. And you can get that map by accounting for seasons in your smoothing. It's just a step more complicated than regular old exponential smoothing. Your seasonal forecast is based not only on the most recent observation, but also on the last time this season came through on the calendar.

So, as you start to get a ways into the baseline, there are two components to a seasonal forecast, and one is the *level* component. This component is analogous to the previous actual baseline value used in exponential smoothing, described in Chapter 15. The current baseline level needs some adjustment before you apply the smoothing constant, in order to separate out the seasonal effect so that you can focus on the level.

The second component is *seasonal*. The idea is that, every year, the seasons have similar effects on sales. In preparing to make a forecast, you need to quantify those effects. You assign a number to each season — that is, you might've found

that, over time, you experience a \$2,000,000 falloff during spring and a \$4,000,000 boost during winter. You can use that information to improve the accuracy of your forecasts. And not incidentally, you can use it to extend the number of future time periods you can forecast into.

Of course, there is often a third component, *trend*. Trend is often forecast specifically, like the level and the seasonal components. But to keep things straightforward in this book, I discuss differencing as a means of dealing with trend, rather than forecasting it by means of another smoothing constant. Models, often termed *Holt models*, are fine alternatives to differencing as a means of dealing with trended baselines.

Finally, in this chapter, I show you the difference between forecasts that are within the baseline periods and forecasts that extend past the end of the baseline — the ones you’re really interested in. And I show you how to use a utility that I wrote to unburden you: It relieves you of having to put the equations for seasonal baselines into a worksheet.

## Doing Simple Seasonal Exponential Smoothing

Simple seasonal exponential smoothing builds on concepts you use in the examples of exponential smoothing you see earlier in this book. The difference between the two is that you often recognize a seasonal pattern, like colder weather during winter, or increases in retail sales during the winter holidays.

Then you need to consider bringing a more complicated approach to bear on the forecasting problem. By “consider,” I mean that you need to try both approaches and see which one works best for you.



TIP

I do suggest, though, that if you haven’t looked through this book’s chapters on simple exponential smoothing, you do so before you get into this chapter. It will be a lot easier to follow if you’ve already thought about things like smoothing constants and diminishing influences.

### Relating a season to its ancestors

Think back to how exponential smoothing works. It uses a formula like this one to base the next forecast in part on the prior actual and in part on the prior forecast:

$$\text{New Forecast} = (0.3 \times \text{Prior Actual}) + (0.7 \times \text{Prior Forecast})$$

This amounts to a weighted average of two prior figures — the actual and the forecast. This particular formula gives a good bit more of the weight to the forecast than to the actual. You have to experiment around some with a particular baseline to get the right smoothing constant (that's the 0.3 in the formula) and the right damping factor (that's the 0.7 in the formula).

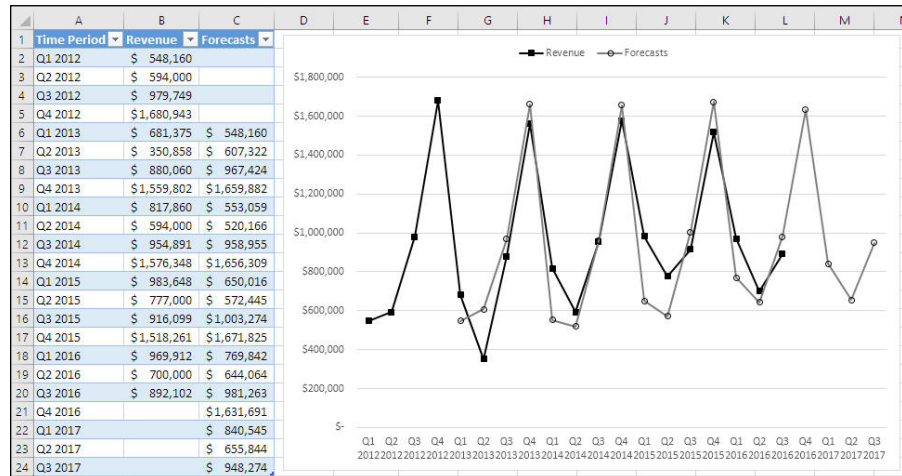
The idea here is that one time period in the baseline is going to be closely related to the following time period. If today's high temperature were 70°F, you'd have to show me an approaching cold front to convince me that tomorrow's high will be 50°F. Without additional, contradictory information, I'd prefer to bet on 70°F. Yesterday tends to forecast today, and today tends to forecast tomorrow.

But let's shift to months. A given month's average temperature is much more closely related to the historical average for that month than it is to the prior month's average temperature. If May's average daily high were 70°F, I'd still lean toward 70°F for June, but before I put any money down on it I'd want to know what *last* June's average daily high was.

So here's what I'm going to do: Instead of using just one smoothing constant, I'll use two. Instead of using only one constant in conjunction with the immediately prior baseline value, I'll use one for the prior value (smoothing May to help forecast June), and one for the season that's one year back from this one (smoothing last June to help forecast next June).

Figure 18-1 shows a seasonal sales baseline, and the associated forecasts, in practice.

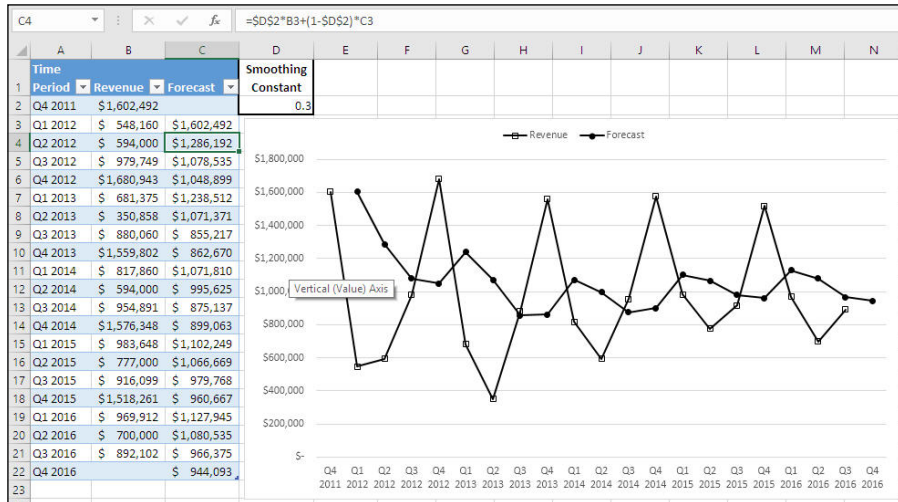
**FIGURE 18-1:**  
The seasonal forecasts cannot start until one sequence of baseline seasons has passed.



Notice in Figure 18-1 how the sales invariably head up during the third quarter of each year, and spike during the fourth quarter. Then the bottom falls out during the first and second quarters. The figure also shows the forecasts, which have captured the seasonal pattern in a smoothing equation, making the forecasts that much more accurate.

What if I used simple exponential smoothing, the sort covered in Chapter 15? Figure 18-2 gives some of the bad news.

**FIGURE 18-2:**  
The forecasts smooth through the signal in the baseline.



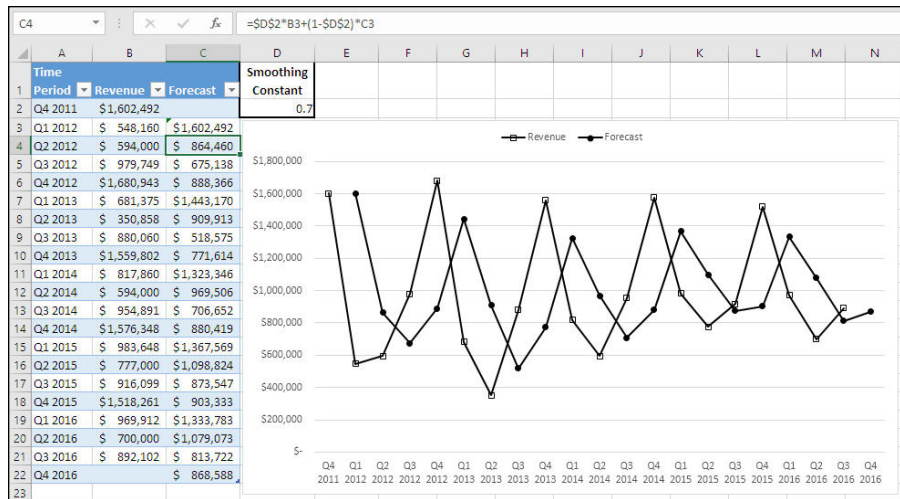
Here, the smoothing constant is 0.3, and the forecasts are relatively insensitive to fluctuations in the actuals from the baseline. The forecasts do nod in passing to the peaks and valleys in the baseline, but it's a dismissive sort of nod.

What if I boosted the smoothing constant so that the forecasts track the actuals more than they smooth them? That situation is shown in Figure 18-3, where the smoothing constant is 0.7.

In Figure 18-3, the peaks and valleys are represented more clearly — but they lag one period behind their actual occurrence. Compare Figure 18-3 and its tardy forecasts with Figure 18-1 and its on-time forecasts. The forecasts in Figure 18-1 can show up on time because they pay attention to what happened last year. And showing up is 85 percent of life.

Figure 18-4 shows how I combine the components to get a forecast value. Don't worry, the source of the components and what they mean become clear as I walk through developing the seasonal forecast.





**FIGURE 18-3:**  
The forecasts are late to reflect the changes in the baseline.

**FIGURE 18-4:**  
The seasonal effects are above (positive values) and below (negative values) the current overall level of the baseline.

Year	Season	t	Observations	Forecast for this period	Forecast Level	Seasonal Index	Forecast for next period	Alpha	Delta
2012	1	1	\$ 548,160			\$ (402,553)		0.1	0.3
	2	2	\$ 594,000			\$ (356,713)			
	3	3	\$ 979,749			\$ 29,036			
	4	4	\$ 1,680,943		\$ 950,713	\$ 730,230	\$ 548,160		=F5+G2
2013	1	5	\$ 681,375	\$ 548,160	\$ 964,035	\$ (366,585)	\$ 607,322		

The formula in cell F5 of Figure 18-4 gives the level of the baseline as of Q4 2012. The formula is:

=AVERAGE(D2:D5)

At the outset of the smoothing process, this is our best estimate of the current level of the baseline. It's just the average of the four quarterly revenue results for 2012. It is analogous to using the first observation as the first forecast in simple exponential smoothing.

From examining the formula in cell H5 of Figure 18-4:

=F5+G2

you can see that the forecast for Quarter 1 of 2013 is the sum of two quantities:

- » The forecast level of the baseline for Q1 2013 as of Q4 2012 (see cell F5)
- » The effect of being in Quarter 1 as of 2012 (see cell G2)

Every forecast in column E and column H in this chapter's figures is the sum of the forecast level of the baseline and the effect of the season from the preceding year. A good sanity check compares the seasonal smoothing forecasts in Figure 18-1 with the ordinary smoothing forecasts in Figures 18-2 and 18-3. Clearly, you're better off if you can estimate the seasonal effect *before* it takes place. This is what is happening in Figure 18-4, which combines the level that's attributable to a season with the general level of the baseline to get the current season's forecast *before* the next instance of the season takes place.

That's the reason, in Figure 18-4, that I put the forecast for the *next* period in column H, and for the *current* period in column E. Doing so helps me remember that I can assemble the forecast for a given period at the end of the preceding period. Notice, for example, that cell H5 has the forecast for the next period, that cell E6 has the forecast for the current period, and that they both equal \$548,160.

## Using the smoothing constants

This section walks you through a demonstration of how you get forecasts that are seasonally smoothed, based on the discussion of Figure 18-4. There's some math in it, but nothing more complicated than arithmetic. It's a little tedious — or it would be if you had to smooth each forecast in the way I show you here. But there's a workbook named *Seasonal Smoothing.xlsx*, with code in it that you can download from the publisher's website. When you run that code, you'll be prompted for some information, like where your baseline data is and the values you want to use for the smoothing constants. When you click OK, you'll get the seasonally smoothed forecasts.

If you decide you'd like to do the forecasting on the worksheet, using the formulas I set out in this section, there's just the upfront work to do. When you're a couple of periods into the forecast region, the process turns into a simple copy and paste.

So getting a seasonally smoothed forecast is not as onerous as it's going to seem by the time you've reached the end of this section. When you've gotten there, you'll have a better understanding of what's going on, and you can steam ahead.

Seasonal smoothing uses not one but two smoothing constants: one for the current level of the baseline (*alpha*) and one for the current seasonal effect (*delta*).



TECHNICAL  
STUFF

Actually, there are sometimes three smoothing constants: one for the current level, one for the current season, and one for the slope in the baseline, and a smoothing model that uses all three is called a *Holt-Winters model*. To keep from getting things tangled up, I assume that either there's no slope in the baseline, or I'm working with a baseline that I've already differenced and, thus, made

stationary. If you read other books on forecasting, you may see the level constant referred to as *alpha* and the seasonal constant referred to as *delta*. (There's not a lot of standardization in the Greek names for the constants, but it appears that the literature on smoothing tends to prefer *alpha* for the level constant and *delta* for the seasonal constant.)

Figure 18–5 shows an example of the smoothing equation for the forecast's *level* component.

**FIGURE 18-5:**  
The first estimate of the baseline's level is the average of the revenues in the first year, in cell F5.

	C	D	E	F	G	H	I	J	K
	t	Observations	Forecast for this period	Forecast Level	Seasonal Index	Forecast for next period	Alpha	Delta	
1							0.1		
2	1	\$ 548,160			\$ (402,553)			0.3	
3	2	\$ 594,000			\$ (356,713)				
4	3	\$ 979,749			\$ 29,036				
5	4	\$1,680,943		\$ 950,713	\$ 730,230	\$ 548,160			
6	5	\$ 681,375	\$ 548,160	\$ 964,035	\$ (366,585)	\$ 607,322			
7	6	\$ 350,858	\$ 607,322						
8					=Alpha*(D6-G2)+(1-Alpha)*F5				

Bear in mind that, to forecast revenue in Q2 2013, you want to do your data gathering and apply your formulas during Q1 2013. So, you're working with information that's available to you by the end of Q1 2013. More generally, you can forecast revenue for your next period as soon as the data for the current period is available. And with seasonal smoothing you can legitimately forecast out as far as one full turn of the seasons beyond the most recent actual result in your baseline.

### Estimating the season's effect

Work forward from cell G2. That cell contains the initial estimate of the effect of being in the first quarter. It's nothing more than the actual revenue in Q1 2012 less the average for all quarters in 2012. In terms of formulas, that's the following:

$$=D2-\$F\$5$$

where D2 contains the actual revenue for Q1 2012, and \$F\$5 returns the quarterly average of the first year's revenues, as discussed earlier in this chapter. (I've made the reference to cell F5 an absolute reference, so that the formula can be copied from cell G2 into G3:G5 without losing the reference to the quarterly average for 2012.)

So, by subtracting the average quarterly revenue in the first year from the first quarter's actual revenue, you arrive at an estimate of the effect of being in the first quarter. That effect may be positive — if the first quarter is good for sales, its effect is a positive number — or negative — if the first quarter is bad for sales, it's a negative number. Whichever, that seasonal effect is also frequently called the *seasonal index*.

## Starting the smoothing process

By now you have an estimate of the baseline level as of Q4 2012 in cell F5 and of the effect of Q1 in cell G2. You add them together in cell H5 to get your first forecast of quarterly revenue, for Q1 2013, in cell H5.

To show that forecast in the quarter and year that it's intended to estimate, you use this formula in cell E6:

```
=H5
```

to show the forecast both at the end of the quarter in which it's made and at the start of the quarter that it's meant to forecast.

In cell F6 you now move into the smoothing portion of the forecast. The formula in F6 is:

```
=A1pha*(D6-G2)+(1-A1pha)*F5
```

Does that look at all familiar? It may if you've looked at Chapter 15 recently. It's a smoothing equation, and it has two parts:

- » In the first part, you multiply the smoothing constant (designated as *alpha*) times the difference between D6 and G2: the actual revenue in Q1 2013 less the actual effect of being in Q1. Just as with the (non-seasonal) exponential smoothing discussed in Chapter 15, you're multiplying the smoothing constant times an actual. But because you can't observe the current level of the baseline directly, you get it by subtracting the seasonal index in G2 from the observed revenue in D6.
- » In the second part, you're multiplying the damping factor times the previous forecast — again, just as in simple exponential smoothing. With seasonal smoothing, though, the second part of the formula in cell F6 forecasts only the baseline level, not the seasonal effect.



REMEMBER

The damping factor is 1 minus alpha, the smoothing constant.

Finally, you get your forecast for Q2 2013 in cell H6 by adding the estimate of the seasonal effect to the estimate of the level of the baseline:

$$=F6+G3$$

or \$607,322.

## Reviewing the process

Here's a review of what you've done: In cell G2 you have an estimate of the effect on revenue of being in Q1. Your best estimate of that, in Q1 2013, is what happened in Q1 2012. Then, \$ (402,553) was the effect of being in Q1.



REMEMBER

There are several ways to display a negative amount in currency. The default U.S. English method is to enclose the amount in parentheses, and as a default kinda guy, I'm using that method. So \$ (402,553) means a negative number of dollars.

Put another way, in Q1 2012 the company made \$402,553 *less* than the average quarter for all of 2012. Relative to each quarter's average, Q1 was a bad quarter to be in, to the tune of \$402,553. In terms of cell addresses:

$$\text{\$ (402,553)} = D2 - \text{\$F\$5}$$

and in terms of the dollar values in those cells:

$$\text{\$ (402,553)} = \$548,160 - \$950,713$$

so that's your seasonal effect, at least as of Q1 2013.

Also, you've smoothed an actual and a forecast to get a measure of the current level of the baseline:

- » You used the level smoothing constant, alpha, of 0.1 on the difference between revenue for Q1 2013 and the seasonal effect. You take revenue to be the combined effect of the current level of the baseline and the seasonal effect.

$$\text{Revenue} = \text{Level} + \text{Season}$$

So the difference between the revenue and the seasonal effect is a measure of the level of the baseline:

$$\text{Revenue} - \text{Season} = \text{Level}$$

» You used the damping factor of  $(1.0 - 0.1)$ , or  $0.9$ , on the previous forecast of the baseline level, just as in Chapter 15. In this case, that's the average of all four quarters in 2012: your best estimate of the level of the baseline at this point. You add those two — the actual and the forecast, each multiplied by the appropriate constant — together. In terms of cell addresses:

$$\text{\$964,035} = \text{Alpha} * (\text{D6} - \text{G2}) + (1 - \text{Alpha}) * \text{F5}$$

and in terms of the dollar values in those cells:

$$\text{\$964,035} = 0.1 * (\text{\$681,375} - \text{\$ (402,553)}) + 0.9 * \text{\$950,713}$$

Finally, you total the smoothed estimate of the baseline level and the estimate of the seasonal effect, to get the forecast for Q2 2013. In terms of cell addresses:

$$\text{\$607,322} = \text{F6} + \text{G3}$$

and in terms of the dollar values in those cells:

$$\text{\$607,322} = \text{\$964,035} + \text{\$ (356,713)}$$

## Getting Farther into the Baseline

Chapter 15 is where this book first really digs into exponential smoothing. There you can find that the first forecast made by exponential smoothing is just the first value in the baseline. Nothing is available earlier than the first value on which to base a forecast, so exponential smoothing uses the first value instead.

### Calculating the first forecast

If you prefer to look here at the effect of there being no prior value, instead of going to Chapter 15 for it, Figure 18-6 shows an example of what I'm talking about.

Notice first cell C5 in Figure 18-6. It contains the classic formula for exponential smoothing:

$$=\text{Alpha} * \text{B4} + (1 - \text{Alpha}) * \text{C4}$$

	A	B	C	D	E	F
2	Month	Sales Revenue	Forecast		Alpha	0.3
3	Jan-15	844976				
4	Feb-15	457693	\$844,976	=B3		
5	Mar-15	297644	\$728,791	=Alpha*B4+(1-Alpha)*C4		
6	Apr-15	239536	\$599,447			
7	May-15	765852	\$491,474			
8	Jun-15	692499	\$573,787			
9	Jul-15	322603	\$609,401			
10	Aug-15	306883	\$523,361			
11	Sep-15	832918	\$458,418			
12	Oct-15	588323	\$570,768			
13	Nov-15	828963	\$576,034			
14	Dec-15	510947	\$651,913			
15	Jan-16	269966	\$609,623			
16	Feb-16	265274	\$507,726			
17	Mar-16	575070	\$434,990			
18	Apr-16	562041	\$477,014			
19	May-16	282268	\$502,522			
20	Jun-16	557839	\$436,446			
21	Jul-16	325820	\$472,864			
22	Aug-16	844698	\$428,751			
23	Sep-16	635165	\$553,535			

**FIGURE 18-6:** The first forecast in smoothing is usually the first value in the baseline.

That is:

- » The smoothing constant alpha, 0.3 in this case, times the prior actual value in cell B4
- » The damping factor (1.0 minus the smoothing constant alpha) times the prior forecast in cell C4
- » The sum of the two multiples

Suppose you copied and pasted that formula one cell higher in the worksheet — which would amount to starting to use the formula one month earlier, in February 2015 rather than March 2015. It would be:

```
=Alpha*B3+(1-Alpha)*C3
```

From the point of view of the worksheet structure, the problem is that there's no value in cell C3 to multiply by the damping factor. That is, this portion of the formula:

```
0.7*C3
```

is 0 because C3 is empty.

From the point of view of the logic of forecasting by way of exponential smoothing, the problem is that you've gone back too far in the baseline. Exponential smoothing needs a weighted average of a prior actual and a prior forecast. But as

of the first time period in the baseline, no forecast is available: There is no value prior to the first time period on which to base a forecast.

So, the first forecast, found in cell C4 in Figure 18-6, is not a smoothed forecast. It takes the first value in the baseline to be the best estimate of the first forecast, and in this case its formula is:

=B3

This is called *initializing* the forecast.



TECHNICAL  
STUFF

A method called *backcasting* forecasts backward into the baseline's first time period, and — if the backcaster wants — even farther back. I don't get into it here, but at some later point you may want to know that the technique exists, or at least recognize the term. There are methods that act as though the baseline starts earlier than it really does. Although they can be useful, none of those methods is completely satisfactory.

What does all this have to do with *seasonal* smoothing? The problem at the start of the series is extended, because you have not just one but several forecasts to initialize. Those are the initial estimates of the seasonal effects (also termed *seasonal indexes*).

The preceding section mentions briefly how this is done, but I want to put it in the context of the worksheet and the formulas to give you a better feel for what's going on.

To initialize the forecasts for the level of the baseline and the seasons, you start by getting the average of the quarterly revenues for the earliest full year, in cell F5 of Figure 18-7. You use this value in two places:

- » Later, when you start smoothing to get new estimates of the level of the baseline
- » Now, when you initialize the seasonal effects based on the first year

**FIGURE 18-7:**  
The seasonal effects show how the individual revenue figures vary around the *average* revenue figure.

	A	B	C	D	E	F	G	H	I	J	K
1	Year	Season	t	Observations	Forecast for this period	Forecast Level	Seasonal Index				
2	2012	1	1	\$ 548,160			\$ (402,553)	=D2-\$F\$5		Alpha	0.1
3		2	2	\$ 594,000			\$ (356,713)	=D3-\$F\$5		Delta	0.3
4		3	3	\$ 979,749			\$ 29,036	=D4-\$F\$5			
5		4	4	\$ 1,680,943		\$ 950,713	\$ 730,230	=D5-\$F\$5			
6											
7							\$0	=SUM(G2:G5)			



To get your initial estimates of the seasonal effects, you calculate in cells G2:G5 the results of subtracting the average quarterly revenue for 2012 from the actual revenue during each quarter. No offense to the numbers, but these are called *deviations*.

Notice in cell G7 that the sum of these deviations is zero, and this is always true (and is easily proven). So each deviation isolates from the current level of the baseline the effect of its season above or below that level:

- » The effect of Q1 on revenue is \$ (402,553). The level of the baseline, \$950,713 plus the seasonal effect of \$ (402,553), is \$548,160, the revenue for Q1 2012.
- » The effect of Q2 on revenue is \$ (356,713). The level of the baseline, \$950,713 plus the seasonal effect of \$ (356,713), is \$594,000, the revenue for Q2 2012.
- » The effect of Q3 on revenue is \$29,035. The level of the baseline, \$950,713 plus the seasonal effect of \$29,036 is \$979,749, the revenue for Q3 2012.
- » The effect of Q4 on revenue is \$730,230. The level of the baseline, \$950,713 plus \$730,230 is \$1,680,943, the revenue for Q4 2012.

Because these four deviations sum to zero, adding them to the level of the baseline has no effect on that level for the full year — just for the revenue during each quarter.



TECHNICAL  
STUFF

I've started the calculations using Q1 through Q4 of 2012, even though Q4 2011 could have been available. I did this to keep the notion of seasons in a year straightforward: It's easier to think of a year as starting in Q1 than to think of it as starting in Q4. But the quarterly designations are just labels and have no effect on the forecasting process. Besides, your company might well have a fiscal year that begins on October 1. So there's no special technical reason to start the forecasting in Q1, and there can be good reasons to start it in some other quarter (or monthly period, or bimonthly period).

## Smoothing through the baseline level

After the initial estimates are made as described in the preceding section, you're ready to start getting three actual forecasts:

- » The smoothed forecast of the revenue itself
- » The smoothed forecast of the level of the baseline
- » The smoothed forecast of the level of the season

Figure 18–8 shows how the smoothed forecast of the baseline level is calculated. Column G shows the contents of the formulas that are used in column F. Column G

shows that the formulas follow the pattern you've established for smoothing. In cell F7 and as shown in G7, the formula uses:

- » A smoothing constant, often called *alpha* and found in cell L1, times an actual value — the actual revenue for Q2 2013 less the actual seasonal effect of Q2 2012
- » A damping factor times the prior estimate of the same variable — that is, 1.0 minus the smoothing factor, times the forecast of the baseline level as of Q1 2013

	A	B	C	D	E	F	G	H	I	J	K	L
	Year	Season	t	Observations	Forecast for this period	Forecast Level		Seasonal Index	Forecast for next period		Alpha	Delta
1	2012	1	1	\$ 548,160							0.1	0.3
2		2	2	\$ 594,000				\$ (402,553)				
3		3	3	\$ 979,749				\$ (356,713)				
4		4	4	\$ 1,680,943		\$ 950,713	=AVERAGE(D2:D5)	\$ 29,036	\$ 730,230	\$ 548,160		
5	2013	1	5	\$ 681,375	\$ 548,160	\$ 964,035	=Alpha*(D6-H2)+(1-Alpha)*F5	\$ (366,585)	\$ 607,322			
6		2	6	\$ 350,858	\$ 607,322	\$ 938,388	=Alpha*(D7-H3)+(1-Alpha)*F6	\$ (425,958)	\$ 967,424			
7		3	7	\$ 880,060	\$ 967,424	\$ 929,652	=Alpha*(D8-H4)+(1-Alpha)*F7	\$ 5,448	\$ 1,659,882			
8		4	8	\$ 1,559,802	\$ 1,659,882	\$ 919,644	=Alpha*(D9-H5)+(1-Alpha)*F8	\$ 703,208	\$ 553,059			
9	2014	1	9	\$ 817,860	\$ 553,059	\$ 946,124	=Alpha*(D10-H6)+(1-Alpha)*F9	\$ (295,089)	\$ 520,166			
10		2	10	\$ 594,000	\$ 520,166	\$ 953,507	=Alpha*(D11-H7)+(1-Alpha)*F10	\$ (406,023)	\$ 958,955			
11		3	11	\$ 954,891	\$ 958,955	\$ 953,101	=Alpha*(D12-H8)+(1-Alpha)*F11	\$ 4,350	\$ 1,656,309			
12		4	12	\$ 1,576,348	\$ 1,656,309	\$ 945,105	=Alpha*(D13-H9)+(1-Alpha)*F12	\$ 681,619	\$ 650,016			
13	2015	1	13	\$ 983,648	\$ 650,016	\$ 978,468	=Alpha*(D14-H10)+(1-Alpha)*F13	\$ (205,008)	\$ 572,445			
14		2	14	\$ 777,000	\$ 572,445	\$ 998,923	=Alpha*(D15-H11)+(1-Alpha)*F14	\$ (350,793)	\$ 1,003,274			
15		3	15	\$ 916,099	\$ 1,003,274	\$ 990,206	=Alpha*(D16-H12)+(1-Alpha)*F15	\$ (19,187)	\$ 1,671,825			
16		4	16	\$ 1,518,261	\$ 1,671,825	\$ 974,850	=Alpha*(D17-H13)+(1-Alpha)*F16	\$ 640,157	\$ 769,842			
17	2016	1	17	\$ 969,912	\$ 769,842	\$ 994,857	=Alpha*(D18-H14)+(1-Alpha)*F17	\$ (150,989)	\$ 644,064			
18		2	18	\$ 700,000	\$ 644,064	\$ 1,000,450	=Alpha*(D19-H15)+(1-Alpha)*F18	\$ (335,690)	\$ 981,263			
19		3	19	\$ 892,102	\$ 981,263	\$ 991,534	=Alpha*(D20-H16)+(1-Alpha)*F19	\$ (43,260)	\$ 1,631,691			

**FIGURE 18-8:** Wait for a year's worth of actuals in the baseline before starting seasonal smoothing.

These two values are summed to get the smoothed forecast of the baseline level made at Q2 2013 and a portion of the forecast of the actual Q3 2013 revenue.

## 'Tis the seasonal component

Figure 18-9 shows how the smoothed forecast of the seasonal effect is calculated.

The same approach is used: the smoothing constant times an actual, plus the damping factor times the prior estimate. Starting in cell G6 and continuing down through G20, the same pattern is in place, differing only in the location of the prior actuals and the prior estimates. In cell G6, the formula uses:

- » A smoothing constant, often called *delta* and found in cell L2, times an actual value — the actual revenue for Q1 2013 less the forecast baseline level for Q1 2013. Note that this leaves the seasonal effect, based on actuals.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Year	Season	t	Observations	Forecast for this period	Forecast Level	Seasonal Index		Forecast for next period		Alpha	0.1
2	2012	1	1	\$ 548,160			\$ (402,553)	=D2-\$F55			Delta	0.3
3		2	2	\$ 594,000			\$ (356,713)	=D3-\$F55				
4		3	3	\$ 979,749			\$ 29,036	=D4-\$F55				
5		4	4	\$ 1,680,943		\$ 950,713	\$ 730,230	=D5-\$F55	\$ 548,160			
6	2013	1	5	\$ 681,375	\$ 548,160	\$ 964,035	\$ (366,585)	=Delta*(D6-F6)+(1-Delta)*G2	\$ 607,322			
7		2	6	\$ 350,858	\$ 607,322	\$ 938,388	\$ (425,958)	=Delta*(D7-F7)+(1-Delta)*G3	\$ 967,424			
8		3	7	\$ 880,060	\$ 967,424	\$ 929,652	\$ 5,448	=Delta*(D8-F8)+(1-Delta)*G4	\$ 1,659,882			
9		4	8	\$ 1,559,802	\$ 1,659,882	\$ 919,644	\$ 703,208	=Delta*(D9-F9)+(1-Delta)*G5	\$ 553,059			
10	2014	1	9	\$ 817,860	\$ 553,059	\$ 946,124	\$ (295,089)	=Delta*(D10-F10)+(1-Delta)*G6	\$ 520,166			
11		2	10	\$ 594,000	\$ 520,166	\$ 953,507	\$ (406,023)	=Delta*(D11-F11)+(1-Delta)*G7	\$ 958,955			
12		3	11	\$ 954,891	\$ 958,955	\$ 953,101	\$ 4,350	=Delta*(D12-F12)+(1-Delta)*G8	\$ 1,656,309			
13		4	12	\$ 1,576,348	\$ 1,656,309	\$ 945,105	\$ 681,619	=Delta*(D13-F13)+(1-Delta)*G9	\$ 650,016			
14	2015	1	13	\$ 983,648	\$ 650,016	\$ 978,468	\$ (205,008)	=Delta*(D14-F14)+(1-Delta)*G10	\$ 572,445			
15		2	14	\$ 777,000	\$ 572,445	\$ 998,923	\$ (350,793)	=Delta*(D15-F15)+(1-Delta)*G11	\$ 1,003,274			
16		3	15	\$ 916,099	\$ 1,003,274	\$ 990,206	\$ (19,187)	=Delta*(D16-F16)+(1-Delta)*G12	\$ 1,671,825			
17		4	16	\$ 1,518,261	\$ 1,671,825	\$ 974,850	\$ 640,157	=Delta*(D17-F17)+(1-Delta)*G13	\$ 769,842			
18	2016	1	17	\$ 969,912	\$ 769,842	\$ 994,857	\$ (150,989)	=Delta*(D18-F18)+(1-Delta)*G14	\$ 644,064			
19		2	18	\$ 700,000	\$ 644,064	\$ 1,000,450	\$ (335,690)	=Delta*(D19-F19)+(1-Delta)*G15	\$ 981,263			
20		3	19	\$ 892,102	\$ 981,263	\$ 991,534	\$ (43,260)	=Delta*(D20-F20)+(1-Delta)*G16	\$ 1,631,691			

**FIGURE 18-9:** After the end of the first full year, you switch from deviations to smoothing for seasonal estimates.

» A damping factor times the prior estimate of the same variable — that is, 1.0 minus the smoothing factor, times the estimate of the seasonal effect as of Q1 2012.

Again, these two values are summed to get the smoothed forecast of the seasonal effect for the first quarter, as of Q1 2013. Although the seasonal index in cell G6 is calculated in Q1 2013, it is not actually used in a revenue forecast until Q1 2014 (see cells G10 and H10).

You now have the two components that determine the seasonally smoothed forecasts: a smoothed baseline level and a smoothed seasonal effect. You can add them together to come up with the forecast for the subsequent time period. For example, the formula for the forecast for Q2 2013 is:

=F6+G3

That forecast is calculated in cell I6 and repeated as a cell link in cell E7.

Referring to Figure 18-9, you have four columns with static values: the label that identifies the time period in columns A:C and the actual revenue for each time period in column D. These columns would have to be filled in manually (usually by copy-and-paste methods). They may also get their values from an external data range that points to another data source such as a database, or via a pivot table that summarizes to quarter, month, or some other time period.

The remaining columns in Figure 18-9, columns E through I, contain formulas. These formulas are special for rows 2 through 5, where you're calculating the seasonal effects directly in G2:G5 and the baseline level directly in F5.

The remaining rows, row 6 through row 20, contain formulas that you can copy and paste.

## Finishing the Forecast

When you get to the end of the baseline, a minor modification is needed to complete the forecasts. It's these forecasts, the ones that extend beyond the end of the baseline, that are the real focus of your work. After all, by now you have the actual revenue values for the periods that have already come and gone.

### Modifying the formulas

Figure 18-10 illustrates what's going on at this point.

	A	B	C	D	E	F	G	H	I	J	K
1	Year	Season	t	Observations	Forecast for this period	Forecast Level	Seasonal Index	Forecast for next period		Alpha	0.1
2	2012	1	1	\$ 548,160			\$ (402,553)			Delta	0.3
3		2	2	\$ 594,000			\$ (356,713)				
4		3	3	\$ 979,749			\$ 29,036				
5		4	4	\$1,680,943		\$ 950,713	\$ 730,230	\$ 548,160			
6	2013	1	5	\$ 681,375	\$ 548,160	\$ 964,035	\$ (366,585)	\$ 607,322			
7		2	6	\$ 350,858	\$ 607,322	\$ 938,388	\$ (425,958)	\$ 967,424			
8		3	7	\$ 880,060	\$ 967,424	\$ 929,652	\$ 5,448	\$1,659,882			
9		4	8	\$1,559,802	\$1,659,882	\$ 919,644	\$ 703,208	\$ 553,059			
10	2014	1	9	\$ 817,860	\$ 553,059	\$ 946,124	\$ (295,089)	\$ 520,166			
11		2	10	\$ 594,000	\$ 520,166	\$ 953,507	\$ (406,023)	\$ 958,955			
12		3	11	\$ 954,891	\$ 958,955	\$ 953,101	\$ 4,350	\$1,656,309			
13		4	12	\$1,576,348	\$1,656,309	\$ 945,105	\$ 681,619	\$ 650,016			
14	2015	1	13	\$ 983,648	\$ 650,016	\$ 978,468	\$ (205,008)	\$ 572,445			
15		2	14	\$ 777,000	\$ 572,445	\$ 998,923	\$ (350,793)	\$1,003,274			
16		3	15	\$ 916,099	\$1,003,274	\$ 990,206	\$ (19,187)	\$1,671,825			
17		4	16	\$1,518,261	\$1,671,825	\$ 974,850	\$ 640,157	\$ 769,842			
18	2016	1	17	\$ 969,912	\$ 769,842	\$ 994,857	\$ (150,989)	\$ 644,064			
19		2	18	\$ 700,000	\$ 644,064	\$ 1,000,450	\$ (335,690)	\$ 981,263			
20		3	19	\$ 892,102	\$ 981,263	\$ 991,534	\$ (43,260)	\$1,631,691			
21		4	20		\$1,631,691	\$ 991,534	\$ 150,649	\$ 840,545			
22	2017	1	21		\$ 840,545	\$ 991,534	\$ (403,153)	\$ 655,844			
23		2	22		\$ 655,844	\$ 991,534	\$ (532,443)	\$ 948,274			
24		3	23		\$ 948,274	\$ 991,534	\$ (327,743)	\$1,142,184			
25		4	24		\$1,142,184						

**FIGURE 18-10:** Past the baseline, the estimates of baseline levels become a constant.

The forecasts of the level component that extend beyond the end of the baseline are in the range E21:E25. They're different from earlier forecasts in that they don't incorporate an estimate of the baseline's level from the prior period. The most recent updated level estimate is in cell F20, for Q3 2016.

And this continues to be your best estimate for any forecasts that extend past Q3 2016, when you got your most recent actual. So subsequent forecasts use it as the level estimate. You can see that in the forecasts for Q4 2016 through Q4 2017. Each of them uses the value calculated in cell F20 for the level component of the forecast.

In contrast, you have estimates of the seasonal effects from Q1 2016 through Q4 2016. These are available to you for use in the forecasts of Q1 2017 through Q4 2017. And that's what the formulas in H21:H24 do: They make use of the most recent estimate of the level of the baseline, and add to it the current values of the seasonal effects.

Of course, as the next quarter's actuals become available, you can reestimate the current level of the baseline and update the forecast for Q1 2017, using the new information. You can also get a new season effect estimate and extend the future smoothed forecasts.

## Using the worksheet

If you want to view the intermediate calculations that come into play — in particular, the calculation of the seasonal effects and the baseline level — you can plug a new baseline into the worksheets illustrated in Figure 18-10. You just need to keep a few matters in mind:

- » **If you're using a different time period in your baseline than quarters, you'll want to change the time period labels in columns A and B.** That's not necessary for the actual forecasting, but it'll help keep you straight with which data goes where.
- » **Your forecasts, the values in column E, will start no sooner than the first row after the first year's worth of actuals have completed.**
- » **You may want to adjust the way you calculate the first estimate of the baseline level, shown in Figure 18-10 in cell F5.** There, the formula is the average of cells D2:D5. But if you were using monthly actuals in your baseline, it might be the average of cells D2:D13. And in that case, your forecasts would begin in cells E14 — all due to the fact that there are eight more months in a year than there are quarters.
- » **Be sure that your estimates of the level effect in column F, and the seasonal effects in column G, point back to the correct time period.** For example, in Figure 18-10, the formulas in cells F6 and G6 make use of the value in cell G2. That's the most recent estimate, as of Q1 2013, of the effect of being in Q1, so it's the appropriate estimate to use for a forecast of the next Q1 revenue. If you're using months as your time period, your level and seasonal estimates for, say, January will need to point back to the preceding January's seasonal effect.



TIP

In this chapter, I have named the cells that contain the smoothing constants — it’s convenient to name one of the cells Alpha and the other cell Delta. Names are by default absolute references, so you don’t need to worry about putting dollar signs in your formulas, and it helps if the formulas document themselves — for example:

```
=Delta*(B7-E7)+(1-Delta)*F3
```

## Using the workbook

You can download an Excel workbook from the publisher’s website. It’s named Seasonal Smoothing.xlsx and was mentioned in “Using the smoothing constants” earlier in this chapter. The workbook contains code that will do all the seasonal exponential smoothing calculations for you. It’s much faster than entering all the worksheet formulas, but it’s also less informative if you really want to dig into what’s happening. Its output includes this information:

- » The baseline values
- » The forecast values
- » The values of alpha and delta that you used

All you need is a baseline. Open the Seasonal Smoothing.xlsx file, and then open the workbook with your baseline. Click Smooth in the Ribbon’s Add-ins group to display the Seasonal Exponential Smoothing dialog box. Figure 18-11 shows how it might look.

**FIGURE 18-11:**  
If you use text as a list header, don’t include it in the baseline range.

With the dialog box and your baseline showing, follow these steps:

- 1. In the dialog box, click the reference edit box labeled Baseline Range, and drag through the cells containing your baseline.**

Given the layout in Figure 18-6, that would be B3:B26.

**2. In the Number of Periods in Each Season box, enter the appropriate number.**

For example, if you're treating a quarter as a season and your baseline has one period for each quarter, you would enter the number 1. If you're treating a quarter as a season and your baseline has monthly data, you would enter the number 3.

**3. In the Number of Seasons in Each Year box, enter the appropriate number.**

If you want to treat each quarter as a season, enter **4** in the box — regardless of whether your baseline range shows the data on a monthly basis or a quarterly basis.

**4. Put the values you want to use for alpha and delta in the appropriate boxes, and click OK.**

The results are put in a new worksheet, inserted before the active worksheet (the one where you have the baseline). Figure 18-12 shows the results for the baseline values shown in the range B3:B26 of Figure 18-6, and assumes that the seasons are quarters and that each time period in the baseline represents one quarter.

	A	B	C	D
1	Baseline	Forecast	Smoothing constant for trend: 0.1	Smoothing constant for seasons: 0.3
2	548160			
3	594000			
4	979749			
5	1680943			
6	681375	548160.0		
7	350858	607321.5		
8	880060	967424.2		
9	1559802	1659881.7		
10	817860	553058.8		
11	594000	520165.7		
12	954891	958955.0		
13	1576348	1656309.4		
14	983648	650016.1		
15	777000	572445.1		
16	916099	1003273.9		
17	1518261	1671824.9		
18	969912	769841.6		
19	700000	644063.6		
20	892102	981263.5		
21		1631690.8		
22		840545.1		
23		655843.9		
24		948273.7		

**FIGURE 18-12:** Compare the forecasts from the add-in with the forecasts from the worksheet in Figure 18-10.

Notice that the smoothing constants are shown in row 1. You can run the analysis repeatedly, for different values of alpha and delta. Then calculate the errors in the forecasts and put them through the same analysis discussed in Chapter 15. This will help you determine which combination of alpha and delta gives you the smallest amount of error in your forecasts.



## Excel 2016's new Forecast Sheet

There's a new utility included with Excel 2016, called the Forecast Sheet. You can get at it from the Ribbon's Data tab, in the Forecast group.

I can't recommend that you use this utility. It does create forecasts using a smoothing approach that appears to be similar to the Holt-Winters methods that I mention earlier in this chapter. There's a check box you can select that directs the utility to try to detect trend in your baseline. The results include a chart of the actuals and the forecasts, and a confidence interval around the forecasts. Optional forecast statistics include values for an alpha, a beta, and a gamma constant, terms that are sometimes used to refer to level, trend, and seasonality smoothing constants.

All that sounds great. But a slightly closer look reveals defects. I have benchmarked the Forecast Sheet against several well-known data sets, including Series G on number of airline passengers provided by the seminal Box-Jenkins book on autoregressive integrated moving averages, or *ARIMA*.

I'm sorry to report that the Forecast Sheet does not return results that correspond to the results returned repeatedly by the literature on time series analysis. Nor do the smoothing constants reported by the utility result in forecasts that are consistent with those provided on the forecast chart. In fairness, this may be due to a different method of choosing the initial forecast values than is used in the literature. But the documentation of the Forecast Sheet, sketchy to begin with, does not address the issue of initialization. It's impossible to back into the algorithm because only result values are reported, not worksheet formulas.

I would have more to say here, but I'm restricted by the terms of a nondisclosure agreement with Microsoft. Let me put it this way:

Several statisticians, including me, have followed the evolution of the `LINEST()` worksheet function since the mid-1990s. We're still not completely satisfied with it, but at least the really erroneous aspects of the function were fixed in Excel's 2003 version. Still, problems that were known and reported in 1995 took *eight years* to fix. The Forecast Sheet utility is a good idea. Let's hope that Microsoft takes less than eight years to fix it.



# 5

## The Part of Tens

### **IN THIS PART . . .**

This Part of Tens doesn't focus specifically on sales forecasting. It does hit on some problems, traps, tips, and tricks that you'll find helpful in setting up your forecasts. I take this chance to tell you about some very helpful tools that you may not have explored before. I also cover some issues with array formulas: a special kind of worksheet formula that's particularly important when it's time to use the important forecasting functions LINEST and TREND.

## Chapter 19

# Ten Fun Facts to Know and Tell about Array Formulas

If you find yourself using regression to forecast sales — and regression is one of the three principal methods used for forecasting — then you’re also going to find yourself using array formulas. Excel has quite a few functions that *require* you to use what’s called “array entry,” and several of them are intended for regression analysis.

Array formulas are quirky little beasts and at least one entire book has been written about their use. At least two forms of array formulas proved so popular a few years ago that Microsoft coded new functions of the normal type so that users wouldn’t have to use array formulas to get the correct results.

To use array formulas effectively, and to use functions in array formulas, you need to know more than just the arguments to the functions. You need to know the dimensions of the range that the results will occupy. And you’d best know how to edit them, because you’re going to want to.

# Entering Array Formulas

The term “array formula” itself is a fuzzy one. It’s true that many array formulas are intended to fill an array of cells on the worksheet. But it’s also true that many array formulas are intended to occupy one cell only. You might find it helpful to think of an array formula as one that processes one or more arrays of data — arrays that might or might not appear on the worksheet. If the arrays aren’t visible on the sheet, they’re used in Excel’s internal processing locations, away from prying eyes.

You need to inform Excel that what you’re entering is an array formula, not a normal one such as this:

```
=AVERAGE(B2:B25)
```

To enter that formula and the function it employs, you select a cell, type the formula, and press Enter. Suppose you want to *array enter* a formula such as this one:

```
=AVERAGE(IF(A2:A25="Zig",B2:B25,""))
```

To do so, you select a cell, type the formula, and simultaneously hold down the Ctrl and the Shift keys as you press Enter. If you have done things right, the formula shows up in the formula bar surrounded by curly brackets. And the cell where you array-entered the formula will show the average of the values, perhaps sales revenues, in A2:A25 for any record with the text “Zig” in B2:B25.

AVERAGEIF and SUMIF are two of the functions coded by Microsoft to give you an alternative to the array formula methods. Maybe I’m just being reactionary, but I’ve always preferred the array formulas. I think it gives me more control over what’s going on.

## Using the Shift Key

There’s an important difference between entering a formula by means of the keyboard combination Ctrl+Enter, and by means of Ctrl+Shift+Enter. Both result in an array of results if you begin by selecting a range of cells, but only the formula entered using Ctrl+Shift+Enter is what Excel conventionally terms an “array formula.” Here’s the difference.

When you array-enter a formula, using Ctrl+Shift+Enter, you are entering *one* formula in multiple cells. Those cells normally display different results, often because your formula uses a function such as LINESST or TRANSPOSE that is

designed to return different values in different cells. Or the different results can come about because the formula's arguments include a range of cells.

It's a subtle difference (and I admit that at first it struck me as a *transcendentally* subtle difference), but using Ctrl+Enter results in different formulas in the selected cells, where Ctrl+Shift+Enter results in the same formula in each of the selected cells. See Figure 19-1.

	A	B	C	D	E
1	1		1		
2	2		2		
3	3		3		
4	4		4		
5	5		5		
6					
7	1		1	=A7	
8	2		2	=A8	
9	3		3	=A9	
10	4		4	=A10	
11	5		5	=A11	
12					
13	1		1	=A13:A17	
14	2		2	=A14:A18	
15	3		3	=A15:A19	
16	4		4	=A16:A20	
17	5		5	=A17:A21	
18					

**FIGURE 19-1:**  
Each cell in C1:C5  
contains the same array  
formula.

The cells in the range C1:C5 share the same array formula. The *results* that appear in the individual cells are different because the array formula returns an array of five different values.

The cells in the range C7:C11 have different formulas, although only one was entered via the keyboard. I selected C7:C11 and typed this formula:

```
=A7
```

Then I entered the formula with Ctrl+Enter. That had the same effect as if I had entered the formula normally in C7, copied it, and pasted it into C8:C11. That is, the formula adjusted its references to point at A8, A9, A10, and A11. Different cells, different formulas.

Cells C13:C17 were populated the same way as C7:C11, except that the formula I typed was:

```
=A13:A17
```

Again, I used Ctrl+Enter to fill all five cells in C13:C17. The formula also adjusted its references as shown in E13:E17. The results employ what Excel calls the *implicit intersection*. That is, the location of the formula implies the location of the intersection with an array of cells.

So, C13 points at A13:A17 and C13's row intersects A13:A17 at A13. The formula in C13 returns the value found at that intersection, 1. Similarly. The formula in C17 points at A17:A21. It intersects A17:A21 at A17 and the implicit intersection causes the formula to return the value in the intersection, 5.

## Noticing the Curly Brackets

I mentioned in the previous section that when Excel accepts your formula as an array formula, it surrounds what you typed with curly brackets to signify the array formula status. But you see the brackets in the formula bar only, not in the cell that contains the array formula (if you choose to show formulas rather than results in cells), not in the evaluation box displayed by the Evaluate Formula command, not elsewhere.

But even if your array formula returns an error value, the formula bar displays the curly brackets. For example, this formula can return the #DIV/0! error:

```
=SUM(A1 : A6/B1 : B6)
```

If, say, cell B3 contains a zero, then it causes a #DIV/0! error when it's divided into the contents of cell A3 by the array formula. That division by zero error propagates up the line to the SUM function and the full formula returns #DIV/0!. But the formula as seen in the formula bar still has the curly brackets around it.

The lesson: It's often important to know that Excel has treated a formula as an array formula. The curly brackets around the formula in the formula bar are an indicator that it's happened that way. But don't assume that the curly brackets mean the formula is doing what you want it to. Whatever value it returns has to pass your giggle test.

## Using INDEX to Extract a Value from an Array Formula's Result

As you've seen, not all array formulas return arrays with multiple columns and/or multiple rows to the worksheet. But when they do, it can happen that you're

interested in seeing only one value in the array. You can use Excel's INDEX function to help with that.

For example, LINEST is one of the worksheet functions that will work properly only if you array-enter the formula that contains the function. But suppose that you want access to only one cell value in the LINEST results, perhaps to accommodate a worksheet layout in a routine report. In that case, you don't necessarily want the full set of LINEST results, and you can use Excel's INDEX function to pluck out and display only the one you're interested in showing.

For example, here's how you might array-enter LINEST for a multiple regression:

```
=LINEST(A2:A51 , B2:D51 , , TRUE )
```

If you array-enter that formula in a 5 row by 4 column range, the intersection of that range's third row and first column contains the regression's R-squared value. So if you select just a single cell and enter the following formula, you'll get the R-squared value only:

```
=INDEX(LINEST(A2:A51 , B2:D51 , , TRUE ) , 3 , 1)
```

Here, you're supplying INDEX with the array of values returned by the LINEST function. That's the first argument to INDEX. The second and third arguments to INDEX are the numbers 3 and 1, which instruct INDEX to find the value in the third row and first column of the array and return it to the worksheet.

When I first started experimenting with this sort of arrangement I was surprised to find that I could enter the full INDEX formula as just given normally, with an array of LINEST results as its first argument, without the Ctrl and Shift and Enter combination — that is, without array-entering it. (Try it both ways, both array entering it and entering it normally.)

And yet if I tried entering the following single-cell array formula, it produced the error #VALUE! if I tried to enter it normally:

```
=IF (H44639:H44644>0 , G44639:G44644 , 0)
```

The only reason I could come up with is that when the formula calls a function that Excel expects to take an array as an argument, the formula can be entered normally. That's the case with this formula:

```
=INDEX(LINEST(A2:A51 , B2:D51 , , TRUE ) , 3 , 1)
```

The LINEST results are nested within the INDEX function, where they act as its first argument. Excel expects INDEX to take an array of values as its first

argument — parsing an array is what INDEX was born to do. So the formula as given does not need to be array-entered.

In contrast, this single cell array formula must be array-entered:

```
=AVERAGE(IF(A2:A25="Zig",B2:B25,""))
```

In this case, Excel does not expect that the IF function will take an array of values as an argument, but here we're presenting not one but two arrays of values to IF: the range A23:A25 and B2:B25. (You can even take the position that there's an array of 24 instances of "" implied by the first two arguments.) Because the formula does not meet Excel's initial expectation of the arguments to IF, you have to draw Excel's attention to the situation, and you do so by array-entering the formula.

## A Quick Route to Unique Values

From time to time you come across a way to do something faster or in a more graceful way than the Ribbon offers you. It's surprising how often that unexpected method relies on array formulas.

Suppose that you have a table that contains information on revenues in one column and discounts in an adjacent column. As a sales manager you might wonder how often a 5 percent discount is negotiated, how often a 10 percent discount, and so on.

Excel has several ways to get you that information. One good approach is to create a pivot table based on the discount data. You could put the discount data into the row field and also into the values field, using Count as the summary statistic. The pivot table would show you the count of each specific discount. And you could group the discounts to get a count of sales with a discount of from 5 percent to 10 percent, 10 percent to 15 percent, and so on.

That takes a little time and you might use a data filter instead. I go into this in more detail in Chapter 20, but briefly — you might select your list of discounts, and click Advanced in the Ribbon's Data tab to bring up the Advanced Filter. Fill the Unique Values Only check box. Then:

- »» Either click OK to hide duplicate records by hiding the rows they occupy, or
- »» Fill the Copy To Another Location check box, enter that location, and click OK. This will leave the worksheet rows alone and put a copy of the unique discount values elsewhere.



That's quicker than a pivot table but it has some minor drawbacks. Hiding rows can create problems in other columns. And copying unique values to another location can overwrite existing data — you're not warned in this case and Undo doesn't undo it.

The fastest way is to use the FREQUENCY function. You're normally advised to use FREQUENCY in an array formula with two different ranges: a data range and a bins range. The bins range tells Excel how to bracket the values into groups: the number of values between 0.00 and 0.05, between 0.05 and 0.1, and so on.

But if you use the data range as the bins range, FREQUENCY returns the unique records. See Figure 19-2.

	A	B	C	D	E	F	G
1	Revenue	Discount					
2	\$ 536	20%	4				
3	\$ 559	15%	10				
4	\$ 582	25%	4				
5	\$ 585	15%	0				
6	\$ 617	15%	0				
7	\$ 618	15%	0				
8	\$ 624	25%	0				
9	\$ 629	5%	3				
10	\$ 641	10%	3				
11	\$ 642	15%	0				
12	\$ 652	20%	0				
13	\$ 653	10%	0				
14	\$ 696	20%	0				
15	\$ 719	15%	0				
16	\$ 732	5%	0				
17	\$ 751	15%	0				
18	\$ 772	20%	0				
19	\$ 775	5%	0				
20	\$ 777	10%	0				
21	\$ 862	15%	0				
22	\$ 891	15%	0				
23	\$ 898	25%	0				

**FIGURE 19-2:**  
Any value in column B that's paired with a zero in column C is a duplicate.

Just array-enter this formula in C2:C25:

```
=FREQUENCY(B2:B25,B2:B25)
```

The row where a value in column B first appears gets the count of that value in the FREQUENCY column — here, column C. Subsequent instances of that value get a zero in column C. So, simply by entering an array formula you find the unique values and get a count of each. The pivot table approach is doubtless the stronger way to go, but if what you need is quick and dirty and accurate, an array formula might be just what you need.

# Selecting the Range: LINEST

If you're about to array-enter a formula that will occupy a range of cells on the worksheet, you need to know the dimensions of that range. That's because array formulas and the functions that they employ do not automatically populate the necessary range of cells. You need to begin by selecting that range yourself.

Knowing the range's dimensions is largely a matter of experience. If you're about to array-enter the LINEST function, for example, you need to know that you should start by selecting a range that's five rows high and with as many columns as you have predictor variables, plus 1.

So if you have one variable to be predicted in, say, column A and two predictor variables in columns B and C, you're expected to know that you need to start by selecting a range that's at least five rows high and at least three columns wide, such as E1:G5.

# Selecting the Range: TRANSPOSE

But it's not enough to know the size of the range required by the function you're entering via an array formula. You also have to understand its purpose.

A good example is the TRANSPOSE function. From time to time it happens that you want to turn an array of values by 90 degrees. See Figure 19-3 for an example.

	A	B	C	D	E	F
1	Month	Jan	Feb	Mar	Apr	May
2	Revenue	\$ 3,877	\$ 2,722	\$ 4,869	\$ 4,224	\$ 4,392
3						
4			Month	Revenue		
5			Jan	3877		
6			Feb	2722		
7			Mar	4869		
8			Apr	4224		
9			May	4392		

**FIGURE 19-3:**  
There are two good ways to put the array in A1:F2 into C4:D9.

Plenty of more technical situations (for example, various calculations in matrix algebra) require the transposition of an array of values, but the need probably arises more often when you're dealing with worksheet layout issues. Figure 19-3 shows some month name abbreviations in A1:F1 and some currency values in

A2:F2. If you want to put those values into a pivot table, you probably also want to reorient them as shown in C4:D9.



WARNING

I'm about to show you how to use an array formula to copy Figure 19-3's A1:E2 into C5:D9. It might occur to you to make a table of the range C4:D9. Unfortunately, Excel tables cannot include multiple-cell array formulas.

One approach is to select A1:F2, copy it with your method of choice, and select C4. Then choose Paste from the Clipboard group on the Ribbon's Home tab and click the Transpose icon in the first group of Paste commands.

The result is to switch the row-by-column orientation of A1:F2 into C4:D9. This is often exactly what you want, particularly if your purpose was to accommodate the existing page layout in a report.

But suppose that the information in A1:F2 might change from time to time. As more months go by, results from earlier months might be revised. In that sort of case, you would probably want the data in C4:D9 to be updated along with the data in A1:F2. That's one useful aspect of Excel's TRANSPOSE function.

Don't bother to copy A1:F2. Instead, begin by selecting C4:D9. Then array-enter this formula:

```
=TRANSPOSE(A1:F2)
```

The result looks just like what you see in Figure 19-3, but instead of values in C4:D9, that range contains an array formula. Therefore, if the data in A1:F2 changes, the changes are reflected in C4:D9.

The more general point to take from this section is that you need to know what TRANSPOSE does in order to select the range that will contain it, prior to array-entering the function. With a function such as LINES, you need to know to select a range five rows high, with a number of columns that depends on the number of variables you have to analyze. With a function such as TRANSPOSE, you need to derive the rows and columns *and their orientation* from the rows and columns of the original array.

## Selecting a Range: TREND

If you're going to use regression to forecast sales, I hope you've come across my recommendations in earlier chapters that you use TREND to obtain the actual forecasts and LINES to obtain the regression equation itself as well as

additional statistics that inform you about the quality of the equation. These additional statistics include R-squared, which tells you how much variability in the sales figures is shared with the predictor variable or variables.

It's important to use both functions. TREND helps out by relieving you of having to do the arithmetic involved in applying the regression equation. LINEST gives you useful information about whether your forecasts are likely accurate, acceptable, or random garbage.

Your raw sales data is likely in the form of an Excel table or pivot table, or possibly a list.



REMEMBER

A list is just data in a rectangular range. A table is a list that's been enhanced with filter and sort capabilities, a row at the bottom that can show data summaries, and various other handy tools.

If your raw data is oriented in that way, with different variables in different columns and different records in different rows, you'll begin to array-enter the TREND function by selecting a range that's one column wide and with as many rows as you have records in the data source. There's only one variable to be forecast, so only one column is needed. You want to know the forecast value for each record, so your TREND range needs as many rows as there are records.

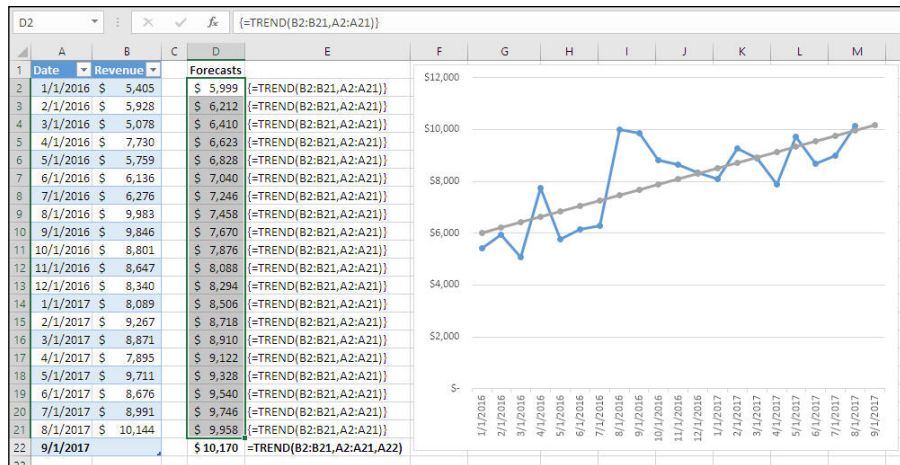
However, most of the forecasts from TREND are for periods that have already passed. You have no special need for those forecasts except as a way to assess the accuracy of the regression equation. So this form of an array formula that uses the TREND function calls for only the known values of the predictor(s) and the variable to be predicted:

```
=TREND(B2 : B21 , A2 : A21 )
```

See Figure 19-4. The range D2:D21 contains that TREND function in its array formula, but it does not forecast the subsequent predictor value in cell A22. It does return the forecasts for the first 19 records, which are useful to have on the worksheet because they can be charted against the actual sales results. The chart in Figure 19-4 gives you a visual sense of the accuracy of the regression equation in generating forecasts.

To get the next forecast value, whose actual will eventually occupy cell B21, you need to supply the next predictor value as TREND's third argument:

```
=TREND(B2 : B21 , A2 : A21 , A22)
```



**FIGURE 19-4:** Using TREND to forecast periods that are now in the past.

So, you need to array-enter the first of the two previous formulas in D2:D21 and the second in D22. Knowing what you're telling TREND to do is essential for knowing the size and orientation of the range into which you'll enter the array formulas.



TIP

You could enter the formula in D22 normally, and I have done so in Figure 19-4. It occupies one cell only and it calls a function that expects arrays as its arguments. Even so I recommend against doing that — instead, array-enter the formula. There's likely to come a time when you have more than just one cell as a true future value: for example, A22:A25 instead of just A22. Then you'll need to array-enter the formula, and you might as well get used to doing so.

## Editing an Array Formula

Inevitably there comes a time when you have to make a change in an array formula. If it's a small change that you want to make, it's usually pretty easy to do so.

Suppose you want to change a range address used in an array formula from A2:A21 to A2:A22. Just select a cell in the array formula's range, make your change, and press **Ctrl+Shift+Enter**. Just as when you array-enter a formula to begin, you need to use all three keys when you edit it. However, you need not select the entire range occupied by the array formula in order to edit it. You have to do that when you establish the formula, but not when you edit it.

Sometimes I need to reduce the dimensions of the range that an array formula occupies. Then, it's a little trickier to manage. Suppose that a LINEST analysis tells me that I might as well forecast column C from column A, instead of from both columns A and B.

Now I want to do two things:

- » Remove the reference to column B from the LINEST arguments.
- » Remove one column from the range that LINEST occupies.

If all I did was to remove the reference to column B from the LINEST arguments, I'd wind up with a bunch of #N/A values in its final column. So, one way to manage things is to select a cell in the array formula's range, delete the equal sign at the start of the formula, and reenter as a conventional formula what has become a text value by Ctrl+Enter.

Then I select the original range except for its final column. I alter the LINEST argument as necessary (including replacing the equal sign) and establish the edit with Ctrl+Shift+Enter. I finish by clearing what had been the range's final column — I can do that now because it's no longer part of an existing array formula. I now have a revised formula with no detritus from the original version hanging around.

It probably seems lots simpler to just delete the entire array formula and reenter it from scratch. Maybe it's just a personal tic, but I seem to screw up less frequently using this method.

## Deleting Array Formulas

By the way, if you want to delete an array formula, you have to start by selecting the entire range of cells that it occupies. Try to delete just one cell and Excel texts you: "You can't change part of an array."

## Chapter 20

# The Ten Best Excel Tools

If you're like most people, definitely including me, you probably don't get the most out of Excel, because you don't know all of what it has to offer. I've used Excel for well over a third of my life, but I didn't know about some of the tools in this chapter until I read about them in online forums and blogs. Cell Comments, AutoComplete, toolbar customization — these aren't the flashiest tools in Excel's kit, but they will all save you time and maybe even grief.

## Cell Comments

When you use Cell Comments, you can make notes about the contents of a worksheet cell. You can document where the information came from, how confident you are of its accuracy, whether it might need revision and when. All this information can be important in building a baseline for a forecast.

If several people are entering data and formulas in a workbook, Cell Comments are really helpful. You don't have to see them unless you want to, but when you want to see them, they can provide great backup.

To enter a Cell Comment, follow these steps:

1. **Select the cell where you want to make a comment.**
2. **Either right-click the cell and choose Insert Comment, or choose New Comment from the Comments group in the Ribbon's Review tab.**
3. **Type into the comments box anything that will help explain what's in that cell and click another cell in your worksheet to close the comment.**

Now the cell will have a small triangle in its upper-right corner, signifying that a comment is there.



TIP

To read a comment, you can just hover your pointer over the cell with the comment in it, and the comment pops up for you to read.

You can deal with comments in a variety of ways. Start by going to the Review tab. In the Comments group, you have several actions available:

- » Click **New Comment** (if the active cell has no comment) or **Edit Comment** (if the active cell has a comment).
- » **Delete:** Click the Delete icon to delete the comment from the active cell.
- » **Previous** and **Next:** Choose one of these to show the previous or the next comment.
- » **Show/Hide Comment:** Use this toggle to show or hide a comment in the active cell.
- » **Show All Comments:** If you can see the text of all comments on a worksheet, click this to show the comment indicators only (the comment indicator is the triangle in the cell's upper-right corner). If you see the indicators only, click to show the comments themselves.

## AutoComplete

I keep my company's checkbook register in an Excel workbook. The deposits and the payments are associated with just a few clients and vendors — for example, my main sources of revenue, my bank, my broker, my office building, my accountant.

Being able to type just a couple of letters in the cell where I show the payee, and having Excel complete the rest of the name is really handy. AutoComplete saves



time and helps keep me from mistyping. Suppose the person you lease an office from is named Maxwell, and your broker's name is Madoff.

In column C, assume you already have a cell that contains Maxwell (but no cell that contains another value beginning with *M*). Now, if you type **M** in column C (the case doesn't matter), Excel offers to finish it off with "axwell." All you have to do now is press Enter or the Tab key. If you don't want to enter Maxwell, for example, and you do want to type Madoff, you just keep typing what you want.

If column C already has a Maxwell *and* a Madoff, and you type **Ma** in a cell in column C, Excel waits — it doesn't know yet whether you mean Maxwell or Madoff. As soon as you type another letter — in this case, *x* or *d* — that distinguishes the two existing entries, Excel finishes the entry for you.



REMEMBER

AutoComplete works only with entries in the same column, not the same row. You cannot selectively turn off AutoComplete (for example, if you wanted to see Maxwell as a possible cell entry but for some reason no longer wanted to see Madoff). But you can turn the feature off completely. Click the File tab and choose Options from the navbar. Click Advanced in the Excel Options navbar and clear the Enable Autocomplete for Cell Values check box.

## Macro Security

Microsoft had a shock a few years ago when it turned out that someone with too much time on his hands could write VBA code that would make Excel (or any Office application, for that matter) go nuts. This little monster is called a *macro virus*. The first one that showed up displayed a message box that said "I think this makes my point."

I confess that I've written one or two macro viruses, but after I calmed down I didn't distribute them. They can be nasty, and you want to protect yourself. An Excel macro's reach extends far beyond Excel. It can delete files, rename files, and just generally do damage.



TIP

It helps to know that more recent versions of Excel save a workbook that has a macro with the extension *.xlsm* or *.xlam* rather than *.xlsx*.

Excel has four levels of macro protection, and you can choose the one you want to use.

If you never share Excel workbooks with anyone else (that means co-workers, clients, your Uncle Joe), you might be able to save some time this way:

1. **Click the Ribbon's File tab.**
2. **Choose Options from the navbar.**
3. **Click Trust Center in the Excel Options navbar.**
4. **Click Trust Center Settings.**
5. **Select the Enable All Macros option button.**

**Note:** Microsoft doesn't recommend this option because, if you ever do get a workbook from somewhere else, you could be in trouble: If the workbook has a macro virus, you won't be warned.

If you *do* share workbooks with other people, or open workbooks that other people have created, you can arrange to be warned that a workbook has potentially hazardous macros by clicking any one of the three Disable option buttons:

- » If you choose to disable macros *with* notification, you will see warnings that editing and content are disabled. You can override that status and enable the macros if you choose to do so.
- » If you choose to disable macros *without* notification, you will not see warnings and the name of the macros will not appear via the Ribbon's Developer tab.
- » You can choose to disable macros unless they have been digitally signed via an application that identifies the source of the digital signature.

It's entirely possible to open workbooks that you think will have any macros disabled, only to find that the macros are enabled. There are various overrides such as trusted publishers and trusted locations — it's likely that your copy of Excel regards you as a trusted publisher. Don't be surprised to see a workbook pass by the sentry unscathed. Just sayin'.



TIP

Opening the workbook but disabling the macros is a good option if you think you know where your workbooks come from, but you're not certain of the good intentions of the source.

## The Customizable Toolbar

Lots of experienced Excel users think, understandably, that the Ribbon and its tabs are inflexible. But they're not. You can remove a tab from the Ribbon, and subsequently replace it. You can put any command on a tab as long as it's a

custom tab, so you can build a tab that contains the commands you use most frequently.

For years, I wasted time by not customizing my toolbars in older versions of Excel. For example, I frequently search for a particular value in a worksheet. I used to use the Find command and type in the value I was looking for. If you do that often, you may want to put the Find command in a more convenient place than tacked onto the extreme right end of the Home tab. Here's how:

- 1. Click the Ribbon's File tab.**
- 2. Choose Options from the navbar.**
- 3. Choose Customize Ribbon from the Excel Options navbar.**
- 4. Click the New Tab button at the bottom of the Main Tabs list box.**  
A new tab and a new group appear near the top of that list box.
- 5. Click on the new tab and drag it to the top of the box.**
- 6. Choose All Commands from the Choose Commands From drop-down.**
- 7. Scroll down and click Find in the Choose Commands From list box.**
- 8. Click the New Group (Custom) item in the Main Tabs list box.**
- 9. Click the Add button.**
- 10. Click OK.**

You now have a new tab and group at the left end of the Ribbon, with the Find command on it. To get back to the default Ribbon layout, return to the Customize Ribbon dialog box and click Reset in the lower-right corner of the dialog box.

A good alternative is to customize the Quick Access toolbar, using a similar sequence of steps.

## Evaluate Formula

I admit I didn't start out as a big fan of Excel's formula auditing. In some instances I found it helpful to choose Trace Precedents from the Formula Auditing group of the Ribbon's Formulas tab, to see which cells a formula refers to. For example, if cell C8 has the following formula:

```
=A1 + Q37
```

then Trace Precedents can help a bit if you need to see the values that are involved. The Trace Dependents link is helpful if you want to clear the contents of, say, cell Q74 but you're not sure whether some other, valuable cell depends on whatever formula or value Q74 contains.

Beyond that I never found much use for the Trace commands. The Evaluate Formula command, though, is a different story. I love it.

Start by selecting a cell with a formula. Then go to the Formulas tab and choose Evaluate Formula from the Formula Auditing group, and you can see step by step how Excel evaluates that formula. This is a wonderful way to help figure out how you've gone wrong in entering a formula. And believe me: It's happened to me a lot.

Click the Evaluate button to show the value of the underlined expression. Click Step In to evaluate a cell that the active cell references, which itself might contain a formula. (If the referenced cell contains a constant instead, its value is returned.) If your formula is returning an unexpected value, Evaluate Formula is likely to show you where you screwed up.

## Worksheet Protection

These days, almost everyone has workbooks that other people use. Whether the workbooks are simultaneously opened and shared by others (I don't recommend this) or they're just available to others through shared folders, your colleagues probably have access to your work. And that means you need to look after your work.

Excel makes protecting your work pretty easy. Open a worksheet that you want to protect and go to the Home tab. Choose Format in the Cells group and click Protect Sheet. The Protect Sheet dialog box appears, giving you various options for what to protect. If you want to protect any cells against being changed, you need to lock them first. Select those cells and choose Lock Cells in the Protect Sheet drop-down. Then protect the worksheet.



TIP

Make sure to select the Protect Worksheet and Contents of Locked Cells check box. Provide a password in the Password to Unprotect Sheet box. If you do this, you're better protected against other users changing your values and formulas.



REMEMBER

Worksheet protection isn't completely secure. You can buy a *password cracker* — software that figures out your password — and if you can buy it, that means other people can buy it, too. Free VBA code is floating around the Internet that will do the same thing. But unless you need to protect your work from someone who's

both knowledgeable and really determined, Excel's protection is probably sufficient. If you want something that's *easy* to hack, buy an iPhone.

## Unique Records Only

Excel has a way to show you individual unique values that are in a table or list. This feature can be helpful when you have repeated values, and you want to view the specific unique values. For example, if you have a table of dates when sales were recorded, those dates are probably repeated. Or if you want a list of all your sales reps for the last five years, a sales file is a good place to look, but most of the names are likely to appear more than just once.

To get a list of unique individual values, follow these steps:

- 1. Select the cells in your existing table or list, or, if no other data is directly adjacent to the data, just select any cell in your table or list.**
- 2. Go to the Ribbon's Data tab and choose Advanced from the Sort & Filter group.**

The Advanced Filter dialog box appears.

- 3. Select the Unique Records Only check box.**
- 4. Click OK.**

Excel hides any rows with duplicate records. I don't care for that behavior, so I generally use the Copy to Another Location option so my original data is left as is — but be careful if you do so. Suppose you choose to copy to another location, and you specify, say, cell F1. If there's already data in column F, it can be overwritten by the new filtered list — and you can't undo it. You'll find a good alternative to this approach, using FREQUENCY, described in Chapter 19.

## Using the Fill Handle

The active cell on a worksheet has a small square on its lower-right corner. That square is called the *fill handle*. You can use that fill handle to copy and paste the cell's data in any direction. Just put your mouse pointer over the fill handle, press the mouse button, and drag down, right, left, or up. This technique works both with formulas and constants. If you start with a multiple selection, you can extend a series of values such as 1, 2 to go as far as you'd like.

Suppose you have a set of revenue figures in cells A1:A500. You'd like to enter a 15 percent commission on each of those figures in B1:B500. You could enter this formula in cell B1:

```
=A1 * 0.15
```

Then click the fill handle in B1 and drag down through B500. The problem is that it's tedious. I've even had the process, which Excel calls *autofill*, run away with me and I find myself filling into B422865 before I know it.

The solution — and I've loved this ever since running into it on a blog — is to *double-click* the fill handle. Excel autofills whatever is in the active cell as far down as an adjacent list extends. In the example I just cited, double-clicking B1's fill handle autofills the formula down through B500.

## Quick Data Summaries

If you don't like typing formulas — and who does? — you can quickly get data summaries on the Status Bar. These summaries just show you the result — they don't save anything. Suppose you want to know the smallest number (or the largest number, or the average, or the sum, and so on) in a list. Just follow these steps:

- 1. Right-click the Status Bar — you'll find it at the bottom of the Excel window — to get a shortcut menu.**
- 2. Choose None (to display no summary value) or Average, Count, Numerical Count, Min, Max, or Sum.**

Now, when you select a range of cells on the worksheet, the Status Bar will display the summary value that you chose. You can also use the shortcut menu to enable or suppress other information that might show up on the Status Bar, such as the status of the Caps Lock and Scroll Lock keys.

The Status Bar itself can be suppressed and reenabled, but only through code such as VBA.

## Help with Functions

Ever had trouble remembering what information you need to provide to an Excel function? For some functions, like SUM and AVERAGE, no problem: You just enter

the name of the function, and then an open parenthesis, drag through the cells involved, and enter a close parenthesis.

But for more-complicated functions such as PMT (which shows you the amount of a recurring payment for a given loan amount, interest rate, and number of payments) or HYPERLINK (which you use to put a hyperlink into a worksheet), remembering what the inputs are and what order to enter them can be difficult. The order in which you enter the inputs is important because, for example, Excel will interpret this:

```
=PMT(360, .005, 100000)
```

as meaning that the interest rate is 360 percent and that you're going to make only 0.005 payments. Instead, you need to enter something like this:

```
=PMT(.005, 360, 100000)
```

which means that the interest rate is one half of 1 percent per payment period, and you're going to make 360 payments.

The Insert Function tool can be helpful for these more-complex functions. To the left of the Formula Bar, you'll see the Insert Function button (labeled  $f_x$ ). Click the Insert Function button to get to the Insert Function dialog box, which lets you first select the function you want to use, and then guides you through entering the inputs you want to use — in the correct order.





# Index

## Symbols

#DIV/0! error, 354

#N/A value, 282

## A

absolute references (cells)

defined, 193

names, 346

ACF (autocorrelation function), 307–308

actual revenues, 183–184

actuals, 13, 28

Add Chart Element option (Chart Design menu), 110

add-ins

Data Analysis

avoiding problems with, 115–117

charting residuals, 234–235

Correlation tool, 34, 63–67

defined, 13

Exponential Smoothing tool, 34, 117, 165–168, 269–275

general discussion, 20–21, 112–115

installing, 160–162

Moving Average tool, 117, 162–164, 230–234

overview, 33–34, 159–160, 229–230

Regression tool, 34–35, 67–68, 115–117, 168–172, 182–186, 292–301

tables, 112–117

Solver, 167, 326

Adjusted R-squared (Shrinkage Estimator), 293–295

Advanced Filter dialog box (tables), 98–100

advanced tools, 20–21

alpha smoothing constant, 334–335, 342

Analysis ToolPak (ATP), 112, 159–160. *See also* Data Analysis add-in

Analyze tab (Ribbon), 138

area charts, 145

arguments

defined, 199

Function Arguments dialog box, 202–203

TREND function, 217

ARIMA (autoregressive integrated moving averages), 26–27, 348

array formulas

changing, 209

choosing range for, 206–207

Ctrl+Shift+Enter keyboard sequence, 207–208

curly brackets, 354

deleting, 362

editing, 361–362

entering, 352

INDEX function, 354–356

LINEST function, 358

overview, 205–206, 351

recognizing, 208

Shift key, 352–354

TRANSPOSE function, 358–359

TREND function, 359–361

unique values, 356–357

ATP (Analysis ToolPak), 112, 159–160. *See also* Data Analysis add-in

AutoComplete, 364–365

autocorrelation

calculating, 63, 253–254

defined, 27

diagnosing, 254–257

general discussion, 244–253

autocorrelation function (ACF), 307–308

autofill, 367

autoregression, 76, 306–309

autoregressive integrated moving averages (ARIMA), 26–27, 348

Average summary statistic option (tables), 93

AVERAGEIF function, 352

- averages
  - autoregressive integrated moving averages, 26–27, 348
  - moving
    - calculating and charting, 228–229
    - Data Analysis add-in, 229–235
    - defined, 9
    - noise, 221–224
    - ordering your data, 72–74
    - overview, 14–15, 29, 219
    - signal, 220–221
    - smoothing, 226–228
    - step function, 224–226
- axes (charts)
  - category axis, 104–105, 143
  - date-scaled category axis, 148
  - time-scaled category axis, 148
  - true category axis, 148, 149
  - vertical axis, 104–105

## B

- backcasting, 264, 283, 340
- bar charts, 145
- baselines
  - charting
    - date and time data, 142–146
    - Line charts, 146–150
    - overview, 31–32, 141–142
    - pivot charts, 152–157
    - value axes, 157–158
    - XY charts, 151–152
  - defined, 8
  - differencing, 315–317, 321–325
  - first differencing, 235
  - integrating, 325–327
  - making out of sales data
    - getting subsets with filter fields, 126–128
    - overview, 123–125
    - using column fields, 125–126
    - using row fields, 125
  - organizing data
    - ordering your data, 72–77
    - overview, 71
    - time periods, 77–86

- overview, 27, 30–31, 41
- qualitative data
  - asking right questions, 42–43
  - purpose of forecast, 43–45
- reasons to remove trend from, 312–315
- recovering from mistakes, 45–47
- seasonality
  - overview, 47
  - understanding, 50–51
- signal, 220–221
- time periods
  - accounting for, 45
  - choosing, 81–82
  - deciding how far to forecast, 78–81
  - length of, 44
  - organizing by date, 45
  - overview, 77–78
  - regression, 177
  - spacing equally, 82–86
- trends
  - identifying, 48–50
  - overview, 32–33, 47
  - trends and, 55–58
- bubble charts, 145

## C

- calculating
  - autocorrelation, 63, 253–254
  - exponential smoothing, 338–344
  - moving averages, 228–229
  - seasonal forecast, 338–344
- category axis (charts), 104–105, 143
- cells
  - absolute references
    - defined, 193
    - names, 346
  - Cell Comments feature, 363–364
- Change Chart Type option (Chart Design menu), 110
- Chart Design menu
  - Add Chart Element option, 110
  - Change Chart Type option, 110
  - Chart Stylest option, 110
  - Move Chart option, 110

- Select Data option, 110
- Switch Row/Column option, 110
- charting data
  - moving averages, 228–229
  - regression, 301–305
  - residuals, 234–235
- charts
  - area, 145
  - bar, 145
  - bubble, 145
  - category axis, 104–105, 143
  - column, 145
  - cone, 145
  - creating, 108–109
  - cylinder, 145
  - date-scaled category axis, 148
  - Design tab, 110
  - doughnut, 145
  - Format tab, 111
  - Line
    - axes, 104
    - charting baseline, 31, 32, 146–150
    - overview, 145
  - Line Fit Plot
    - plotting actual revenues, 183–184
    - Regression tool, 293
  - overview, 19
  - pie, 145
  - pivot, 152–157
  - pyramid, 145
  - radar, 145
  - refining
    - formatting axes, 112
    - using Chart menu, 109–111
  - residual plot, 293
  - surface, 145
  - time-scaled category axis, 148
  - true category axis, 148, 149
  - turning tables into
    - chart types, 104–108
    - creating, 108–109
    - overview, 103–104
    - refining charts, 109–112
    - types of, 104–108
    - vertical, 104–105
    - vertical axis, 104–105
    - XY (Scatter)
      - charting baseline, 31, 32, 151–152
      - charting correlated data, 243–244
      - defined, 145–146
      - overview, 105
- coefficients
  - correlation coefficients
    - defined, 242
    - testing statistical significance of, 59–60
  - LINEST function, 289
  - Regression tool, 292, 296–301
  - standard error of each coefficient, 210
  - standard errors and, 297
- column charts, 145
- columns
  - pivot tables, 125–126
  - tables, 88
- commodities, 24–25
- cone charts, 145
- confidence intervals
  - overview, 184
  - Regression tool, 298
- confidence levels, 184–185, 293
- constants
  - alpha smoothing constant, 334–335, 342
  - avoiding zero constant, 185–186
  - defined, 18
  - delta smoothing constant, 334–335, 342
  - LINEST function, 289
  - Regression tool, 292
  - smoothing constants
    - comparison standard, 275–278
    - damping factor and, 28
    - defined, 259
    - minimizing square root of mean square error, 278–282
    - modifying, 273–275
    - seasonal forecast, 334–338
    - zero constant, 185–186
- CORREL function, 27, 65, 243, 253–254

- correlation
  - autocorrelation
    - calculating, 63, 253–254
    - defined, 27
    - diagnosing, 254–257
    - general discussion, 244–253
  - charting data, 243–244
  - negative, 35
  - overview, 27–28, 240–243
  - partial autocorrelation function, 307–309
  - positive, 35
  - quantitative forecast, 67–68
- correlation coefficient
  - defined, 242
  - testing statistical significance of, 59–60
- correlation matrix, 65
- correlation ratio (Eta Squared statistic), 303
- Correlation tool (Data Analysis add-in), 34, 63–67
- Correlograms. xlsx workbook, 307
- cost data, 35–39
- COUNT function, 277
- Count Numbers summary statistic option (tables), 94
- count of observations (Regression tool), 295
- Count summary statistic option (tables), 93
- counting units (pivot tables), 129–130
- Ctrl+Shift+Enter keyboard sequence, 207–208
- curly brackets, 208, 354
- curvilinear regression, 226
- customizing toolbar, 366–367
- cycles
  - defined, 248
  - forecasting, 28
  - product life cycle, 25
- cyclical baseline, 50
- cylinder charts, 145

## D

- damping factor, 28, 260, 272
- data
  - charting
    - moving averages, 228–229
    - regression, 301–305
    - residuals, 234–235
  - data labels, 150

- data series, 147
- date and time, 142–146
- grouping
  - creating groups, 135–137
  - know when to group, 135
  - overview, 134–135
- importing to table to database, 100–102
- ordering your data
  - exponential smoothing, 75–76
  - importance of, 72
  - moving averages, 72–74
  - overview, 12
  - regression, 76–77
- organizing
  - baseline, 71–86
  - tables, 87–117
- qualitative, 42–45
- quantitative, 42
- revenue, 35–39
- smoothing
  - adjusting forecast, 260–262
  - autocorrelation, 244–257
  - damping factor, 260
  - defined, 9–10, 28
  - expanding the equation, 263–264
  - exponents, 264–265
  - losing early averages, 238–240
  - ordering your data, 75–76
  - overview, 16, 34, 237–238
  - problems with, 282–283
  - smoothing constants, 259, 275–282
- summarizing sales data with pivot tables
  - avoiding problems, 137–139
  - making baselines out of sales data, 123–128
  - overview, 121–123
  - totaling data, 128–130
- tables, 10–12
- totaling, 128–130
- Data Analysis add-in
  - avoiding problems with, 115–117
  - charting residuals, 234–235
  - Correlation tool, 34, 63–67
  - defined, 13
  - Exponential Smoothing tool, 34, 117, 165–168, 269–275

- general discussion, 20–21, 112–115
- installing, 160–162
- Moving Average tool
  - general discussion, 230–234
  - overview, 162–164
  - reporting results, 117
  - stationary baselines and, 117
- overview, 33–34, 159–160, 229–230
- Regression tool
  - Adjusted R-squared, 293–295
  - analyzing correlations, 67–68
  - avoiding problems with, 115–117
  - avoiding zero constant, 185–186
  - checking forecast errors, 182–183
  - confidence levels, 184–185
  - general discussion, 168–172
  - multiple regression, 292–301
  - overview, 34–35
- tables
  - avoiding problems, 115–117
  - general discussion, 112–115
- Data Analysis dialog box, 114
- Data tab (Ribbon)
  - accessing Data Analysis tools, 159
  - Advanced Filter dialog box, 98–100, 356
  - Data Analysis dialog box, 65, 113
  - Select Data Source dialog box, 100
  - Sort & Filter group, 89–90
  - using Forecast Sheet, 348
  - using Moving Average tool, 230
  - using Regression tool, 174
- database, importing data to table from, 100–102
- date and time data, 142–146
- date-scaled category axis (charts), 148
- degrees of freedom (df)
  - LINEST function, 210, 292
  - Regression tool, 295
- deleting array formulas, 362
- delta smoothing constant, 334–335, 342
- Design tab (Ribbon)
  - Add Chart Element, 109
  - Convert to Range option, 96
  - overview, 110
  - setting up tables, 88
- Developer tab (Ribbon), 280, 366
- deviations, 341
- df (degrees of freedom)
  - LINEST function, 210, 292
  - Regression tool, 295
- diagnosing
  - autocorrelation, 254–257
  - trends, 313–315
- dialog boxes
  - Advanced Filter, 98–100
  - Function Arguments, 202–203
  - Moving Average, 114, 115, 230
  - Select Data Source, 100–101
  - Top 10 AutoFilter, 97–98
- differencing
  - defined, 311
  - disadvantages of, 321–325
  - first differencing, 235, 249–250, 316
  - second differencing, 316, 317
  - subtracting one value from next value, 316–317
- doughnut charts, 145
- downward trend, 49
- dummy coding, 188
- dynamic range names, 11

## E

- editing array formulas, 361–362
- Einstein, Albert, 43
- entering formulas
  - array
    - changing, 209
    - choosing range for, 206–207
    - Ctrl+Shift+Enter keyboard sequence, 207–208
    - overview, 205–206
    - recognizing, 208
    - Shift key, 352–354
  - defined, 202–203
  - Insert Function, 199–205
  - overview, 191–192
  - reasons for using, 192–198
  - regression functions
    - LINEST function, 210–215
    - TREND function, 215–218

- entering formulas (*continued*)
  - syntax, 198–199
  - using table name in formula, 96
- equation
  - expanding, exponential smoothing, 263–264
  - trendlines and, 107–108
- errors
  - #DIV/0! error, 354
  - baselines, 71
  - charting, 234–235
  - defined, 13, 39
  - exponential smoothing
    - adjusting forecast, 260–262
    - damping factor, 260
    - expanding the equation, 263–264
    - exponents, 264–265
    - problems with, 282–283
    - smoothing constant, 259, 275–282
    - Smoothing tool's formula, 269–275
  - moving averages, 221–224
  - recovering from, 45–47
  - Regression tool, 182–183
  - removing trend from baseline, 312–313
  - root mean square error
    - defined, 277
    - minimizing, 278–282
  - square root of mean square error
    - defined, 277
    - minimizing, 278–282
  - standard error of estimate
    - LINEST function, 290–291
    - multiple regression, 290–291
  - standard errors
    - coefficients and, 297
    - LINEST function, 289
  - tables, 115
- Eta Squared statistic (correlation ratio), 303
- Evaluate Formula command, 367–368
- evaluating regression
  - autoregression, 306–309
  - regressing one trend onto another, 309–310
- Excel 2016 For Dummies (Harvey), 192
- Excel formulas
  - array
    - changing, 209
    - choosing range for, 206–207
    - Ctrl+Shift+Enter keyboard sequence, 207–208
    - curly brackets, 354
    - deleting, 362
    - editing, 361–362
    - entering, 352
    - INDEX function, 354–356
    - LINEST function, 358
    - overview, 205–206, 351
    - recognizing, 208
    - Shift key, 352–354
    - TRANSPOSE function, 358–359
    - TREND function, 359–361
    - unique values, 356–357
  - defined, 202–203
  - Insert Function, 199–205
  - layout, 198
  - overview, 191–192
  - reasons for using, 192–198
  - regression functions
    - LINEST function, 210–215
    - TREND function, 215–218
  - seasonal forecast, 344–345
  - Smoothing tool
    - modifying smoothing constant, 273–275
    - overview, 269–272
  - static values, 197
  - syntax, 198–199
  - using table name in, 96
- Excel tables
  - Advanced Filter dialog box, 98–100
  - Average summary statistic option, 93
  - columns, 88
  - Count Numbers summary statistic option, 94
  - Count summary statistic option, 93
  - creating
    - converting table to list, 96
    - overview, 91–93
    - resizing table, 96
    - Total row, 93–95
    - using table name in formula, 96

- Data Analysis add-in
  - avoiding problems, 115–117
  - general discussion, 112–115
- filtering lists
  - Advanced Filter, 98–100
  - overview, 96–98
- importing data from database, 100–102
- Max summary statistic option, 94
- Min summary statistic option, 94
- More functions option, 94
- None summary statistic option, 93
- overview, 87–88
- pivot tables
  - avoiding problems, 137–139
  - building, 130–134
  - grouping data, 134–137
  - making baselines out of sales data, 123–128
  - overview, 12, 90–91, 121–123
  - totaling data, 128–130
- Select Data Source dialog box, 100–101
- StdDev summary statistic option, 94
- structure, 88–91
- structured references, 96
- Sum summary statistic option, 94
- Top 10 AutoFilter dialog box, 97–98
- Total row, 93–95
- turning into charts
  - chart types, 104–108
  - creating, 108–109
  - overview, 103–104
  - refining charts, 109–112
- Var summary statistic option, 94
- Excel toolbar, 366–367
- Excel tools
  - advanced tools, 20–21
  - AutoComplete, 364–365
  - Cell Comments feature, 363–364
  - Evaluate Formula command, 367–368
  - Exponential Smoothing tool, 34, 117, 165–168, 269–275
  - fill handle, 369–370
  - Insert Function tool, 370–371
  - macro security, 365–366
  - Moving Average tool
    - general discussion, 230–234
    - overview, 162–164
    - reporting results, 117
    - stationary baselines and, 164–165
  - Regression tool
    - Adjusted R-squared, 293–295
    - analyzing correlations, 67–68
    - avoiding problems with, 115–117
    - avoiding zero constant, 185–186
    - checking forecast errors, 182–183
    - confidence levels, 184–185
    - general discussion, 168–172
    - multiple regression, 292–301
    - overview, 34–35, 178–182
    - plotting actual revenues, 183–184
  - Status Bar, 370
  - toolbar, 366–367
  - unique values, 369
  - worksheet protection, 368–369
- exponential smoothing
  - adjusting forecast, 260–262
  - autocorrelation
    - calculating, 253–254
    - diagnosing, 254–257
    - general discussion, 244–253
  - correlation
    - charting data, 243–244
    - overview, 240–243
  - damping factor, 260
  - defined, 9–10, 28
  - expanding the equation, 263–264
  - exponents, 264–265
  - losing early averages, 238–240
  - ordering your data, 75–76
  - overview, 16, 34, 237–238
  - problems with, 282–283
  - seasonal forecast
    - calculating, 342–344
    - calculating first forecast, 338–341
    - relating season to previous forecast, 330–334
    - smoothing constants, 334–338
    - through baseline level, 341–342

- exponential smoothing (*continued*)
  - smoothing constants
    - comparison standard, 275–278
    - defined, 259
    - minimizing square root of mean square error, 278–282
  - Smoothing tool’s formula
    - modifying smoothing constant, 273–275
    - overview, 269–272
- Exponential Smoothing tool (Data Analysis add-in), 34, 117, 165–168, 269–275

## F

- F ratio
  - LINEST function, 291
  - Regression tool, 296
- fill handle, 44, 232, 271, 318, 369–370
- filter fields (pivot tables), 126–128
- filtering (tables)
  - lists
    - Advanced Filter, 98–100
    - overview, 96–98
  - overview, 90
- fire doors, 18
- first differencing, 235, 249–250, 316
- Fisher transformation, 59
- FORECAST function, 218
- forecast period, 28–29, 36
- Forecast Sheet, 348
- forecasting
  - advanced tools, 20–21
  - asking right questions, 42–43
  - autoregressive integrated moving averages, 26–27
  - baseline
    - charting, 31–32
    - defined, 8
    - overview, 27, 30–31
    - trends, 32–33
  - charts, 19
  - correlation, 27–28
  - cycles, 28
  - damping factor, 28
  - Data Analysis add-in, 13, 20–21
  - data preparation
    - ordering your data, 12
    - tables, 10–12
  - exponential smoothing, 9–10, 16, 28, 34
  - forecast period, 28–29
  - general discussion, 7–9
  - moving averages, 9, 14–15, 29
  - naïve, 282
  - overview, 23–24
  - predictor variables, 29
  - purpose of, 43–45
  - quantitative forecasting
    - correlation analysis, 67–68
    - overview, 63–64
    - predictors, 64–67
    - trends, 54–62
  - reasons for
    - planning sales strategies, 24–25
    - sizing inventories, 25–26
  - regression, 10, 16–18, 29, 34–35
  - with revenue and cost data, 35–39
  - seasonality, 30
  - trends, 30
- forecasting, advanced
  - exponential smoothing
    - adjusting forecast, 260–262
    - autocorrelation, 244–257
    - correlation, 240–244
    - damping factor, 260
    - expanding the equation, 263–264
    - exponents, 264–265
    - losing early averages, 238–240
    - overview, 237–238
    - problems with, 282–283
    - smoothing constant, 259, 266–269, 275–282
    - Smoothing tool’s formula, 269–275
- formulas
  - array, 205–209
  - Insert Function, 199–205
  - overview, 191–192
  - reasons for using, 192–198
  - regression functions, 210–218
  - syntax, 198–199



- managing trends
  - differencing, 315–317, 321–325
  - integrating, 325–327
  - link relatives, 318–320
  - overview, 311
  - rates, 320–321
  - reasons to remove trend from baseline, 312–315
- moving averages
  - calculating and charting, 228–229
  - Data Analysis add-in, 229–235
  - noise, 221–224
  - overview, 219
  - signal, 220–221
  - step function, 224–226
  - tracking and smoothing, 226–228
- regression
  - charting, 301–305
  - evaluating, 306–310
  - multiple regression, 286–301
  - overview, 285
- seasonal forecast
  - baseline, 338–344
  - exponential smoothing, 330–344
  - Forecast Sheet, 348
  - modifying formulas, 344–345
  - overview, 329–330
  - using workbook, 346–347
  - using worksheet, 345–346
- forecasting, basic
  - charting baseline
    - date and time data, 142–146
    - Line charts, 146–150
    - overview, 141–142
    - pivot charts, 152–157
    - value axes, 157–158
    - XY charts, 151–152
  - Data Analysis add-in
    - Exponential Smoothing tool, 34, 117, 165–168, 269–275
    - installing, 160–162
    - Moving Average tool, 117, 162–165, 230–234
    - overview, 159–160
    - Regression tool, 34–35, 67–68, 115–117, 168–172, 182–186, 292–301
  - pivot tables
    - avoiding problems, 137–139
    - building, 130–134
    - grouping data, 134–137
    - making baselines out of sales data, 123–128
    - overview, 121–123
    - totaling data, 128–130
  - regression
    - choosing predictor variables, 176–177
    - Data Analysis add-in, 178–186
    - multiple predictor variables, 177–178
    - multiple regression, 186–188
    - overview, 173–175
    - related variables, 176
    - time periods, 177
  - Format tab (charts), 111
  - Format tab (Ribbon), 109–110
- formulas
  - array
    - changing, 209
    - choosing range for, 206–207
    - Ctrl+Shift+Enter keyboard sequence, 207–208
    - curly brackets, 354
    - deleting, 362
    - editing, 361–362
    - entering, 352
    - INDEX function, 354–356
    - LINEST function, 358
    - overview, 205–206, 351
    - recognizing, 208
    - Shift key, 352–354
    - TRANSPOSE function, 358–359
    - TREND function, 359–361
    - unique values, 356–357
  - defined, 202–203
  - Insert Function, 199–205
  - layout, 198
  - overview, 191–192
  - reasons for using, 192–198
  - regression functions
    - LINEST function, 210–215
    - TREND function, 215–218
  - seasonal forecast, 344–345

formulas (*continued*)

Smoothing tool

    modifying smoothing constant, 273–275

    overview, 269–272

static values, 197

syntax, 198–199

using table name in, 96

Formulas tab (Ribbon), 196, 367

FREQUENCY function, 357

Function Arguments dialog box, 202–203

functions

    AVERAGE function, 191

    AVERAGEIF function, 352

    CORREL function, 27, 65, 243, 253–254

    COUNT function, 277

    defined, 191, 203

    FORECAST function, 218

    FREQUENCY function, 357

    HYPERLINK function, 371

    INDEX function, 354–356

    Insert function, 199–205

    LINEST function

        array formulas, 358

        autoregression and, 306–307

        choosing predictors, 65

        defined, 323

        degrees of freedom, 210, 292

        differencing and, 323

        F ratio, 291

        forecasting statistics, 213–215

        F-ratio, 210

        multiple regression, 288–292

        overview, 108, 210–213

        regression sum of squares, 210

        residual sum of squares, 210, 292

        R-squared value, 210, 289–290

        selecting range, 213, 358

        standard error of each coefficient, 210

        standard error of estimate, 210, 290–291

        sum of squares regression, 292

    MINVERSE function, 323

    MMULT function, 323

    OFFSET function, 193–196, 199, 204, 206

    PMT function, 371

    RSQ function, 303

    SUM function, 198

    SUMIF function, 352

    SUMXMY2 function, 276

    T.DIST function, 299–300

    T.INV function, 300–301

    TRANSPOSE function, 200–201, 206–207, 358–359

    TREND function

        array formulas, 359–361

        differencing and, 322, 324–325

        general discussion, 215–218

        multiple regression, 286–288

        regression approach and, 37

## G

Goal Seek tool, 167

grouping data

    creating groups, 135–137

    know when to group, 135

    overview, 134–135

## H

Harvey, Greg, 192

Hawking, Stephen, 43

Holt models, 330

Holt-Winters model, 334, 348

Home tab (Ribbon), 66, 73, 207, 359

HYPERLINK function, 371

## I

implicit intersection., 354

importing data, to table from database, 100–102

INDEX function, 354–356

inflation, trends and, 54

initializing forecast, 340

Insert function, 199–205, 370–371

Insert tab (Ribbon), 87, 91, 95, 103, 131, 153,  
252, 272

integrating baselines, 325–327

intercepts

    alpha smoothing constant, 334–335, 342

    avoiding zero constant, 185–186

    defined, 18

    delta smoothing constant, 334–335, 342

- LINEST function, 289
- Regression tool, 292
- smoothing constants
  - comparison standard, 275–278
  - damping factor and, 28
  - defined, 259
  - minimizing square root of mean square error, 278–282
  - modifying, 273–275
  - seasonal forecast, 334–338
- zero constant, 185–186
- intervals
  - confidence intervals
    - overview, 184
    - Regression tool, 298
  - defined, 230
- inventories
  - just-in-time inventory management, 26
  - sizing, 25–26

## J

- just-in-time (JIT) inventory management, 26

## K

- known x's, 217
- known y's, 217

## L

- layout (formulas), 198
- life cycle (products), 25
- Line charts
  - axes, 104
  - charting baseline, 31, 32, 146–150
  - defined, 145
- Line Fit Plot chart
  - overview, 183–184
  - Regression tool, 293
- linear (straight-line) regression, 58, 225
- LINEST function
  - array formulas, 358
  - autoregression and, 306–307
  - choosing predictors, 65
  - degrees of freedom, 210, 292

- differencing and, 323
- F ratio, 291
- forecasting statistics, 213–215
- F-ratio, 210
- multiple regression, 288–292
- overview, 108, 210–213
- regression sum of squares, 210
- residual sum of squares, 210, 292
- R-squared value, 210, 289–290
- selecting range, 213, 358
- standard error of each coefficient, 210
- standard error of estimate, 210, 290–291
- sum of squares regression, 292
- link relatives, 318–320
- lists
  - converting tables to, 96
  - converting to tables, 87–88, 91–93
  - defined, 11, 87
  - filtering
    - Advanced Filter, 98–100
    - overview, 96–98

## M

- macro security, 365–366
- macro virus, 365
- MAD (Mean Absolute Deviation), 277
- Max summary statistic option (tables), 94
- Mean Absolute Deviation (MAD), 277
- mean square (MS)
  - Regression tool, 296
  - square root of mean square error
    - defined, 277
    - minimizing, 278–282
- Min summary statistic option (tables), 94
- MINVERSE function, 323
- mistakes
  - #DIV/0! error, 354
  - baselines, 71
  - charting, 234–235
  - defined, 13, 39
  - exponential smoothing
    - adjusting forecast, 260–262
    - damping factor, 260
    - expanding the equation, 263–264

- mistakes (*continued*)
  - exponents, 264–265
  - problems with, 282–283
  - smoothing constant, 259, 275–282
  - Smoothing tool's formula, 269–275
- moving averages, 221–224
- recovering from, 45–47
- Regression tool, 182–183
- removing trend from baseline, 312–313
- root mean square error
  - defined, 277
  - minimizing, 278–282
- square root of mean square error
  - defined, 277
  - minimizing, 278–282
- standard error of estimate
  - LINEST function, 290–291
  - multiple regression, 290–291
- standard errors
  - coefficients and, 297
  - LINEST function, 289
- tables, 115
- MMULT function, 323
- More functions option (tables), 94
- Move Chart option (Chart Design menu), 110
- Moving Average dialog box, 114, 115, 230
- Moving Average tool (Data Analysis add-in)
  - general discussion, 230–234
  - overview, 162–164
  - reporting results, 117
  - stationary baselines and, 164–165
- moving averages
  - calculating and charting, 228–229
  - Data Analysis add-in
    - charting residuals, 234–235
    - Moving Average tool, 230–234
    - overview, 229–230
  - defined, 9
  - noise, 221–224
  - ordering your data, 72–74
  - overview, 14–15, 29, 219
  - signal, 220–221
  - smoothing, 226–228
  - step function, 224–226

- MS (mean square)
  - Regression tool, 296
  - square root of mean square error
    - defined, 277
    - minimizing, 278–282
- multicollinearity, 65
- Multiple R (Regression tool), 292, 294
- multiple regression
  - LINEST function, 288–292
  - overview, 186–187
  - predictor variables
    - with existing variable, 188
    - with forecast variable, 187–188
  - predictors, 286–289
  - Regression tool
    - Adjusted R-squared, 295
    - coefficients, 292, 296–301
    - confidence levels, 293
    - count of observations, 295
    - degrees of freedom, 295
    - F ratio, 296
    - intercept, 292
    - line fit plot chart, 293
    - mean square, 296
    - Multiple R, 292
    - residual plot chart, 293
    - significance of F, 296
    - R-squared value, 289–290
    - standard error of estimate, 290–291
    - TREND function, 286–288

## N

- naïve forecasting, 282
- negative correlation, 35
- new x's, 217
- noise
  - #DIV/0! error, 354
  - baselines, 71
  - charting, 234–235
  - defined, 13, 39
  - exponential smoothing
    - adjusting forecast, 260–262
    - damping factor, 260

- expanding the equation, 263–264
- exponents, 264–265
- problems with, 282–283
- smoothing constant, 259, 275–282
- Smoothing tool's formula, 269–275
- moving averages, 221–224
- recovering from, 45–47
- Regression tool, 182–183
- removing trend from baseline, 312–313
- root mean square error
  - defined, 277
  - minimizing, 278–282
- square root of mean square error
  - defined, 277
  - minimizing, 278–282
- standard error of estimate
  - LINEST function, 290–291
  - multiple regression, 290–291
- standard errors
  - coefficients and, 297
  - LINEST function, 289
- tables, 115
- None summary statistic option (tables), 93

## O

- OFFSET function, 193–196, 199, 204, 206
- opportunity costs, 36
- ordering your data
  - exponential smoothing, 75–76
  - importance of, 72
  - moving averages, 72–74
  - overview, 12
  - regression, 76–77
- organizing data
  - baseline
    - ordering your data, 72–77
    - overview, 71
    - time periods, 77–86
  - tables
    - creating, 91–96
    - Data Analysis add-in, 112–117
    - filtering lists, 96–100
    - importing data from database, 100–102

- overview, 87–88
- structure, 88–91
- turning into charts, 103–112

## P

- PACF (partial autocorrelation function), 307–309
- parsimony, 64, 298
- partial autocorrelation function (PACF), 307–309
- per-capita ratios, 320
- periodic relationships, 83–85
- pie charts, 145
- pivot charts, 152–157
- pivot tables
  - avoiding problems
    - blank dates, 137–138
    - multiple groups, 138–139
  - building, 130–134
  - column fields, 125–126
  - counting units, 129–130
  - filter fields, 126–128
  - grouping data
    - creating groups, 135–137
    - know when to group, 135
    - overview, 134–135
  - making baselines out of sales data
    - getting subsets with filter fields, 126–128
    - overview, 123–125
    - using column fields, 125–126
    - using row fields, 125
  - overview, 12, 90–91, 121–123
  - summing revenues, 128–129
  - totaling data
    - counting units, 129–130
    - summing revenues, 128–129
  - value field, 122
- PMT function, 371
- positive correlation, 35
- predictor variables
  - choosing, 176–177
  - defined, 29
  - with existing variable, 188
  - with forecast variable, 187–188
  - using multiple regression, 177–178

- predictors
  - multiple regression, 286–289
  - quantitative forecast, 64–67
- products, life cycle, 25
- p-values (significance of F), 296, 299
- pyramid charts, 145

## Q

- qualitative data (baseline)
  - asking right questions, 42–43
  - purpose of forecasting, 43–45
- quantitative forecasting
  - correlation analysis, 67–68
  - overview, 63–64
  - predictors, 64–67
  - quantitative data, 42
  - trends
    - baselines and, 55–58
    - causes of, 54–55
    - characteristics of, 54
    - testing for, 59–62
- Quick Access toolbar, 366–367

## R

- radar charts, 145
- random shock (step function), 224–226
- range
  - for array formulas, 206–207
  - LINEST function, 213, 358
- rates (trends), 320–321
- ratios
  - correlation, 303
  - F ratio
    - LINEST function, 291
    - Regression tool, 296
  - per-capita, 320
  - Student's t statistic, 299–300
  - turns, 25
- records, grouping
  - creating groups, 135–137
  - know when to group, 135
  - overview, 134–135

- regression
  - charting, 301–305
  - curvilinear, 226
  - defined, 10
  - evaluating
    - autoregression, 306–309
    - regressing one trend onto another, 309–310
  - linear, 58, 225
  - multiple regression
    - LINEST function, 288–292
    - new predictor with existing variable, 188
    - new predictor with forecast variable, 187–188
    - overview, 186–187
    - predictors, 286–289
    - Regression tool output, 292–301
    - R-squared value, 289–290
    - standard error of estimate, 290–291
    - TREND function, 286–288
  - ordering your data, 76–77
  - overview, 16–18, 29, 285
  - predictor variables
    - choosing, 176–177
    - using multiple, 177–178
  - related variables, 176
  - relationships and, 34–35
  - time periods, 177
- regression functions
  - LINEST function
    - array formulas, 358
    - autoregression and, 306–307
    - choosing predictors, 65
    - degrees of freedom, 210, 292
    - differencing and, 323
    - F ratio, 291
    - forecasting statistics, 213–215
    - F-ratio, 210
    - multiple regression, 288–292
    - overview, 108, 210–213
    - regression sum of squares, 210
    - residual sum of squares, 210, 292
    - R-squared value, 210, 289–290
    - selecting range, 213, 358
    - standard error of each coefficient, 210

- standard error of estimate, 210, 290–291
- sum of squares regression, 292
- TREND function
  - array formulas, 359–361
  - differencing and, 322, 324–325
  - general discussion, 215–218
  - multiple regression, 286–288
  - regression approach and, 37
- Regression tool (Data Analysis add-in)
  - Adjusted R-squared, 293–295
  - analyzing correlations, 67–68
  - avoiding problems with, 115–117
  - avoiding zero constant, 185–186
  - checking forecast errors, 182–183
  - confidence levels, 184–185
  - general discussion, 168–172
  - multiple regression
    - Adjusted R-squared, 293–295
    - coefficients, 292, 296–301
    - confidence levels, 293
    - count of observations, 295
    - degrees of freedom, 295
    - F ratio, 296
    - intercept, 292
    - line fit plot chart, 293
    - mean square, 296
    - Multiple R, 292, 294
    - residual plot chart, 293
    - significance of F, 296
  - overview, 34–35, 178–182
  - plotting actual revenues, 183–184
- related variables, 176
- Remember icon, 3
- residual plot chart, 293
- residual sum of squares, 292
- residuals
  - #DIV/0! error, 354
  - baselines, 71
  - charting, 234–235
  - defined, 13, 39
  - exponential smoothing
    - adjusting forecast, 260–262
    - damping factor, 260
    - expanding the equation, 263–264
    - exponents, 264–265
    - problems with, 282–283
    - smoothing constant, 259, 275–282
    - Smoothing tool's formula, 269–275
  - moving averages, 221–224
  - recovering from, 45–47
- Regression tool, 182–183
- removing trend from baseline, 312–313
- root mean square error
  - defined, 277
  - minimizing, 278–282
- square root of mean square error
  - defined, 277
  - minimizing, 278–282
- standard error of estimate
  - LINEST function, 290–291
  - multiple regression, 290–291
- standard errors
  - coefficients and, 297
  - LINEST function, 289
- tables, 115
- resizing tables, 96
- revenues
  - actual, 183–184
  - revenue data, 35–39
  - summing, 128–129
  - trends and, 55–56
- Review tab (Ribbon), 364
- Ribbon
  - Analyze tab, 138
  - Data tab
    - accessing Data Analysis tools, 159
    - Advanced Filter dialog box, 98–100, 356
    - Data Analysis dialog box, 65, 113
    - Select Data Source dialog box, 100
    - Sort & Filter group, 89–90
    - using Forecast Sheet, 348
    - using Moving Average tool, 230
    - using Regression tool, 174
  - Design tab
    - Add Chart Element, 109
    - Convert to Range option, 96
    - overview, 110
    - setting up tables, 88
  - Developer tab, 280, 366
  - Format tab, 109–110

## Ribbon (*continued*)

- Formulas tab, 196, 367
- Home tab, 66, 73, 207, 359
- Insert tab, 87, 91, 95, 103, 131, 153, 252, 272
- Review tab, 364

## root mean square error (RMSE)

- defined, 277
- minimizing, 278–282

## row fields

- pivot tables
  - defined, 122
  - making baselines out of sales data, 125
- tables, 88

## RSQ function, 303

## R-squared value

- defined, 68
- multiple regression, 289–290
- trendlines and, 107

# S

## sales data

- sales history, 8
- summarizing with pivot tables
  - avoiding problems, 137–139
  - making baselines out of sales data, 123–128
- overview, 121–123
- totaling data, 128–130

## sales strategies, planning, 24–25

## Scatter (XY) charts

- charting baseline, 31, 32, 151–152
- charting correlated data, 243–244
- defined, 145–146
- overview, 105

## seasonal effects (seasonal indexes), 335–336, 343

## seasonal forecast

- exponential smoothing
  - calculating, 342–344
  - calculating first forecast, 338–341
  - relating season to previous forecast, 330–334
  - smoothing constants, 334–338
  - through baseline level, 341–342

## Forecast Sheet, 348

- modifying formulas, 344–345
- overview, 329–330

## using workbook, 346–347

## using worksheet, 345–346

## seasonal indexes (seasonal effects), 335–336, 343

## seasonality, 30, 47, 50–51

## second differencing, 316, 317

## Select Data option (Chart Design menu), 110

## Select Data Source dialog box (tables), 100–101

## Shift key (array formulas), 352–354

## Shrinkage Estimator (Adjusted R-squared), 293–295

## Sign Test, 314–315

## signals

- defined, 13, 71
- moving averages, 29, 220–221
- tables, 115

## significance of F (p-values), 296, 299

## simple regression. *See also* regression

- defined, 173
- predictor variables, 176–177
- related variables, 176
- time periods, 177

## smoothing constants

- alpha, 334–335, 342
- comparison standard, 275–278
- damping factor and, 28
- defined, 259
- delta, 334–335, 342
- intercepts, 289
- minimizing square root of mean square error, 278–282
- modifying, 273–275
- seasonal forecast
  - estimating seasonal effects, 335–336
  - overview, 334–335
  - reviewing process, 337–338
  - starting smoothing process, 336–337

## smoothing data

- adjusting forecast, 260–262
- autocorrelation
  - calculating, 253–254
  - diagnosing, 254–257
  - general discussion, 244–253

## correlation

- charting data, 243–244
- overview, 240–243



- damping factor, 260
- defined, 9–10, 28
- expanding the equation, 263–264
- exponents, 264–265
- losing early averages, 238–240
- moving averages, 226–228
- ordering your data, 75–76
- overview, 16, 34, 237–238
- problems with, 282–283
- seasonal forecast
  - calculating, 342–344
  - calculating first forecast, 338–341
  - relating season to previous forecast, 330–334
  - through baseline level, 341–342
- smoothing constants
  - alpha, 334–335, 342
  - comparison standard, 275–278
  - damping factor and, 28
  - defined, 259
  - delta, 334–335, 342
  - intercepts, 289
  - minimizing square root of mean square error, 278–282
  - modifying, 273–275
  - seasonal forecast, 334–338
- Smoothing tool's formula
  - modifying smoothing constant, 273–275
  - overview, 269–272
- Solver add-in, 167, 326
- spacing time periods (baseline)
  - missing data and, 85–86
  - overview, 83
  - periodic relationships, 83–85
- spifs, 245
- square root of mean square error
  - defined, 277
  - minimizing, 278–282
- SS (sum of squares) regression, 292
- standard error of estimate
  - LINEST function, 290–291
  - multiple regression, 290–291
- standard errors
  - coefficients and, 297
  - LINEST function, 289
- static values (formulas), 197
- stationary baselines, 164–165
- stationary series of values, 316
- statistical significance, 299
- statistics
  - Average summary statistic option, 93
  - Count Numbers summary statistic option, 94
  - Count summary statistic option, 93
  - Max summary statistic option, 94
  - Min summary statistic option, 94
  - None summary statistic option, 93
  - StdDev summary statistic option, 94
  - Sum summary statistic option, 94
  - Var summary statistic option, 94
- Status Bar, 370
- StdDev summary statistic option (tables), 94
- step function (random shock), 224–226
- Stout, Rex, 36
- straight-line (linear) regression, 58, 225
- structured references (tables), 96
- Student's t statistic (t statistic; t ratio), 299–300
- SUM function, 198
- sum of squares (SS) regression, 292
- Sum summary statistic option (tables), 94
- SUMIF function, 352
- SUMMARY OUTPUT table, 295
- summary statistics
  - Average summary statistic option, 93
  - Count Numbers summary statistic option, 94
  - Count summary statistic option, 93
  - Max summary statistic option, 94
  - Min summary statistic option, 94
  - None summary statistic option, 93
  - StdDev summary statistic option, 94
  - Sum summary statistic option, 94
  - Var summary statistic option, 94
- summing revenues (pivot tables), 128–129
- SUMXMY2 function, 276
- surface charts, 145
- Switch Row/Column option (Chart Design menu), 110
- syntax
  - formulas, 198–199
  - LINEST function, 172, 211

## T

t ratio (Student's t statistic), 299–300

### tables

Advanced Filter dialog box, 98–100

Average summary statistic option, 93

columns, 88

Count Numbers summary statistic option, 94

Count summary statistic option, 93

creating

converting table to list, 96

overview, 91–93

resizing table, 96

Total row, 93–95

using table name in formula, 96

Data Analysis

avoiding problems, 115–117

general discussion, 112–115

filtering lists

Advanced Filter, 98–100

overview, 96–98

importing data from database, 100–102

Max summary statistic option, 94

Min summary statistic option, 94

More functions option, 94

None summary statistic option, 93

overview, 87–88

pivot tables

avoiding problems, 137–139

building, 130–134

grouping data, 134–137

making baselines out of sales data, 123–128

overview, 12, 90–91, 121–123

totaling data, 128–130

Select Data Source dialog box, 100–101

StdDev summary statistic option, 94

structure, 88–91

structured references, 96

Sum summary statistic option, 94

Top 10 AutoFilter dialog box, 97–98

Total row, 93–95

turning into charts

chart types, 104–108

creating, 108–109

overview, 103–104

refining charts, 109–112

Var summary statistic option, 94

T.DIST function, 299–300

t-distribution, 59

Technical Stuff icon, 3

technology, trends and, 55

testing

Sign Test, 314–315

for trends, 59–62

time periods (baseline)

accounting for, 45

choosing, 81–82

deciding how far to forecast, 78–81

length of, 44

organizing by date, 45

overview, 77–78

regression, 177

spacing equally

missing data and, 85–86

overview, 83

periodic relationships, 83–85

time series, 27. *See also* baselines

charting

date and time data, 142–146

Line charts, 146–150

overview, 31–32, 141–142

pivot charts, 152–157

value axes, 157–158

XY charts, 151–152

defined, 8

differencing, 315–317, 321–325

first differencing, 235

integrating, 325–327

making out of sales data

getting subsets with filter fields, 126–128

overview, 123–125

using column fields, 125–126

using row fields, 125

organizing data

ordering your data, 72–77

overview, 71

time periods, 77–86

- overview, 27, 30–31, 41
- qualitative data
  - asking right questions, 42–43
  - purpose of forecast, 43–45
- reasons to remove trend from, 312–315
- recovering from mistakes, 45–47
- seasonality
  - overview, 47
  - understanding, 50–51
- signal, 220–221
- time periods
  - accounting for, 45
  - choosing, 81–82
  - deciding how far to forecast, 78–81
  - length of, 44
  - organizing by date, 45
  - overview, 77–78
  - regression, 177
  - spacing equally, 82–86
- trends
  - effect of, 55–58
  - identifying, 48–50
  - overview, 32–33, 47
- time-scaled category axis (charts), 148
- T.INV function, 300–301
- Tip icon, 2
- toolbar (Excel), 366–367
- tools (Excel)
  - advanced tools, 20–21
  - AutoComplete, 364–365
  - Cell Comments feature, 363–364
  - Evaluate Formula command, 367–368
  - Exponential Smoothing tool, 34, 117, 165–168, 269–275
  - fill handle, 369–370
  - Insert Function tool, 370–371
  - macro security, 365–366
  - Moving Average tool
    - general discussion, 230–234
    - overview, 162–164
    - reporting results, 117
    - stationary baselines and, 164–165
  - Regression tool
    - Adjusted R-squared, 293–295
    - analyzing correlations, 67–68
    - avoiding problems with, 115–117
    - avoiding zero constant, 185–186
    - checking forecast errors, 182–183
    - confidence levels, 184–185
    - general discussion, 168–172
    - multiple regression, 292–301
    - overview, 34–35, 178–182
    - plotting actual revenues, 183–184
  - Status Bar, 370
  - toolbar, 366–367
  - unique values, 369
  - worksheet protection, 368–369
- Top 10 AutoFilter dialog box (tables), 97–98
- Total row (tables), 93–95
- totaling data (pivot tables)
  - counting units, 129–130
  - summing revenues, 128–129
- tracking
  - moving averages, 228–229
  - smoothing constants and, 266–269
- TRANSPOSE function, 200–201, 206–207, 358–359
- TREND function
  - array formulas, 359–361
  - differencing and, 322, 324–325
  - general discussion, 215–218
  - multiple regression, 286–288
  - regression approach and, 37
- trendlines
  - charts and, 106
  - defined, 35, 85–86
  - regression, 301–305
- trends
  - baseline
    - changing revenues, 55–58
    - identifying, 48–50
    - overview, 47
  - causes of, 54–55
  - characteristics of, 54
  - defined, 30
  - diagnosing, 313–315
  - downward, 49

- trends (*continued*)
  - managing
    - differencing, 316–317, 321–325
    - integrating, 325–327
    - link relatives, 318–320
  - overview, 311
  - rates, 320–321
  - reasons to remove trend from baseline, 312–315
- rates, 320–321
- regressing one trend onto another, 309–310
- Sign Test, 314–315
- testing for, 59–62
- upward, 49

true category axis (charts), 148, 149

turns ratios, 25

## U

- unique values
  - array formulas, 356–357
  - listing, 369
- upward trend, 49

## V

- value axes
  - charting baseline, 157–158
  - charting date and time, 143
- value field (pivot tables), 122
- values
  - #N/A value, 282
  - p-values (significance of F), 296, 299
  - R-squared value
    - defined, 68
    - multiple regression, 289–290
    - trendlines and, 107
  - static, 197
  - stationary series of values, 316
  - unique
    - array formulas, 356–357
    - listing, 369
  - z-value (z-statistic), 61
- Var summary statistic option (tables), 94

- variables
  - autocorrelation
    - calculating, 253–254
    - diagnosing, 254–257
    - general discussion, 244–253
  - correlation
    - charting data, 243–244
    - overview, 240–243
  - predictor variables
    - choosing, 176–177
    - defined, 29
    - with existing variable, 188
    - with forecast variable, 187–188
    - using multiple, 177–178
  - related, 176
  - tables, 88
  - VBA (Visual Basic for Applications) procedure, 278–282
  - vertical axis (charts), 104–105
  - Visual Basic for Applications (VBA) procedure, 278–282

## W

- Warning icon, 3
- workbooks
  - Correlograms. xlsx workbook, 307
  - seasonal forecast, 346–347
- worksheets
  - seasonal forecast, 345–346
  - worksheet protection, 368–369

## X

- XY (Scatter) charts
  - charting baseline, 31, 32, 151–152
  - charting correlated data, 243–244
  - defined, 145–146
  - overview, 105

## Z

- zero constant, 185–186
- z-value (z-statistic), 61

## About the Author

---

**Conrad Carlberg** is the author of more than ten books about Microsoft Excel. As a multi-time recipient of Microsoft's MVP designation for Excel, he is a nationally recognized expert on that application.

Carlberg's Ph.D. in statistics involves work in forecasting, as does his work in telecommunications and the health-care industry. He used the techniques in this book to reduce a crushing \$24 million inventory owned by a Baby Bell to under \$10 million in 18 months. The carrying costs for \$24 million in equipment are significant. The point: This forecasting stuff works.

As preparation for starting his consultancy, Carlberg spent two years as a sales engineer for a Fortune 500 company. He lives near San Diego, where he tries his best to keep from crashing into other sailboats.

## Dedication

---

For Joe Frazier, Mike Kobluk, and Chad Mitchell: Show me a pretty little number.

## Author's Acknowledgments

---

I want to thank Kathy Ivens, who suggested my name to Wiley for this book and who has been the best coauthor one could hope for on our prior books; Katie Mohr, the acquisitions editor who thought that this one would make sense; Maureen Tullis, the project manager who brought the book about; and the technical editor for this book, who patiently pointed out my numeric missteps, Mike Talley.

## **Publisher's Acknowledgments**

**Executive Editor:** Katie Mohr

**Project and Copy Editor:** Scott Tullis

**Technical Editor:** Mike Tally

**Sr. Editorial Assistant:** Cherie Case

**Production Editor:** Antony Sami

**Project Manager:** Maureen Tullis

**Cover Image:** Chad McDermott/Shutterstock

## **Apple & Mac**

iPad For Dummies,  
6th Edition

978-1-118-72306-7

iPhone For Dummies,  
7th Edition

978-1-118-69083-3

Macs All-in-One  
For Dummies, 4th Edition

978-1-118-82210-4

OS X Mavericks  
For Dummies

978-1-118-69188-5

## **Blogging & Social Media**

Facebook For Dummies,  
5th Edition

978-1-118-63312-0

Social Media Engagement  
For Dummies

978-1-118-53019-1

WordPress For Dummies,  
6th Edition

978-1-118-79161-5

## **Business**

Stock Investing  
For Dummies, 4th Edition

978-1-118-37678-2

Investing For Dummies,  
6th Edition

978-0-470-90545-6

Personal Finance  
For Dummies, 7th Edition

978-1-118-11785-9

QuickBooks 2014  
For Dummies

978-1-118-72005-9

Small Business Marketing  
Kit For Dummies,  
3rd Edition

978-1-118-31183-7

## **Careers**

Job Interviews  
For Dummies, 4th Edition

978-1-118-11290-8

Job Searching with Social  
Media For Dummies,  
2nd Edition

978-1-118-67856-5

Personal Branding  
For Dummies

978-1-118-11792-7

Resumes For Dummies,  
6th Edition

978-0-470-87361-8

Starting an Etsy Business  
For Dummies, 2nd Edition

978-1-118-59024-9

## **Diet & Nutrition**

Belly Fat Diet For Dummies

978-1-118-34585-6

Mediterranean Diet  
For Dummies

978-1-118-71525-3

Nutrition For Dummies,  
5th Edition

978-0-470-93231-5

## **Digital Photography**

Digital SLR Photography  
All-in-One For Dummies,  
2nd Edition

978-1-118-59082-9

Digital SLR Video &  
Filmmaking For Dummies

978-1-118-36598-4

Photoshop Elements 12  
For Dummies

978-1-118-72714-0

## **Gardening**

Herb Gardening  
For Dummies, 2nd Edition

978-0-470-61778-6

Gardening with Free-Range  
Chickens For Dummies

978-1-118-54754-0

## **Health**

Boosting Your Immunity  
For Dummies

978-1-118-40200-9

Diabetes For Dummies,  
4th Edition

978-1-118-29447-5

Living Paleo For Dummies

978-1-118-29405-5

## **Big Data**

Big Data For Dummies

978-1-118-50422-2

Data Visualization  
For Dummies

978-1-118-50289-1

Hadoop For Dummies

978-1-118-60755-8

## **Language & Foreign Language**

500 Spanish Verbs  
For Dummies

978-1-118-02382-2

English Grammar  
For Dummies, 2nd Edition

978-0-470-54664-2

French All-in-One  
For Dummies

978-1-118-22815-9

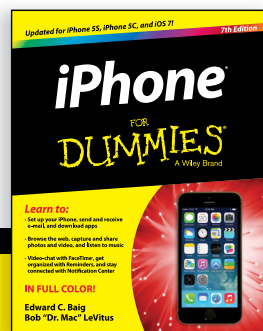
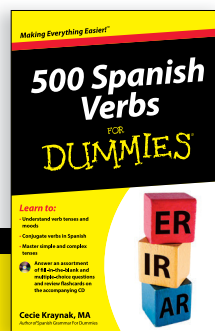
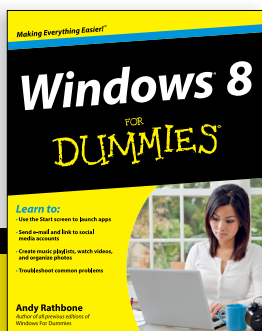
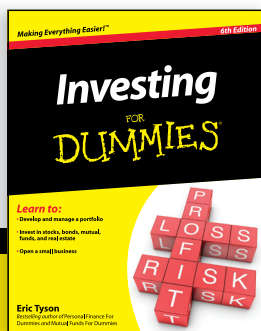
German Essentials  
For Dummies

978-1-118-18422-6

Italian For Dummies,  
2nd Edition

978-1-118-00465-4

 Available in print and e-book formats.



Available wherever books are sold. For more information or to order direct visit [www.dummies.com](http://www.dummies.com)

## **Math & Science**

Algebra I For Dummies,  
2nd Edition  
978-0-470-55964-2

Anatomy and Physiology  
For Dummies, 2nd Edition  
978-0-470-92326-9

Astronomy For Dummies,  
3rd Edition  
978-1-118-37697-3

Biology For Dummies,  
2nd Edition  
978-0-470-59875-7

Chemistry For Dummies,  
2nd Edition  
978-1-118-00730-3

1001 Algebra II Practice  
Problems For Dummies  
978-1-118-44662-1

## **Microsoft Office**

Excel 2013 For Dummies  
978-1-118-51012-4

Office 2013 All-in-One  
For Dummies  
978-1-118-51636-2

PowerPoint 2013  
For Dummies  
978-1-118-50253-2

Word 2013 For Dummies  
978-1-118-49123-2

## **Music**

Blues Harmonica  
For Dummies  
978-1-118-25269-7

Guitar For Dummies,  
3rd Edition  
978-1-118-11554-1

iPod & iTunes  
For Dummies, 10th Edition  
978-1-118-50864-0

## **Programming**

Beginning Programming  
with C For Dummies  
978-1-118-73763-7

Excel VBA Programming  
For Dummies, 3rd Edition  
978-1-118-49037-2

Java For Dummies,  
6th Edition  
978-1-118-40780-6

## **Religion & Inspiration**

The Bible For Dummies  
978-0-7645-5296-0

Buddhism For Dummies,  
2nd Edition  
978-1-118-02379-2

Catholicism For Dummies,  
2nd Edition  
978-1-118-07778-8

## **Self-Help & Relationships**

Beating Sugar Addiction  
For Dummies  
978-1-118-54645-1

Meditation For Dummies,  
3rd Edition  
978-1-118-29144-3

## **Seniors**

Laptops For Seniors  
For Dummies, 3rd Edition  
978-1-118-71105-7

Computers For Seniors  
For Dummies, 3rd Edition  
978-1-118-11553-4

iPad For Seniors  
For Dummies, 6th Edition  
978-1-118-72826-0

Social Security  
For Dummies  
978-1-118-20573-0

## **Smartphones & Tablets**

Android Phones  
For Dummies, 2nd Edition  
978-1-118-72030-1

Nexus Tablets  
For Dummies  
978-1-118-77243-0

Samsung Galaxy S 4  
For Dummies  
978-1-118-64222-1

Samsung Galaxy Tabs  
For Dummies  
978-1-118-77294-2

## **Test Prep**

ACT For Dummies,  
5th Edition  
978-1-118-01259-8

ASVAB For Dummies,  
3rd Edition  
978-0-470-63760-9

GRE For Dummies,  
7th Edition  
978-0-470-88921-3

Officer Candidate Tests  
For Dummies  
978-0-470-59876-4

Physician's Assistant Exam  
For Dummies  
978-1-118-11556-5

Series 7 Exam For Dummies  
978-0-470-09932-2

## **Windows 8**

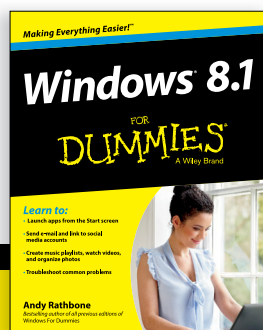
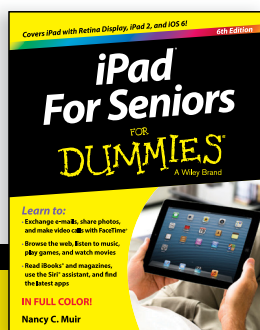
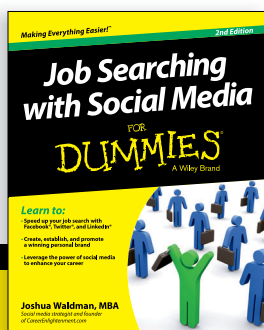
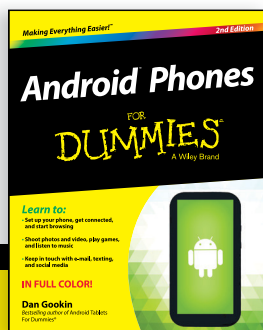
Windows 8.1 All-in-One  
For Dummies  
978-1-118-82087-2

Windows 8.1 For Dummies  
978-1-118-82121-3

Windows 8.1 For Dummies,  
Book + DVD Bundle  
978-1-118-82107-7



Available in print and e-book formats.





# **WILEY END USER LICENSE AGREEMENT**

Go to [www.wiley.com/go/eula](http://www.wiley.com/go/eula) to access Wiley's ebook EULA.