Quantitative Approaches to Decision Making Collection

Donald N. Stengel, *Editor*

Business Applications of Multiple Regression

Second Edition

Ronny Richardson

BEP BUSINESS EXPERT PRESS

Business Applications of Multiple Regression

Second Edition

Business Applications of Multiple Regression

Second Edition

Ronny Richardson
Kennesaw State University

BEP BUSINESS EXPERT PRESS

Business Applications of Multiple Regression, Second Edition
Copyright © Business Expert Press, LLC, 2015.

First published in 2015 by
Business Expert Press, LLC
222 East 46th Street, New York, NY 10017
www.businessexpertpress.com

ISBN-13: 978-1-63157-059-9 (paperback)
ISBN-13: 978-1-63157-060-5 (e-book)

Business Expert Press Quantitative Approaches to Decision Making Collection

Collection ISSN: 2163-9515 (print)
Collection ISSN: 2163-9582 (electronic)

Cover and interior design by S4Carlisle Publishing Services Private Ltd., Chennai, India

First edition: July 2011

10 9 8 7 6 5 4 3 2 1

Printed in the United States of America.

Abstract

This book describes the use of the statistical procedure called multiple regression in business situations, including forecasting and understanding the relationships between variables. The book assumes a basic understanding of statistics but reviews correlation analysis and simple regression to prepare the reader to understand and use multiple regression.

The techniques described in the book are illustrated using both Microsoft Excel and a professional statistical program. Along the way, several real-world data sets are analyzed in detail to better prepare the reader for working with actual data in a business environment.

This book will be a useful guide to managers at all levels who need to understand and make decisions based on data analysis performed using multiple regression. It also provides the beginning analyst with the detailed understanding required to use multiple regression to analyze data sets.

Keywords

correlation, forecasting, model building, multiple regression, regression

Contents

Introduction

Imagine that you are a business owner with a couple of years' worth of data. You have monthly sales figures, your monthly marketing budget, a rough estimate of the monthly marketing budget for your major competitors, and a few other similar variables. You desperately want this data to tell you something. Not only that, you are sure it can give you some business insights if you know more. But what exactly can the data tell you? And once you have a clue what the data might tell you, how do you get to that information?

Really large companies have sophisticated computer software to do data mining. Data mining refers to extracting or "mining" knowledge from large amounts of data.[1] Stated another way, data mining is the process of analyzing data and converting that data into useful information. But how, specifically?

While data mining uses a number of different statistical techniques, the one we will focus on in this book is multiple regression. Why study multiple regression? The reason is the insight that the analysis provides. For example, knowing how advertising, promotion, and packaging might impact sales can help you decide where to budget your marketing dollars. Or knowing how price, advertising, and competitor spending affect demand can help you decide how much to produce. In general, we use multiple regression either to explain the behavior of a single variable, such as consumer demand, or to forecast the future behavior of a single variable, such as sales.

Before you can understand the operation of multiple regression and how to use it to analyze large data sets, you must understand the operation of two simpler techniques: correlation analysis and simple regression. Understanding these two techniques will greatly aid your understanding of multiple regression.

Correlation analysis measures the strength of the linear relationship between a pair of variables. Some pairs of variables, such as sales and advertising or education and income, will have a strong relationship whereas others, such as education and shoe size, will have a weak relationship. We will explore correlation analysis in more detail in chapter 1. As part of that discussion, we will see what it means for a relationship to be linear as well as what it means for the relationship to be strong or weak and positive or negative.

When a pair of variables has a linear relationship, *simple regression* calculates the equation of the line that describes that relationship. As part of simple regression, one variable will be designated as an independent, or explainer, variable and the other will be designated as a dependent, or explained, variable. We will explore simple regression in more detail in chapter 2.

Sometimes, a single variable is all we need to explain the behavior of the dependent variable. However, in business situations, it almost always takes multiple variables to explain the behavior of the dependent variable. For example, due to the economy and competitor actions, it would be a rare business in which advertising alone would adequately explain sales. Likewise, height alone is not enough to explain someone's weight. *Multiple regression* is an extension of simple regression that allows for the use of multiple independent or explainer variables. We will explore multiple regression in more detail in chapter 3.

When using multiple regression with its multiple independent variables, we face the issue of deciding which variables to leave in the final model and which variables to drop from the final model. This issue is made complex by the "diseases" that can affect multiple regression models. We will explore building complex multiple regression models in more detail in chapter 4. It is when we get to model building that we will begin to see the real-world use of multiple regression.

This book assumes you have a background in statistics. Specifically, we will use the normal distribution, Student *t*-distribution, and *F* distribution to perform hypothesis tests on various model parameters to see if they are significant. While it is helpful if you are familiar with these concepts, it is not essential. The software today is advanced enough to present the results in such a way that you can easily judge the significance of a parameter without much statistical background. A brief review is provided in chapter 1.

Correlation, simple regression, and multiple regression can all be performed using any version of Microsoft Excel. Most readers will be able to perform all their analyses in Excel. However, some of the advanced features of multiple regression require an actual statistical package. There are many fine ones on the market, and any of them will perform all the techniques we will discuss. The examples in this book are all either from Excel or from a statistical package called SPSS.

CHAPTER 1

Correlation Analysis

We begin preparing to learn about multiple regression by looking at correlation analysis. As you will see, the basic purpose of correlation analysis is to tell you if two variables have enough of a relationship between them to be included in a multiple regression model. Also, as we will see later, correlation analysis can be used to help diagnose problems with a multiple regression model.

Take a look at the chart in Figure 1.1. This scatterplot shows 26 observations on 2 variables. These are actual data. Notice how the points seem to almost form a line? These data have a *strong correlation*—that is, you can imagine a line through the data that would be a close fit to the data points. While we will see a more formal definition of correlation shortly, thinking about correlation as data forming a straight line provides a good mental image. As it turns out, many variables in business have this type of linear relationship, although perhaps not this strong.

Now take a look at the chart in Figure 1.2. This scatterplot also shows actual data. This time, it is impossible to imagine a line that would fit the data. In this case, the data have a very weak correlation.

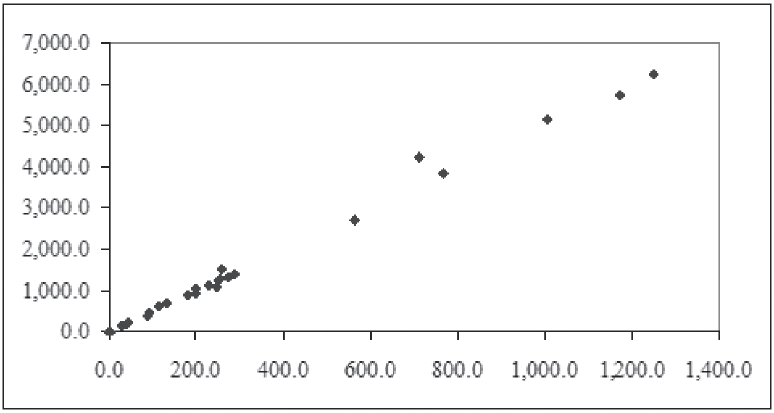


Figure 1.1 A scatterplot of actual data

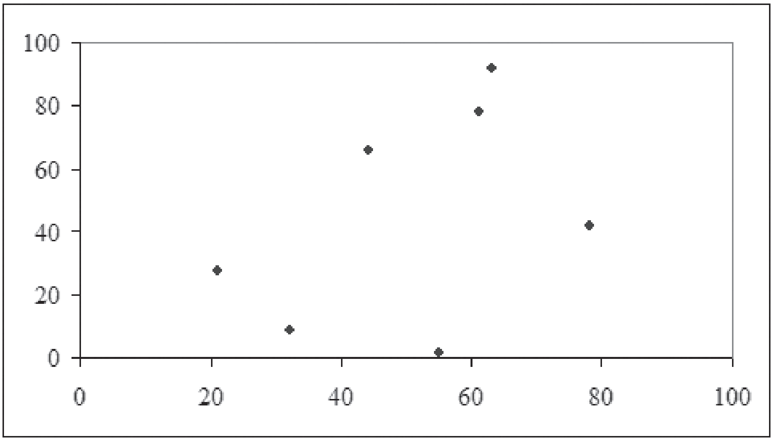


Figure 1.2 Another scatterplot of actual data

Terms

Correlation is only able to find, and simple regression and multiple regression are only able to describe, *linear relationships*. Figure 1.1 shows a linear relationship. Figure 1.3 shows a scatterplot in which there is a perfect relationship between the *X* and *Y* variables, only not a linear one (in this case, a sine wave.) While there is a perfect mathematical relationship between *X* and *Y*, it is not linear, and so there is no linear correlation between *X* and *Y*.

A *positive* linear relationship exists when a change in one variable causes a change in the same direction of another variable. For example, an increase in advertising will generally cause a corresponding increase in sales. When we describe this relationship with a line, that line will have a positive slope. The relationship shown in Figure 1.1 is positive.

A *negative* linear relationship exists when a change in one variable causes a change in the opposite direction of another variable. For example, an increase in competition will generally cause a corresponding decrease in sales. When we describe this relationship with a line, that line will have a negative slope.

Having a positive or negative relationship should not be seen as a value judgment. The terms "positive" and "negative" are not intended to be moral or ethical terms. Rather, they simply describe whether the slope coefficient is a positive or negative number—that is, whether the line slopes up or down as it moves from left to right.

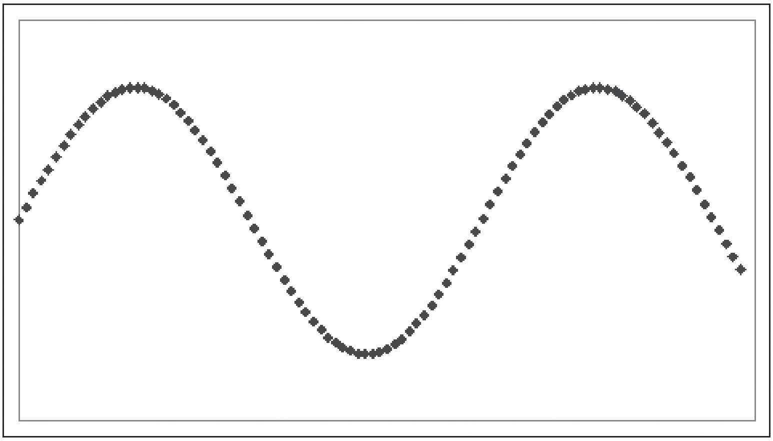


Figure 1.3 A scatterplot of nonlinear (fictitious) data

While it does not matter for correlation, the variables we use with regression fall into one of two categories: *dependent* or *independent* variables. The dependent variable is a measurement whose value is controlled or influenced by another variable or variables. For example, someone's weight likely is influenced by the person's height and level of exercise, whereas company sales are likely greatly influenced by the company's level of advertising. In scatterplots of data that will be used for regression later, the dependent variable is placed on the Y-axis.

An independent variable is just the opposite: a measurement whose value is not controlled or influenced by other variables in the study. Examples include a person's height or a company's advertising. That is not to say that nothing influences an independent variable. A person's height is influenced by the person's genetics and early nutrition, and a company's advertising is influenced by its income and the cost of advertising. In the grand scheme of things, everything is controlled or influenced by something else. However, for our purposes, it is enough to say that none of the other variables in the study influences our independent variables.

While none of the other variables in the study should influence independent variables, it is not uncommon for the researcher to manipulate the independent variables. For example, a company trying to understand the impact of its advertising on its sales might try different levels of advertising in order to see what impact those varying values have on sales. Thus the "independent" variable of advertising is being controlled by the researcher. A medical researcher trying to understand the effect of a drug on a disease might vary the dosage and observe the progress of the disease. A market researcher interested in understanding how different colors and package designs influence brand recognition might perform research varying the packaging in different cities and seeing how brand recognition varies.

When a researcher is interested in finding out more about the relationship between an independent variable and a dependent variable, he must measure both in situations where the independent variable is at differing levels. This can be done either by finding naturally occurring variations in the independent variable or by artificially causing those variations to manifest.

When trying to understand the behavior of a dependent variable, a researcher needs to remember that it can have either a *simple* or *multiple* relationship with other variables. With a simple relationship, the value of the dependent variable is mostly determined by a single independent variable. For example, sales might be mostly determined by advertising. Simple relationships are the focus of chapter 2. With a multiple relationship, the value of the dependent variable is determined by two or more independent variables. For example, weight is determined by a host of variables, including height, age, gender, level of exercise, eating level, and so on, and income could be determined by several variables, including raw material and labor costs, pricing, advertising, and competition. Multiple relationships are the focus of chapters 3 and 4.

Scatterplots

Figures 1.1 through 1.3 are scatterplots. A scatterplot (which some versions of Microsoft Excel calls an *XY chart*) places one variable on the Y-axis and the other on the X-axis. It then plots pairs of values as dots, with the *X* variable determining the position of each dot on the X-axis and the *Y* variable likewise determining the position of each dot on the Y-axis. A scatterplot is an excellent way to begin your investigation. A quick glance will tell you whether the relationship is linear or not. In addition, it will tell you whether the relationship is strong or weak, as well as whether it is positive or negative.

Scatterplots are limited to exactly two variables: one to determine the position on the X-axis and another to determine the position on the Y-axis. As mentioned before, the dependent variable is placed on the Y-axis, and the independent variable is placed on the X-axis.

In chapter 3, we will look at multiple regression, where one dependent variable is influenced by two or more independent variables. All these variables cannot be shown on a single scatterplot. Rather, each independent variable is paired with the dependent variable for a scatterplot. Thus having three independent variables will require three scatterplots. We will explore working with multiple independent variables further in chapter 3.

Data Sets

We will use a couple of data sets to illustrate correlation. Some of these data sets will also be used to illustrate regression. Those data sets, along with their scatterplots, are presented in the following subsections.

All the data sets and all the worksheets and other files discussed in this book are available for download from the Business Expert Press website (http://www.businessexpertpress.com/books/business-applications-multiple- regression). All the Excel files are in Excel 2003 format and all the SPSS files are in SPSS 9.0 format. These formats are standard, and any later version of these programs should be able to load them with no difficulty.

Number of Broilers

Figure 1.1 showed the top 25 broiler-producing states for 2001 by both numbers and pounds, according to the National Chicken Council. The underlying data are shown in Table 1.1.

Age and Tag Numbers

Figure 1.2 was constructed by asking seven people their age and the last two digits of their car tag number. The resulting data are shown in Table 1.2. As you can imagine, there is no connection between someone's age and that person's tag number, so this data does not show any strong pattern. To the extent

Table 1.1 Top 25 Broiler-Producing States in 2001

State	Number of broilers (millions)	Pounds liveweight (millions)
Georgia	1,247.3	6,236.5
Arkansas	1,170.9	5,737.3
Alabama	1,007.6	5,138.8
North Carolina	712.3	4,202.6
Mississippi	765.3	3,826.5
Texas	565.5	2,714.4
Delaware	257.7	1,494.7
Maryland	287.8	1,381.4
Virginia	271.5	1,330.4
Kentucky	253.4	1,292.3
California	250.0	1,250.0
Oklahoma	226.8	1,111.3
Missouri	245.0	1,100.0
South Carolina	198.0	1,049.4
Tennessee	198.3	932.0
Louisiana	180.0	890.0
Pennsylvania	132.3	701.2
Florida	115.3	634.2
West Virginia	89.8	368.2
Minnesota	43.9	219.5
Ohio	40.1	212.5
Wisconsin	31.3	137.7
New York	2.3	12.2
Hawaii	0.9	3.8
Nebraska	0.5	2.7
Other	92.4	451.0

that any pattern at all is visible, it is the result of sampling error and having a small sample rather than any relationship between the two variables.

Return on Stocks and Government Bonds

The data in Table 1.3 show the actual returns on stocks, bonds, and bills for the United States from 1928 to 2009.[1] Since there are three variables (four if you count the year), it is not possible to show all of them in one scatterplot. Figure 1.4 shows the scatterplot of stock returns and treasury bills. Notice that there is almost no correlation.

Table 1.2 Age and Tag Number

Age	Tag no.
55	2
21	28
78	42
61	78
44	66
63	92
32	9

Table 1.3 Return on Stocks and Government Bonds

Year	Stocks (%)	Treasury bills (%)	Treasury bonds (%)
1928	43.81	3.08	0.84
1929	-8.30	3.16	4.20
1930	–25.12	4.55	4.54
1931	–43.84	2.31	–2.56
1932	–8.64	1.07	8.79
1933	49.98	0.96	1.86
1934	–1.19	0.32	7.96
1935	46.74	0.18	4.47
1936	31.94	0.17	5.02
1937	–35.34	0.30	1.38
1938	29.28	0.08	4.21
1939	–1.10	0.04	4.41
1940	–10.67	0.03	5.40
1941	–12.77	0.08	–2.02
1942	19.17	0.34	2.29
1943	25.06	0.38	2.49
1944	19.03	0.38	2.58
1945	35.82	0.38	3.80
1946	–8.43	0.38	3.13
1947	5.20	0.57	0.92
1948	5.70	1.02	1.95
1949	18.30	1.10	4.66
1950	30.81	1.17	0.43
1951	23.68	1.48	–0.30
1952	18.15	1.67	2.27

(*continued*)

***Table 1.3 Return on Stocks and Government Bonds** (continued)*

Year	Stocks (%)	Treasury bills (%)	Treasury bonds (%)
1953	–1.21	1.89	4.14
1954	52.56	0.96	3.29
1955	32.60	1.66	–1.34
1956	7.44	2.56	–2.26
1957	–10.46	3.23	6.80
1958	43.72	1.78	–2.10
1959	12.06	3.26	–2.65
1960	0.34	3.05	11.64
1961	26.64	2.27	2.06
1962	–8.81	2.78	5.69
1963	22.61	3.11	1.68
1964	16.42	3.51	3.73
1965	12.40	3.90	0.72
1966	–9.97	4.84	2.91
1967	23.80	4.33	–1.58
1968	10.81	5.26	3.27
1969	–8.24	6.56	–5.01
1970	3.56	6.69	16.75
1971	14.22	4.54	9.79
1972	18.76	3.95	2.82
1973	–14.31	6.73	3.66
1974	–25.90	7.78	1.99
1975	37.00	5.99	3.61
1976	23.83	4.97	15.98
1977	–6.98	5.13	1.29
1978	6.51	6.93	–0.78
1979	18.52	9.94	0.67
1980	31.74	11.22	–2.99
1981	–4.70	14.30	8.20
1982	20.42	11.01	32.81
1983	22.34	8.45	3.20
1984	6.15	9.61	13.73
1985	31.24	7.49	25.71
1986	18.49	6.04	24.28
1987	5.81	5.72	–4.96
1988	16.54	6.45	8.22

Year	Stocks (%)	Treasury bills (%)	Treasury bonds (%)
1989	31.48	8.11	17.69
1990	–3.06	7.55	6.24
1991	30.23	5.61	15.00
1992	7.49	3.41	9.36
1993	9.97	2.98	14.21
1994	1.33	3.99	–8.04
1995	37.20	5.52	23.48
1996	23.82	5.02	1.43
1997	31.86	5.05	9.94
1998	28.34	4.73	14.92
1999	20.89	4.51	–8.25
2000	–9.03	5.76	16.66
2001	–11.85	3.67	5.57
2002	–21.97	1.66	15.12
2003	28.36	1.03	0.38
2004	10.74	1.23	4.49
2005	4.83	3.01	2.87
2006	15.61	4.68	1.96
2007	5.48	4.64	10.21
2008	–36.58	1.59	20.10
2009	25.92	0.14	–11.12

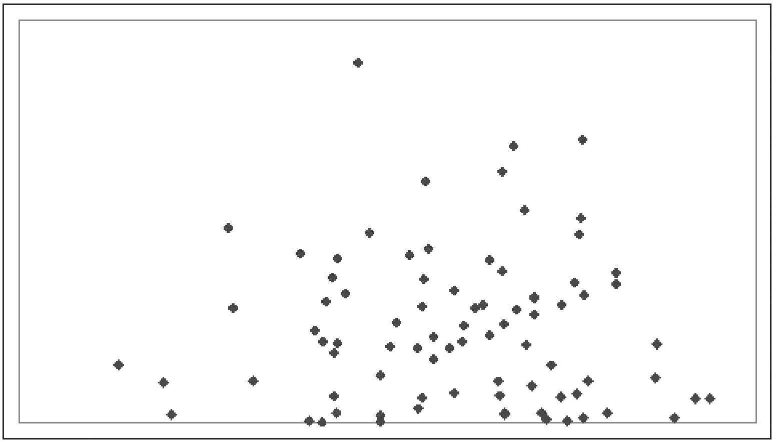


Figure 1.4 Stock returns and treasury bills, 1928 to 2009. X- and Y-axes have been removed for readability

Federal Civilian Workforce Statistics

Table 1.4[2] shows a state-by- state breakdown of the number of federal employees and their average salaries for 2007. Figure 1.5 shows the resulting scatterplot. Notice that there appears to be a fairly weak linear relationship.

Table 1.4 Average Federal Salaries and Number of Employees by State

State	Number of employees	Average salary ($)
Alabama	33,997	64,078
Alaska	11,922	56,525
Arizona	33,871	55,393
Arkansas	12,090	54,176
California	139,804	66,212
Colorado	33,196	67,679
Connecticut	6,854	66,343
Delaware	2,864	57,176
DC	138,622	87,195
Florida	71,858	60,807
Georgia	66,314	61,376
Hawaii	20,759	55,470
Idaho	7,788	58,057
Illinois	42,382	67,385
Indiana	18,577	60,658
Iowa	7,468	55,799
Kansas	15,796	57,528
Kentucky	20,737	52,242
Louisiana	19,011	57,446
Maine	9,128	57,336
Maryland	103,438	79,319
Massachusetts	24,532	67,035
Michigan	23,345	65,576
Minnesota	14,298	62,953
Mississippi	16,576	56,978
Missouri	32,947	56,159
Montana	8,858	55,997
Nebraska	8,826	57,406
Nevada	9,146	59,831

State	Number of employees	Average salary ($)
New Hampshire	3,433	75,990
New Jersey	26,682	72,313
Ohio	41,445	67,638
Oklahoma	33,652	56,603
Oregon	17,649	60,818
Pennsylvania	62,486	59,092
Rhode Island	5,882	73,502
South Carolina	17,158	57,057
South Dakota	7,166	53,000
Tennessee	23,514	57,349
Texas	113,364	59,618
Utah	27,438	54,379
Vermont	3,537	57,279
Virginia	121,337	73,224
Washington	45,948	62,571
West Virginia	13,292	58,964
Wisconsin	11,494	57,404
Wyoming	4,759	54,952

Public Transportation Ridership

Table 1.5[3] shows the largest urbanized areas by population, unlinked passenger trips,[4] and passenger miles for 2008. Figure 1.6 shows the

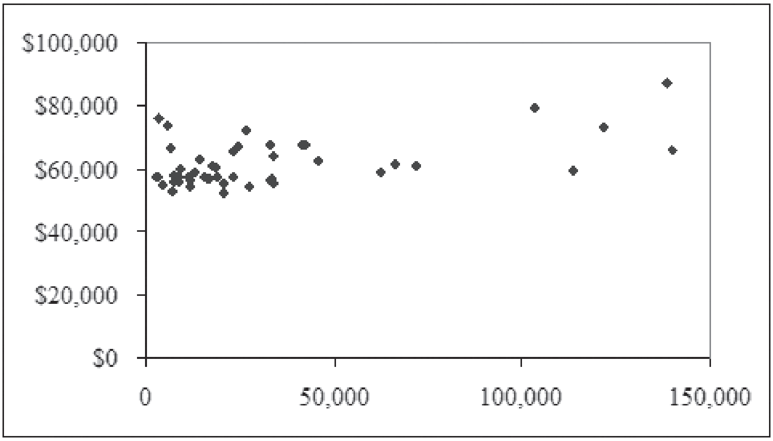


Figure 1.5 Number of federal employees by state and average salaries

Table 1.5 Largest Urbanized Areas by Population, Unlinked Passenger Trips, and Passenger Miles (2008)

Area	Unlinked passenger trips (in thousands)	Passenger miles (in thousands)	Population (2000 Census)
New York-Newark, NY-NJ-CT	4,159,309	21,699,268	17,799,861
Los Angeles-Long Beach-Santa Ana, CA	697,825	3,342,876	11,789,487
Chicago, IL-IN	649,604	4,148,216	8,307,904
Washington, DC-VA-MD	481,776	2,506,203	3,933,920
San Francisco-Oakland, CA	442,185	2,543,376	3,228,605
Boston, MA-NH-RI	377,999	1,881,252	4,032,484
Philadelphia, PA-NJ-DE-MD	361,236	1,726,824	5,149,079
Seattle, WA	195,507	1,284,726	2,712,205
Miami, FL	172,464	1,000,246	4,919,036
Atlanta, GA	162,899	978,010	3,499,840
Baltimore, MD	119,141	764,602	2,076,354
Portland, OR-WA	111,693	467,372	1,583,138
San Diego, CA	104,806	579,977	2,674,436
Denver-Aurora, CO	101,176	554,091	1,984,889
Houston, TX	100,443	632,615	3,822,509
Minneapolis-St. Paul, MN	94,799	490,215	2,388,593
Dallas-Fort Worth-Arlington, TX	76,043	489,618	4,145,659
Phoenix-Mesa, AZ	72,589	315,105	2,907,049
Honolulu, HI	71,310	327,418	718,182
Pittsburgh, PA	69,175	322,026	1,753,136
Las Vegas, NV	66,168	228,917	1,314,357
Cleveland, OH	57,681	263,847	1,786,647
San Juan, PR	56,513	264,342	2,216,616
St. Louis, MO-IL	56,206	315,327	2,077,662
Milwaukee, WI	53,703	178,718	1,308,913
Detroit, MI	53,178	286,301	3,903,377
San Antonio, TX	48,349	218,023	1,327,554
San Jose, CA	44,895	207,074	1,538,312
Salt Lake City, UT	41,714	359,527	887,650
Austin, TX	37,399	161,630	901,920

Area	Unlinked passenger trips (in thousands)	Passenger miles (in thousands)	Population (2000 Census)
Sacramento, CA	37,287	182,727	1,393,498
Cincinnati, OH-KY-IN	30,011	154,207	1,503,262
Virginia Beach, VA	29,268	117,881	1,394,439
Tampa-St. Petersburg, FL	27,710	142,898	2,062,339
Orlando, FL	27,235	166,770	1,157,431
Buffalo, NY	26,173	91,346	976,703
Providence, RI-MA	22,851	110,179	1,174,548
Charlotte, NC-SC	22,721	127,925	758,927
Riverside-San Bernardino, CA	22,605	126,952	1,506,816
Tucson, AZ	18,858	69,853	720,425
Kansas City, MO-KS	17,821	78,210	1,361,744
Rochester, NY	17,653	57,971	694,396
Hartford, CT	17,184	111,520	851,535
Fresno, CA	17,148	37,449	554,923
Columbus, OH	16,662	63,078	1,133,193
New Orleans, LA	16,342	43,726	1,009,283
Louisville, KY-IN	15,593	62,153	863,582
Richmond, VA	14,682	62,340	818,836
Albany, NY	13,903	48,563	558,947
Madison, WI	13,719	48,258	329,533
El Paso, TX-NM	13,180	66,604	674,801
Durham, NC	12,840	61,570	287,796
Memphis, TN-MS-AR	11,514	59,322	972,091
Stockton, CA	5,575	67,948	313,392
Kennewick-Richland, WA	4,894	70,208	153,851

relationship between unlinked passenger trips and population. Notice that almost all the data points are clustered in the bottom left corner of the chart. That is because the New York system has so many more trips (over 4 million versus the next highest of about 700,000) and such a higher population (almost 18 million versus the next highest of almost 12 million) that its observation overpowers the remaining observations. This type of observation outside the usual values is called an *outlier*.[5]

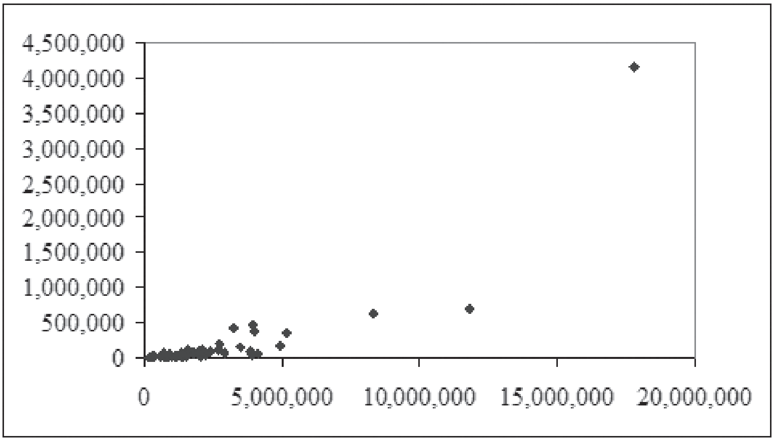


Figure 1.6 Relationship between unlinked passenger trips and population

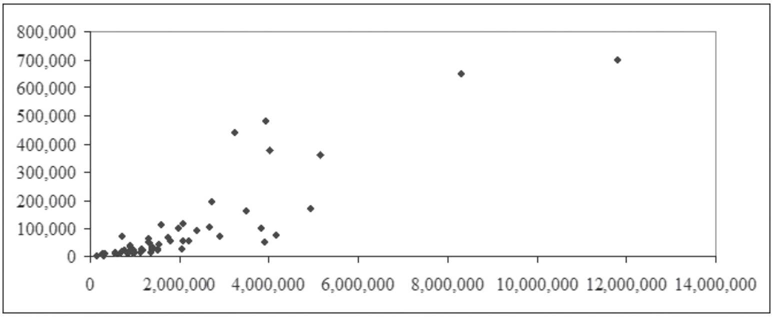


Figure 1.7 Figure 1.6 with the New York observation removed

Figure 1.7 shows the same chart with the New York observation removed. Here you can begin to see how a line might be used to fit the data and how the relationship is positive.

Figure 1.8 shows the relationship between passenger miles and population, again with the New York observation removed. Once again, we see a positive relationship. Figure 1.9 shows the relationship between unlinked passenger trips and passenger miles, again with the New York observation removed. This time, the data form an almost perfectly straight, positive line.

Correlation

Correlation measures the degree of *linear association* between two variables. Correlation can only be measured between pairs of variables, and

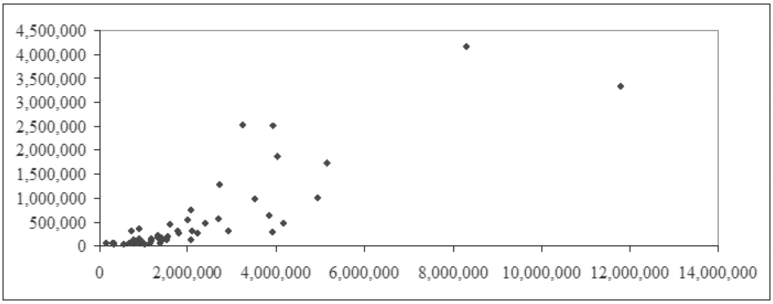


Figure 1.8 Relationship between passenger miles and population with New York observation removed

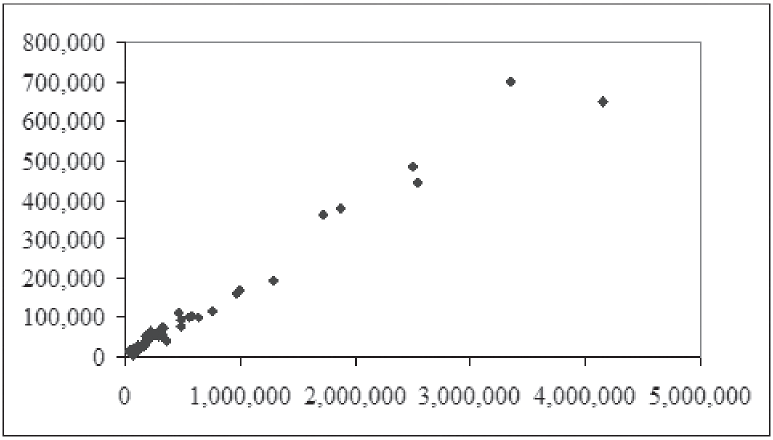


Figure 1.9 Relationship between unlinked passenger trips and passenger miles

it makes no distinction between dependent and independent variables—that is, the correlation between height and weight is exactly the same as between weight and height. The term *correlation analysis* is often used interchangeably with correlation.

Correlation is measured using a statistic called the *correlation coefficient*. The population symbol is the Greek letter rho (ρ), whereas the sample symbol is the letter *r*. The correlation coefficient can take on any value between negative one and positive one. A negative sign indicates a negative relationship, whereas a positive sign indicates a positive relationship. Two variables with a negative relationship will have a line with a negative slope fitted to them, whereas two variables with a positive relationship will have a line with a positive slope fitted to them.

Ignoring the sign, the closer the value is to one (or negative one), the stronger the relationship. The closer the value is to zero, the weaker the relationship. A value of 1 indicates perfect positive linear correlation—that is, all the points form a perfect line with a positive slope. A value of –1 indicates perfect negative linear correlation where all the points form a perfect line with a negative slope. A value of zero indicates no correlation—that is, there is no relationship between the two variables. When this happens, the points will appear to be randomly dispersed on the scatterplot.

It is important to note that correlation only measures *linear* relationships. Even a very strong nonlinear relationship will not be spotted by correlation. So a correlation coefficient near zero only indicates that there is no *linear* relationship, not that there is no relationship. If you look back at Figure 1.3, for example, you can see a clear pattern to the data: a sine wave. The data were generated using a sine wave formula, so a sine wave fits it absolutely perfectly. However, the correlation coefficient for this data is, for all practical purposes, zero.6

Calculating the Correlation Coefficient by Hand

Most likely, you will never need to calculate a correlation coefficient by hand. Excel can easily calculate the value for you, as can any statistical software package. As such, feel free to skip this section if you like. However, seeing and working with the underlying formula can give you some insight into what it means for two variables to be correlated.

The formula to compute the correlation coefficient is as follows:

Correlation Coefficient

$$r = \frac{n\left(\sum X \cdot Y\right) - \left(\sum X\right)\cdot\left(\sum Y\right)}{\sqrt{\left[n\cdot\left(\sum X^2\right) - \left(\sum X\right)^2\right]\cdot\left[n\cdot\left(\sum Y^2\right) - \left(\sum Y\right)^2\right]}}$$

This is a long formula and it looks to be incredibly complex; however, as we will see, it is not all that difficult to compute manually. The first thing to note is that, except for *n* (the sample size), all the terms in this

Table 1.6 Correlation Coefficient Calculations

Age	Tag no.	Age · Tag	Age^2	Tag no.^{2}
55	2	110	3,025	4
21	28	588	441	784
78	42	3,276	6,084	1,764
61	78	4,758	3,721	6,084
44	66	2,904	1,936	4,356
63	92	5,796	3,969	8,464
32	9	288	1,024	81
354	317	17,720	20,200	21,537

equation begin with a summation sign (Σ). It is this characteristic that will allow us to greatly simplify this formula. This is best seen with an example.

Using the data on age and tag numbers from Table 1.2, Table 1.6 shows the interim calculations needed to determine the correlation coefficient. The sample size is seven, so $n = 7$. We will arbitrarily assign Age as X and Tag Number as Y, so, using Table 1.6, $\Sigma X = 354$, $\Sigma Y = 317$, $\Sigma XY = 17{,}720$, $\Sigma X^2 = 20{,}200$, and $\Sigma Y^2 = 21{,}537$.[7] The resulting calculations are as follows:

$$
\begin{aligned}
r &= \frac{n\left(\sum X\cdot Y\right)-\left(\sum X\right)\cdot\left(\sum Y\right)}{\sqrt{\left[n\cdot\left(\sum X^2\right)-\left(\sum X\right)^2\right]\cdot\left[n\cdot\left(\sum Y^2\right)-\left(\sum Y\right)^2\right]}}\\
&= \frac{7(17{,}720)-(354)\cdot(317)}{\sqrt{\left[7\cdot(20{,}200)-(354)^2\right]\cdot\left[7\cdot(21{,}537)-(317)^2\right]}}\\
&= \frac{124{,}040-112{,}218}{\sqrt{[141{,}400-125{,}316]\cdot[150{,}759-100{,}489]}}\\
&\frac{11{,}822}{\sqrt{16{,}084\cdot 50{,}270}} = 0.4157
\end{aligned}
$$

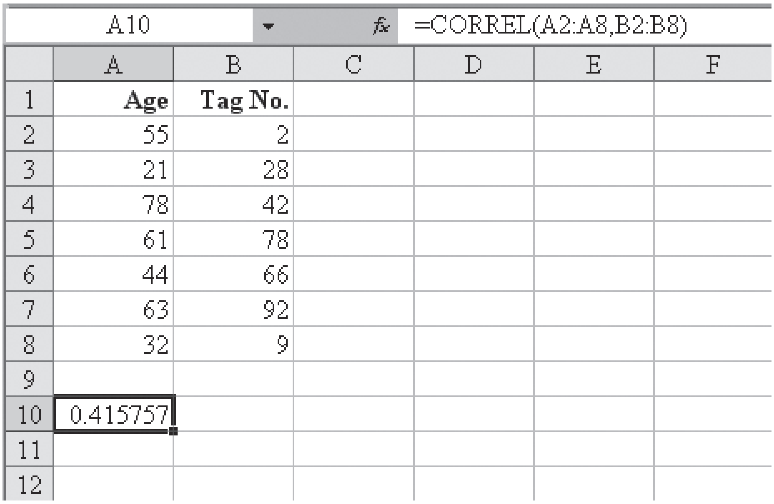
The resulting correlation coefficient of 0.4157 is weak but not zero as you might expect given the lack of a relationship between these two variables. That is a result of the small sampling size and sampling error. However, the real question is if this value is large enough to be *statistically significant*—that is, is this *sample* value large enough to convince us that the *population* value is not zero? We will explore this question in a later section.

Using Excel

Naturally, these calculations can be performed easily using Excel. Excel has two main approaches that can be used to calculate a correlation coefficient: dynamic and static. These two approaches can also be used to compute some of the regression coefficients discussed in later chapters.

The *dynamic approach* uses a standard Excel formula. Doing so has the advantage of automatically updating the value if you change one of the numbers in the data series. For the correlation coefficient, it uses the CORREL function. This function takes two inputs: the range containing the first data series and the range containing the second data series. Since correlation makes no distinction between the independent and dependent variables, they can be entered in either order.

The data in Figure 1.10 are entered in column format—that is, the age variable is entered in one column and the tag number variable is entered in a separate column but side by side. This is the standard format for statistical data: variables in columns and observations in rows. In this case,



A10 | *fx* =CORREL(A2:A8,B2:B8)

	A	B	C	D	E	F
1	Age	Tag No.				
2	55	2				
3	21	28				
4	78	42				
5	61	78				
6	44	66				
7	63	92				
8	32	9				
9						
10	0.415757					
11						
12						

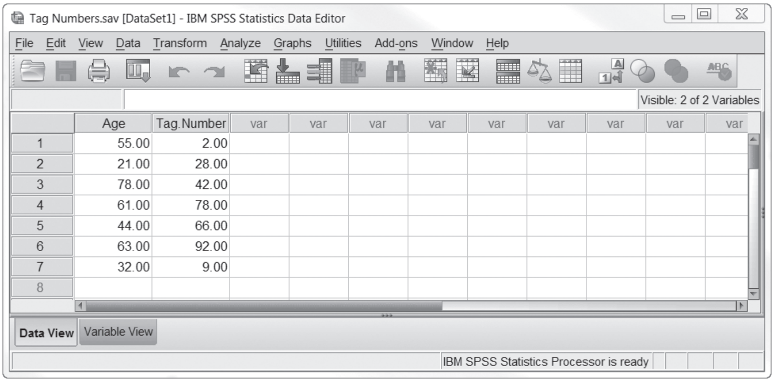
Figure 1.10 Calculating a correlation coefficient using the CORREL *function in Excel*

the data for age are in cells A2 to A8, and the data for tag number are in cells B2 to B8. Row two, for example, represents one observation—that is, someone 55 years old had a tag number that ended in a "02" value.

The following are a few other notes regarding this standard format for statistical data:

1. There should not be any blank columns inside the data set. In fact, having a blank column will cause some of the later procedures we will perform to fail; however, blank columns would not affect correlation analysis.
2. Having blank columns on either side of the data set is a good idea because it sets the data off from the rest of the worksheet and makes data analysis, like sorting, much easier.
3. Having column headings is good; however, they are not required. Some of the later procedures will use these column headings to label the results, which makes those results more readable. Column headings should be meaningful but not too long.
4. There should not be any blank rows inside the data set. While Excel will ignore blank rows in the statistical procedures themselves, blank rows make it more difficult to visualize the data as well as making data sorting difficult.
5. While it does not matter for correlation analysis, since it ignores the dependent/independent status of variables, it is required for multiple regression that all the independent variables be in contiguous columns so the dependent variable should be in either the left or right column of the data set. The generally accepted approach is to use the left column.

While some of these are just "good ideas" in Excel, most statistical software will strictly enforce many of these rules. Figure 1.11 shows the same car tag data inside a professional statistical software package called SPSS.[8] Notice the column format that looks almost identical to Excel, only the variable names are not shown inside a cell the way they are in Excel. Speaking of variable names, notice that the tag number is labeled "Tag.Number" rather than "Tag no". SPSS does not allow spaces in spaces in variable names so a period is used in its place. The data file you can


Tag Numbers.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 2 of 2 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

	Age	Tag.Number	var	var	var	var	var	var	var	var	var
1	55.00	2.00									
2	21.00	28.00									
3	78.00	42.00									
4	61.00	78.00									
5	44.00	66.00									
6	63.00	92.00									
7	32.00	9.00									
8											

Figure 1.11 The car tag data inside SPSS

download is in an older SPSU format that did not allow periods or uppercase letters so it has the variables labeled "age" and "tagno".

SPSS is a very powerful and widely used statistical package. In later chapters, some of the techniques discussed will be too advanced to perform using Excel. These techniques will be illustrated using SPSS. However, any modern statistical package would be able to perform these techniques in a similar manner.

Referring to Figure 1.10, the correlation coefficient is shown in cell A10, and you can see the underlying formula in the formula bar. It is "=CORREL(A2:A8,B2:B8)". The CORREL function is used, and it provides the range of the two variables *without* the column labels.[9]

The *static approach* uses an Excel menu option to perform the calculation of the correlation coefficient. Excel computes the value and then enters it into the worksheet as a *hardwired* number—that is, the actual number is entered into a cell rather than a formula that evaluates to a number. If you then change the data, the correlation coefficient *does not* change since it is just a number. To update its value, you must rerun the menu option.

Before we demonstrate the static approach, we must warn you that not all installations of Excel are ready to perform these calculations. Fortunately, the necessary files are usually installed on the hard drive, and the modification to make Excel ready is quick and only needs to be performed one time. We will see how to prepare Excel before continuing with the example.

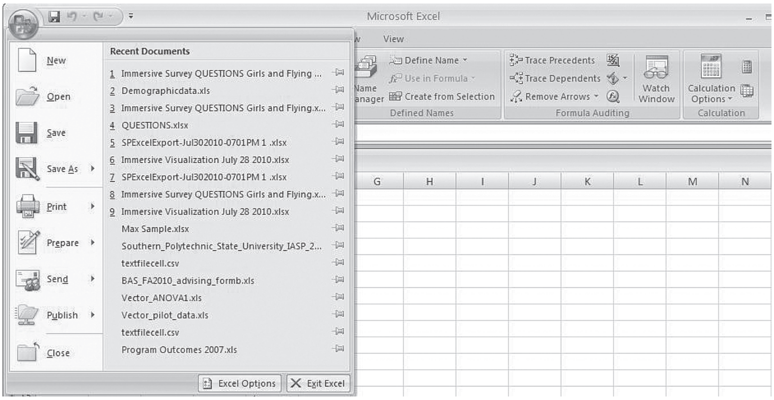
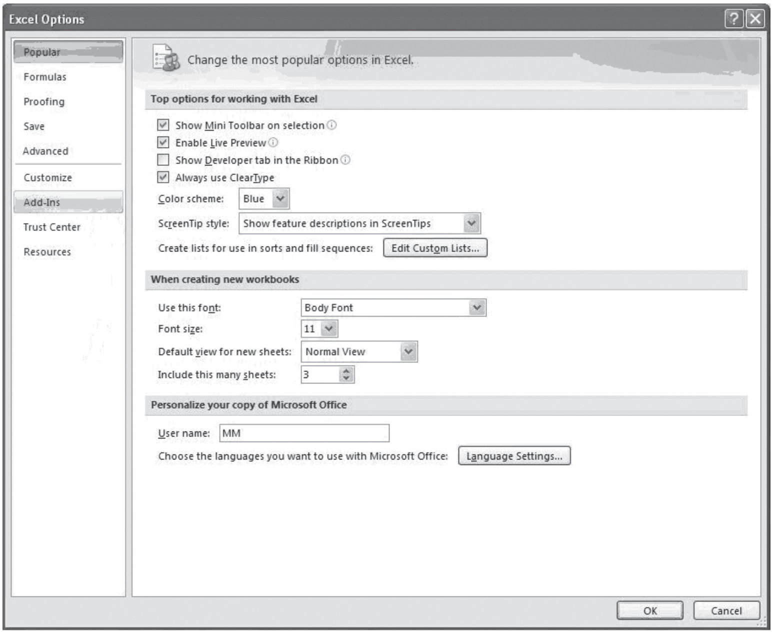

Microsoft Excel
View
New
Open
Save
Save As
Print
Prepare
Send
Publish
Close
Recent Documents
1 Immersive Survey QUESTIONS Girls and Flying ...
2 Demographicdata.xls
3 Immersive Survey QUESTIONS Girls and Flying.x...
4 QUESTIONS.xlsx
5 SPExcelExport-Jul302010-0701PM 1 .xlsx
6 Immersive Visualization July 28 2010.xlsx
7 SPExcelExport-Jul302010-0701PM 1 .xlsx
8 Immersive Survey QUESTIONS Girls and Flying.x...
9 Immersive Visualization July 28 2010.xlsx
Max Sample.xlsx
Southern_Polytechnic_State_University_IASP_2...
textfilecell.csv
BAS_FA2010_advising_formb.xls
Vector_ANOVA1.xls
Vector_pilot_data.xls
textfilecell.csv
Program Outcomes 2007.xls
Excel Options
Exit Excel
Define Name
Use in Formula
Name Manager
Create from Selection
Defined Names
Trace Precedents
Trace Dependents
Remove Arrows
Formula Auditing
Watch Window
Calculation Options
Calculation
G
H
I
J
K
L
M
N

Figure 1.12 The office button dialog box in Excel

Click the Office button, which brings up the dialog box shown in Figure 1.12. Click on the Excel Options button at the bottom, which brings up the *Excel Options* dialog box shown in Figure 1.13. Click on *Add-Ins* on the left and then hit the *GO* button next to *Manage Excel Add-Ins* at the bottom to bring up the *Add-In Manager* shown in Figure 1.14.



Change the most popular options in Excel.

Top options for working with Excel

☑ Show Mini Toolbar on selection

☑ Enable Live Preview

□ Show Developer tab in the Ribbon

☑ Always use ClearType

Color scheme: Blue

ScreenTip style: Show feature descriptions in ScreenTips

Create lists for use in sorts and fill sequences:

When creating new workbooks

Use this font: Body Font

Font size: 11

Default view for new sheets: Normal View

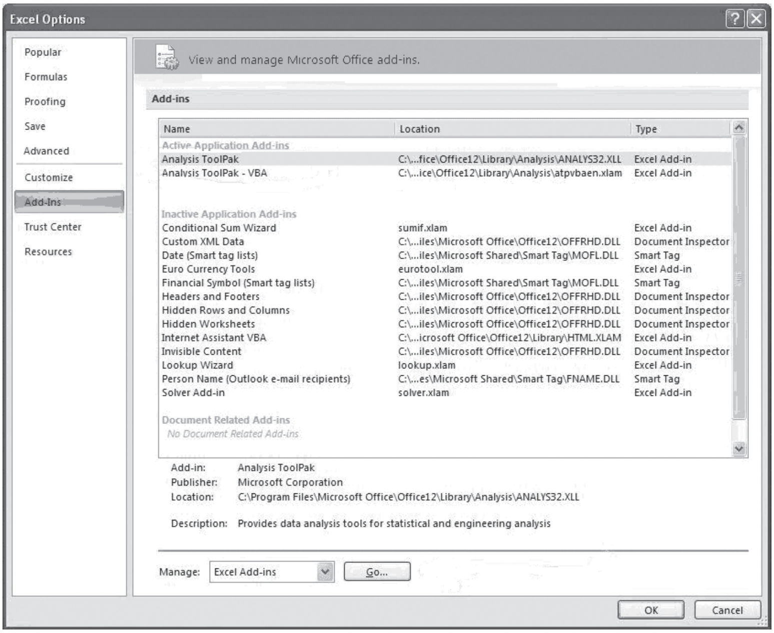
Include this many sheets: 3

Personalize your copy of Microsoft Office

User name: MM

Choose the languages you want to use with Microsoft Office:

Figure 1.13 The Excel options dialog box



Name	Location	Type
Active Application Add-ins		
Analysis ToolPak	C:\...fice\Office12\Library\Analysis\ANALYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...ice\Office12\Library\Analysis\atpvbaen.xlam	Excel Add-in
Inactive Application Add-ins		
Conditional Sum Wizard	sumif.xlam	Excel Add-in
Custom XML Data	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Date (Smart tag lists)	C:\...iles\Microsoft Shared\Smart Tag\MOFL.DLL	Smart Tag
Euro Currency Tools	eurotool.xlam	Excel Add-in
Financial Symbol (Smart tag lists)	C:\...iles\Microsoft Shared\Smart Tag\MOFL.DLL	Smart Tag
Headers and Footers	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Hidden Rows and Columns	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Hidden Worksheets	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Internet Assistant VBA	C:\...icrosoft Office\Office12\Library\HTML.XLAM	Excel Add-in
Invisible Content	C:\...iles\Microsoft Office\Office12\OFFRHD.DLL	Document Inspector
Lookup Wizard	lookup.xlam	Excel Add-in
Person Name (Outlook e-mail recipients)	C:\...es\Microsoft Shared\Smart Tag\FNAME.DLL	Smart Tag
Solver Add-in	solver.xlam	Excel Add-in
Document Related Add-ins		
No Document Related Add-ins		

Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Location: C:\Program Files\Microsoft Office\Office12\Library\Analysis\ANALYS32.XLL

Description: Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins

Figure 1.14 Excel add-in manager under Excel 2010

Click on the *Analysis ToolPak* and *Analysis ToolPak-VBA* and hit *OK* to enable them.

Now that you have the add-ins installed, to compute the correlation coefficient using the static approach, select *Data Analysis* under the Data tab from the Ribbon. This brings up the dialog box shown in Figure 1.15. By default, the procedure you last used during this session will be

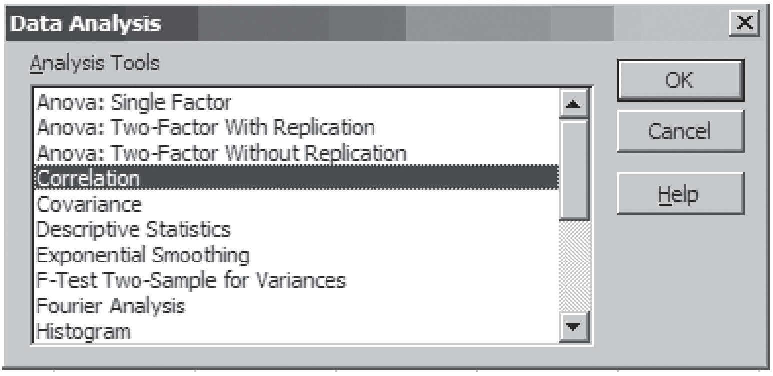
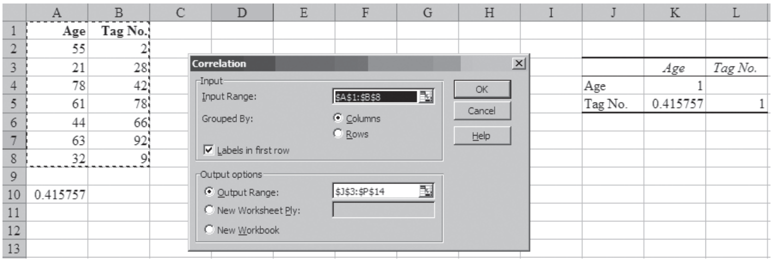


Figure 1.15 Excel data analysis dialog box



	A	B
1	Age	Tag No.
2	55	2
3	21	28
4	78	42
5	61	78
6	44	66
7	63	92
8	32	9
9		
10	0.415757	

Correlation

Input

Input Range: $A$1:$B$8

Grouped By: ● Columns ○ Rows

☑ Labels in first row

Output options

● Output Range: $J$3:$P$14

○ New Worksheet Ply:

○ New Workbook

	Age	*Tag No.*
Age	1	
Tag No.	0.415757	1

Figure 1.16 The Excel correlation dialog box

highlighted. You use this dialog box to select the statistical procedure to perform—*Correlation*, in this case. Selecting *Correlation* and clicking on *OK* brings up the dialog box shown in Figure 1.16.

You use the dialog box in Figure 1.16 to give Excel the information it needs to perform the correlation analysis. At the top of the dialog box, you enter the cells containing the data. If you include the column heading in the range, which is a good idea, Excel will use those titles in the output. If you do, you will need to check the *Labels in first row* box. Excel can perform correlation analysis on data that is stored in either row or column format, so you must tell Excel which format is used. Excel can usually figure it out automatically when headings are included, but it sometimes guesses wrong when only numbers are included in the range.

Finally, you must tell Excel where to store the results. Some statistical procedures, especially regression analysis, take up a lot of space for their output, so it is usually best to store the results in either a large blank area or, even better, a blank worksheet tab, called a *ply* on this dialog box. Just to display the results of this single correlation analysis, for example, Excel required nine worksheet cells. This is shown in Figure 1.16 on the right side of the figure.

At first glance, the output in Figure 1.16 will seem more than a little strange. To understand why this format is used, you need to know that correlation analysis is often applied to many variables all at once. The correlation coefficient itself can only be calculated for pairs of variables, but when applied to many variables, a correlation coefficient is calculated for every possible pair of variables. When more than a couple of variables are used, the format in Figure 1.16 is the most efficient approach to reporting those results.

This format is called a *correlation matrix*. The top row and left column provide the names of the variables. Each variable has a row and column title. Inside this heading row and column are the correlation coefficients. The pair of variables associated with any particular correlation coefficients can be read off by observing the row and column heading for that cell.

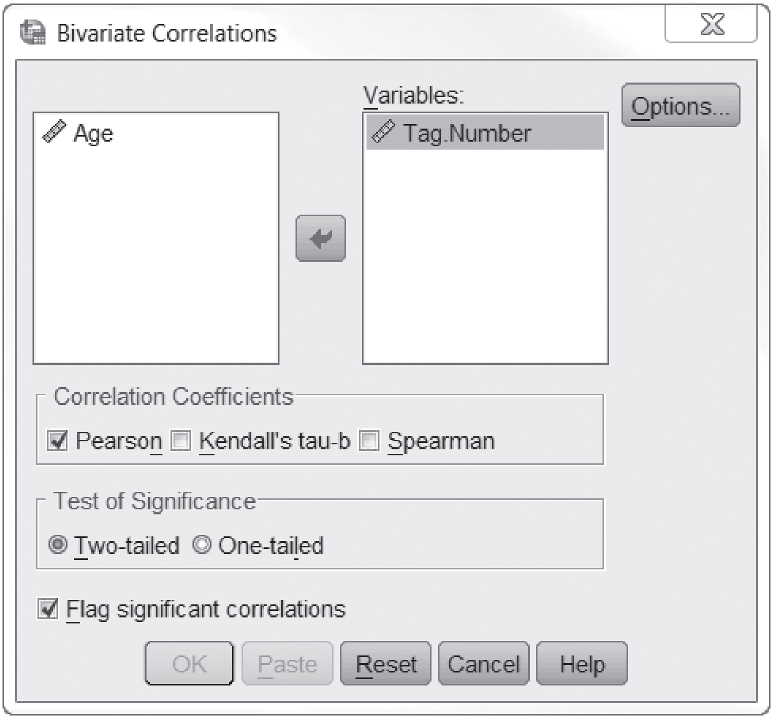
Every correlation matrix will have a value of 1.000 in a diagonal line from the top left cell to the bottom right cell. This is called the *main diagonal*. The cells in the main diagonal have the same row and columns headings and each variable is 100 percent positively correlated with itself, so this diagonal always has a value of one. Notice too that Excel does not show the numbers above the main diagonal. These numbers are a mirror image of the numbers below the main diagonal. After all, the correlation coefficient between *age* and *tag number* would be the same as the correlation coefficient between *tag number* and *age,* so it would be redundant to give the same numbers twice.

The static approach avoids the problem of writing formulas and is especially efficient when the correlation coefficient must be computed for many variables, but it does have a significant drawback. Since the results are static, you must always remember to rerun Correlation if you must change any of the numbers.

Using SPSS

Earlier, we saw how SPSS stores data in column format similar to the way it is stored in Excel. To perform correlation analysis, you click on *Analyze, Correlate*, and *Bivariate*.[10] This brings up the dialog box shown in Figure 1.17 where you select the variables to perform correlation analysis on. A minimum of two variables is required (currently only one is selected), but you can select as many as you like. With more than two variables, SPSS performs correlation on every possible pair of variables.

The result of the correlation analysis in SPSS is shown in Figure 1.18. As discussed, this format for displaying the data is known as a *correlation matrix*. Notice that the correlation matrix has two other numbers in each box besides the actual correlation. The first one is the significances using a two-tailed test. This is also referred to as the *p*-value, which will be discussed later. Notice, too, that the *p*-value is missing on the main diagonal.



Age

Variables: Tag.Number

Correlation Coefficients

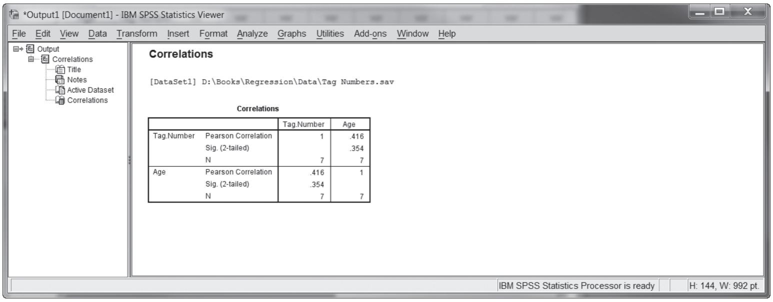
☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance

● Two-tailed ○ One-tailed

☑ Flag significant correlations

Figure 1.17 The SPSS variable selection dialog box



Correlations

```
[DataSet1] D:\Books\Regression\Data\Tag Numbers.sav
```

Correlations

		Tag.Number	Age
Tag.Number	Pearson Correlation	1	.416
	Sig. (2-tailed)		.354
	N	7	7
Age	Pearson Correlation	.416	1
	Sig. (2-tailed)	.354	
	N	7	7

Figure 1.18 Correlation results in SPSS

This will always be the case as these values are always significant. The last number is the sample size, or seven in this case.

Some Correlation Examples

We will now look at some more correlation examples using the data sets discussed earlier.

Broilers

Figure 1.1 shows a set of data that was very strongly correlated. This chart shows the top 25 broiler-producing states for 2001 by both numbers and pounds. The data are shown in Table 1.1. The resulting correlation is 0.9970. As expected, this value is both strong and positive.

Tag Numbers and Sine Wave

Figure 1.2 shows the tag number example, which is discussed earlier. That correlation is 0.4158. Figure 1.3 shows the sine wave. As discussed, that correlation is –0.0424, or about zero.

Stock and Bond Returns

Table 1.3 shows the actual returns on stocks, bonds, and bills for the United States from 1928 to 2009. That dataset has four variables:

1. Year
2. Return on stocks
3. Return on treasury bonds
4. Return on treasury bills

Table 1.7 shows the correlation matrix for this data. Notice that none of the correlations is very high. The value of 0.4898 between "Year" and "Treasury bills" is the highest, whereas the correlation between "Stocks" and "Treasury bonds" is virtually zero.

Federal Employees and Salary

Table 1.4 shows a state-by-state breakdown of the number of federal employees and their average salary for 2007. Figure 1.5 shows this data to

Table 1.7 Correlation Matrix for Stock and Bond Returns

	Year	Stocks	Treasury bills	Treasury bonds
Year	1.0000			
Stocks	0.0212	1.0000		
Treasury bills	0.4898	-0.0189	1.0000	
Treasury bonds	0.2721	-0.0009	0.3146	1.000

Table 1.8 Correlation Matrix for Transit Ridership Including New York

	Unlinked passenger trips	Passenger miles	Population
Unlinked passenger trips	1.0000		
Passenger miles	0.9990	1.0000	
Population	0.8610	0.8633	1.0000

have a weak, positive correlation. This is supported by the resulting correlation value of 0.5350.

Transit Ridership

Table 1.5 shows the largest urbanized areas by population, unlinked passenger trips, and passenger miles for 2008. That dataset has three variables:

1. Unlinked passenger trips in thousands
2. Passenger miles in thousands
3. Population from the 2000 Census

Figure 1.6 shows that the data had an outlier in the values for New York. Its values in all three categories far exceed the values for any other transit system. The single outlier does not affect correlation analysis as much as it does the scatterplots. Table 1.8 shows the correlation matrix using all the data and Table 1.9 shows the correlation matrix while excluding New York. Notice that the correlations are all very strong, all very positive, and not very different with or without New York.

Table 1.9 Correlation Matrix for Transit Ridership Excluding New York

	Unlinked passenger trips	Passenger miles	Population
Unlinked passenger trips	1.0000		
Passenger miles	0.9885	1.0000	
Population	0.8726	0.8544	1.0000

Correlation Coefficient Hypothesis Testing[11]

All the aforementioned data discussed are sample data. The sample correlation coefficients *r* computed on the aforementioned data are just an estimate of the population parameter ρ. As with any other statistic, it makes sense to perform hypothesis testing on the sample value. While the mechanics of the hypothesis for the correlation coefficient are almost identical to the single variable hypothesis tests of means and proportions that you are likely familiar with, the logic behind the test is slightly different.

With hypothesis testing on the sample mean or sample proportion, the test is to see if the sample statistic is statistically different from some hypothesized value. For example, you might test the average weight of cans of peas coming off a production line to see if it is 16 ounces or not. With the correlation coefficient, the hypothesis testing is to see if a significant population linear correlation exists or not. Therefore, our hypotheses become

$$\begin{aligned} &H_0: \text{ The population correlation is not meaningful} \\ &H_1: \text{ The population correlation is meaningful} \end{aligned}$$

Since a nonzero value represents a meaningful correlation, we operationalize these hypotheses as follows:

$$\begin{aligned} H_0&: \rho = 0 \\ H_1&: \rho \neq 0 \end{aligned}$$

If we have reason to expect a positive or negative correlation, we can also perform a one-tailed version of this test.

In virtually all instances, we are testing a one-or two-tailed version of $\rho = 0$. The test we will use for this hypothesis is only good where the null hypothesis assumes a correlation of zero. In the rare case that you wish to test for a value other than zero, the Student *t*-distribution does not apply and the test discussed just after the next paragraph *cannot* be used. Readers needing to test values other than zero are urged to consult an advanced reference for the methodology.

Once the one-or two-tailed version of the hypotheses is selected, the critical value or values are found in the Student *t*-table or from an appropriate worksheet in the normal fashion. However, this test has $n - 2$ degrees of freedom rather than $n - 1$. The test statistic is as follows:

Correlation Coefficient Test Statistics

$$t_{n-2} = \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}}$$

Notice that the hypothesized value is not used in this equation. That is because it is always zero, and subtracting it would have no impact. Also notice that none of the column totals is used in the calculations. All you need is the sample correlation coefficient r and the sample size n.

An Example Using Tag Numbers

In the tag number example, the sample size is seven and the sample correlation is 0.4158. Since we have no reason to believe that tag numbers should be positively or negatively correlated with age, we will perform a two-tailed test—that is,

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0 \end{aligned}$$

With $n = 7$, we have five degrees of freedom ($n - 2$), giving us a critical value of ±2.5706. The test statistic calculates as the following:

Hypothesis Test for Tag Numbers

$$\begin{aligned} t_{n-2} &= \frac{r}{\sqrt{\frac{(1-r^2)}{(n-2)}}} \\ &= \frac{0.4158}{\sqrt{\frac{(1-0.4158^2)}{(7-2)}}} = \frac{0.4158}{\sqrt{\frac{0.8271}{5}}} = 1.0223 \end{aligned}$$

Since 1.0223 is less than the critical value of 2.5706, we accept that the null hypothesis is correct. Accepting the null hypothesis as correct means we conclude that the population correlation is not significantly different from zero. In other words, there is no evidence of a population correlation. Given the nature of the data, this is exactly what we would expect.

Steps to Hypothesis Testing

To summarize, the steps to hypothesis testing of the correlation coefficient are as follows:

1. Select the null and alternative hypothesis based on your belief that the correlation should or should not have a direction. You will always be selecting one of the three following sets of hypotheses:
    a. When you have no reason to believe the correlation will have a positive or negative value

    $$H_0: \rho = 0$$
    $$H_1: \rho \neq 0$$

    b. When you believe the variables will have a positive correlation

    $$H_0: \rho \leq 0$$
    $$H_1: \rho > 0$$

    c. When you believe the variables will have a negative correlation

    $$H_0: \rho \geq 0$$
    $$H_1: \rho < 0$$

2. Set the level of significance, also known as alpha. In business data analysis, this is almost always 0.05. For a more detailed discussion of alpha, consult any introductory statistics textbook.
3. Find the critical value based on the Student *t*-distribution and $n - 2$ degrees of freedom.
4. Compute the test statistic using the Correlation Coefficient Test Statistics formula.
5. Make a decision.
    a. When you have no reason to believe the correlation will have a positive or negative value, you accept the null hypothesis (that

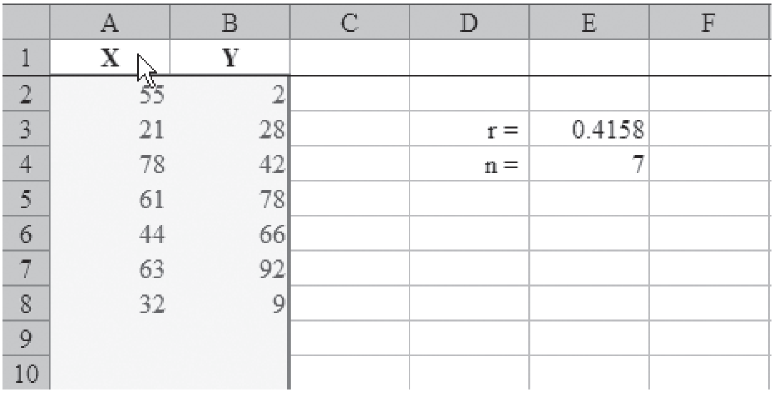
there is no correlation) when the test statistic is between the two values. You reject the null hypothesis (and conclude the correlation is significant) when the test statistic is greater than the positive value or less than the negative value.

b. When you believe the variables will have a positive correlation, you accept the null hypothesis when the test statistic is less than the positive critical value and reject the null hypothesis when the test statistic is greater than the positive critical value.
c. When you believe the variables will have a negative correlation, you accept the null hypothesis when the test statistic is greater than the negative critical value and reject the null hypothesis when the test statistic is less than the negative critical value.

The Excel Template

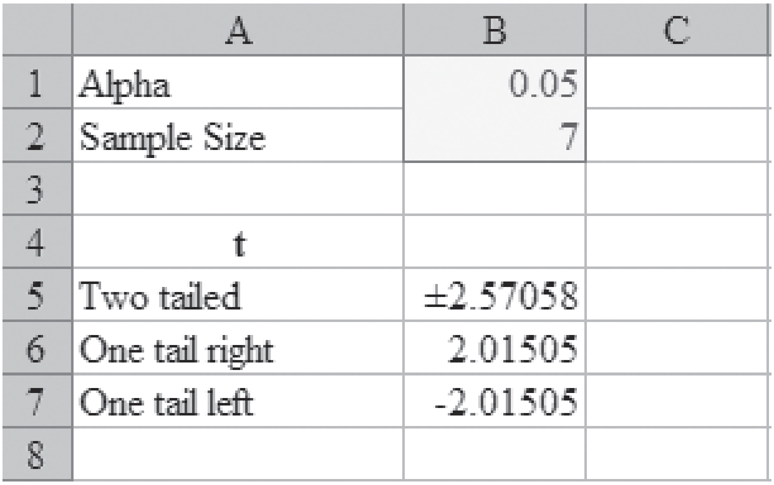
All these calculations can be automated using the *Correlate.XLS* worksheet. We will demonstrate it using the example of the tag data. The *Correlate* tab allows you to enter up to 100 pairs of values in cells A2 to B101. It then shows the correlation coefficient to four decimal points in cell E3 and the sample size in cell E4. This is shown in Figure 1.19.

The red square in the worksheet, shown in gray here, outlines the area where the data are to be entered. Only pairs of values are entered into the calculations and values outside of the red square (shown in gray) are not allowed. This is enforced by the worksheet. It is protected and no changes can be made outside the data entry area.



	A	B	C	D	E	F
1	X	Y				
2	55	2				
3	21	28		r =	0.4158	
4	78	42		n =	7	
5	61	78				
6	44	66				
7	63	92				
8	32	9				
9						
10						

Figure 1.19 Using the Excel template to perform correlation analysis



	A	B	C
1	Alpha	0.05	
2	Sample Size	7	
3			
4	t		
5	Two tailed	±2.57058	
6	One tail right	2.01505	
7	One tail left	-2.01505	
8			

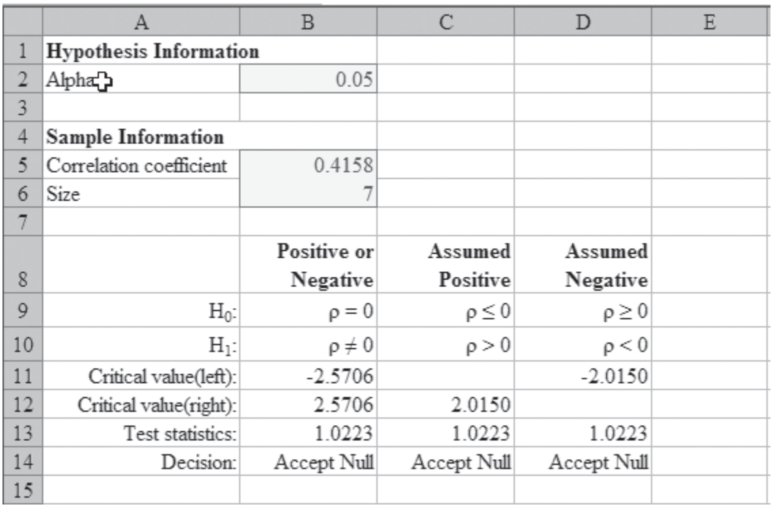
Figure 1.20 The critical values tab of the template

The *Critical Values* tab of the worksheet is shown in Figure 1.20. Since this hypothesis test is always performed using the Student *t*-distribution, those are the only values returned by this worksheet tab. Just as in the hypothesis testing template, these values are not really needed since the tab for hypothesis testing looks up the values automatically.

The *Hypothesis Test* tab of the worksheet is shown in Figure 1.21. This tab automates much of the hypothesis testing. You enter the alpha level in cell B2, the correlation coefficient in cell B5, and the sample size in cell B6. The tab then performs all the calculations. You then simply select the appropriate hypothesis. In this example, the two-tailed test returns the test statistic of 1.0223, as computed in the Hypothesis Test for Tag Numbers, and accepts the null hypothesis.

Using SPSS

Look back at Figure 1.18, which shows the correlation matrix for the tag data. This has everything you need to perform the hypothesis test. In Figure 1.18, the value below the correlation is the two-tailed significance level, or 0.354 in this case. This is also known as the *p*-value. For a two-tailed test, you accept the null hypothesis when the *p*-value is greater than alpha and reject the null hypothesis when the *p*-value is less than alpha. Since the



	A	B	C	D	E
1	**Hypothesis Information**				
2	Alpha	0.05			
3					
4	**Sample Information**				
5	Correlation coefficient	0.4158			
6	Size	7			
7					
8		**Positive or Negative**	**Assumed Positive**	**Assumed Negative**	
9	H_0:	$\rho = 0$	$\rho \leq 0$	$\rho \geq 0$	
10	H_1:	$\rho \neq 0$	$\rho > 0$	$\rho < 0$	
11	Critical value(left):	-2.5706		-2.0150	
12	Critical value(right):	2.5706	2.0150		
13	Test statistics:	1.0223	1.0223	1.0223	
14	Decision:	Accept Null	Accept Null	Accept Null	
15					

Figure 1.21 A template for automating hypothesis testing of correlation values

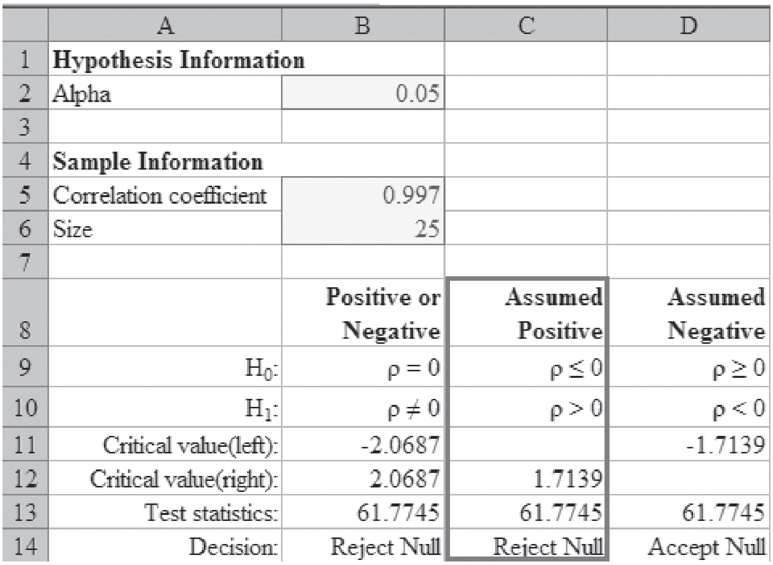
p-value of 0.354 is greater than our alpha value of 0.05, we accept the null hypothesis and again conclude that the correlation is insignificant.

The process is almost as easy for a one-tailed hypothesis test—that is, when you believe the correlation should be either positive or negative. In this case, the hypothesis test is a two-step process. First, you compare the sign of the correlation coefficient. If it does not match your expectations—that is, if your alternative hypothesis is that it is positive but the calculated value is negative—then you always accept the null hypothesis. Second, if the signs match, then you compare alpha and the *p*-value as in the first paragraph of this section, only you divide the *p*-value in half.

So if we believed the tag correlation should have been positive, then we would have passed the first part since the calculated correlation was indeed positive. Now we would compare 0.354 / 2 = 0.177 against the alpha value of 0.05. Since the new *p*-value of 0.117 is still larger than alpha, we would again accept the null hypothesis and conclude the correlation is insignificant.

Broilers

Figure 1.1 shows a set of data that is very strongly correlated. This chart shows the top 25 broiler-producing states for 2001 by both numbers and



	A	B	C	D
1	**Hypothesis Information**			
2	Alpha	0.05		
3				
4	**Sample Information**			
5	Correlation coefficient	0.997		
6	Size	25		
7				
8		**Positive or Negative**	**Assumed Positive**	**Assumed Negative**
9	H_0:	$\rho = 0$	$\rho \leq 0$	$\rho \geq 0$
10	H_1:	$\rho \neq 0$	$\rho > 0$	$\rho < 0$
11	Critical value(left):	-2.0687		-1.7139
12	Critical value(right):	2.0687	1.7139	
13	Test statistics:	61.7745	61.7745	61.7745
14	Decision:	Reject Null	Reject Null	Accept Null

Figure 1.22 Using the Excel template to test the broilers hypothesis

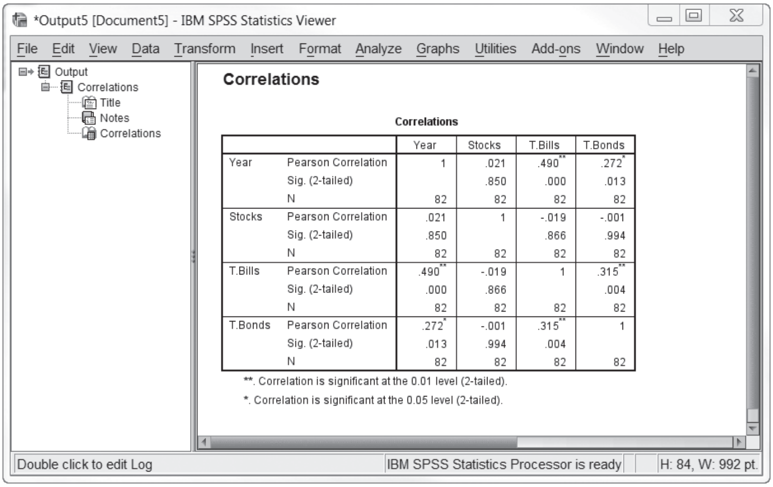
pounds. The data are shown in Table 1.1. The resulting correlation is 0.9970. Since more broilers should weigh more, we would expect a positive correlation. Figure 1.22 shows use of the Excel template to test the correlation coefficient for significances, and it is significant.

Stock and Bond Returns

Table 1.3 shows the actual returns on stocks, bonds, and bills for the United States from 1928 to 2009. Figure 1.23 shows the correlation matrix on these variables from SPSS. As you can see, it flags the combinations that are significant at the 0.05 level with one asterisk, as well as the higher 0.01 level with two asterisks. In this case, the following are significant:

1. "Treasury bills" with "Year"
2. "Treasury bonds" with "Year"
3. "Treasury bills" with "Treasury bonds"

Due to the large sample size, these are significant in spite of their relatively low correlation values. In general, the larger the sample size, the weaker the correlation can be and the correlation still be significant.



Correlations

Correlations

		Year	Stocks	T.Bills	T.Bonds
Year	Pearson Correlation	1	.021	.490**	.272*
	Sig. (2-tailed)		.850	.000	.013
	N	82	82	82	82
Stocks	Pearson Correlation	.021	1	-.019	-.001
	Sig. (2-tailed)	.850		.866	.994
	N	82	82	82	82
T.Bills	Pearson Correlation	.490**	-.019	1	.315**
	Sig. (2-tailed)	.000	.866		.004
	N	82	82	82	82
T.Bonds	Pearson Correlation	.272*	-.001	.315**	1
	Sig. (2-tailed)	.013	.994	.004	
	N	82	82	82	82

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Figure 1.23 SPSS correlation matrix for stock and bond data

Causality

Finding that a correlation coefficient is significant only shows that the two variables have a linear relationship. It does *not show* that changes in one variable cause changes in another variable. This is called *causality*, and showing causality is much more complex than just showing correlation. Consider the example in Box 1.1.

Box 1.1

Spelling and Shoe Size

If you walk into any elementary school in this nation and measure the students' spelling ability and shoe size (yes, shoe size), you will find a strong positive correlation. In fact, the correlation coefficient will be very close to +1 if you compute it separately for boys and girls! Does this mean that big feet cause you to be able to spell better? Can we scrap all the standardized tests that elementary school students take and just measure their feet? Or does it mean that being a good speller causes you to have bigger feet?

To make the matter even more confusing, if you walk into any high school in this nation and make the same measurement, you will find that the correlation coefficient is close to zero and, in fact, is insignificant. Can it be the case that having big feet only helps you in elementary school? Or is correlation analysis telling us something other than big feet cause good spelling?

Have you figured it out? In first grade, most students are poor spellers and have small feet. As they get older and move into higher grades, they learn to spell better and their feet grow. Thus the correlation between foot size and spelling ability really tells us that older elementary school students spell better than younger ones. In addition, since boys and girls grow at different rates, the correlation improves when each gender is computed separately. By high school, much of this effect has finished. Students no longer study spelling and are mostly reasonably competent spellers. Thus any differences in spelling ability are due to factors other than their age. Since age is no longer an indicator of spelling ability, a surrogate measure like foot size is no longer correlated with spelling. In addition, many students have completed the bulk of their growth by high school, so differences in feet are more an indication of the natural variation of foot size in the population than they are of age.

Simply stated, if we wish to show that A causes B, simply showing that A and B are correlated is not enough. However, if A and B are not correlated, that does show that A does not cause B—that is, the lack of correlation between foot size and spelling ability in high school is, by itself, enough to conclusively demonstrate that having larger feet does not cause a student to spell better.

Three things are required in order to show that A causes B:

1. A and B are correlated.
2. If A causes B, then A must come before B. This is called a *clear temporary sequence*.
3. There must be no possible explanation for the existence of B other than A.

Of these three items, only the first item—that A and B are correlated—is demonstrated using statistics. Thus it is never possible to demonstrate causality by just using statistics. Demonstrating the second and third items requires knowledge of the field being investigated. For this reason, they are not discussed in any detail in this textbook.

Think this spelling example is too esoteric to be meaningful? Think again. At many businesses, we can show that as advertising rises, sales go up. Does that mean that increases in advertising cause increases in sales? It could be, but businesses have more income when sales increase, and so they might simply elect to spend more of that income on advertising. In other words, does advertising → sales or does sales → income → advertising? Another example might help.

Now suppose that we have a new marketing campaign that we are testing, and we wish to show that the campaign causes sales of our product to rise. How might this be accomplished?

Showing that the two are correlated would involve computing the correlation of the level of expenditure on the new marketing campaign and our market share in the various regions. Showing a clear temporary sequence would involve looking at historical sales records to verify that the sales in the areas receiving the marketing campaign did not go up until after the marketing campaign had been started. In all likelihood, accomplishing these first two steps would not be too difficult, especially

Box 1.2

Ice Cream Sales

When ice cream sales are high, the number of automobile wrecks is also high. When ice cream sales are low, the number of automobile wrecks is lower. Does this mean that sales of ice cream cause automobile wrecks or that automobile wrecks drive the sale of ice cream?

Actually, it means neither. Just as income might drive advertising, a third variable influences both ice cream sales and automobile wrecks. In the summer, people drive more and so have more wrecks; they also buy more ice cream. In the winter, people drive less and so have fewer wrecks; they also buy less ice cream. Thus it is the season that is influencing both ice cream sales and automobile wrecks. Since they are both influenced by the same variable, they are correlated.

if the marketing campaign truly did cause additional sales. However, the third step might be more difficult.

In deciding if anything other than your new marketing campaign could explain the change in sales, you will need to look at the actions of your competitors, changes in demographics, changes in weather patterns, and much more. Of course, the specific items on this list would depend on the product being investigated. Now imagine how difficult it is to rule out all alternative explanations for more complex areas of study such as something causing cancer. Clearly, showing causality is not a simple undertaking. Fortunately, effective use of regression does not require showing causality. Likewise, using the results of regression either to understand relationships or to forecast future behavior of a variable does not require showing causality.

For another, more detailed discussion of the problems showing causality, see Box 1.3.

Box 1.3

Working Mothers Have Smarter Kids

A few years ago, a rash of television and newspaper reports focused on a research finding that stated that the children of working mothers had

- higher IQ scores,
- lower school absenteeism,
- higher grades,
- more self-reliance.

Any stay-at-home mother who saw these reports might reasonably conclude that the best thing she could do for her children is to put the kids in daycare and get a job!

The problem is the research was seriously flawed! But first, we will review how the research was conducted. Only by knowing how the research was conducted can you begin to see the flaws in that research.

Researchers selected 573 students in 38 states. The students were in first, third, and fifth grades. They divided these students into two groups: those with working mothers and those with stay-at-home mothers. On the measures of success used by the researchers, the first group did better.

Do you see the problem? The researchers made no attempt to figure out why the mothers in the second group were at home. Naturally, some of them were in families who were making the sacrifices necessary so the mother could be home with the kids. If those were the only ones in the second group, then it might make sense to conclude that the mother's staying at home did not improve the child's performance. However, this group of stay-at-home mothers included mothers who were not working for the following reasons:

- They were on welfare.
- They were too sick to work.
- They could not find a job.
- They simply did not want to work.
- They did not speak English.
- They were alcoholics or drug users who were unemployable.
- They were unemployable for other reasons.
- They were under 18 and too young to work.

It is likely that the poor performance from children from these groups was bad enough that it drove down the likely higher performance by children who had loving, concerned mothers who stayed home for their children.

Even if none of these factors were present and the data were completely valid, there is another equally likely explanation for the data: Families are more likely to make the sacrifice for the mother to stay home when the child is having problems. Thus the lower score for the kids of stay-at-home mothers could be due to the mothers' staying at home to help kids with problems rather than the facts that the moms are at home *causing* kids to have problems—that is, it could very well be that poor performance by the children caused their mother to stay at home rather than the mother staying at home causing the poor performance.

This study makes a point that every researcher and every consumer of research should always keep in mind: A statistical relationship—even a strong statistical relationship—does not imply that one thing caused another thing. It also makes another very important point if you wish to be an educated consumer of statistics: It is not enough to know the results of the statistical analysis; in order to truly understand the topic, you must know how the data were collected and what those data collection methods imply.

Summary

In this chapter the topic of correlation was introduced. Beginning with a scatterplot, we considered how two variables could be correlated. We also considered the relationship between causality and correlation.

We saw that correlation measured if there was a significant relationship between a pair of variables. In the next chapter, we will see how to use simple regression to mathematically describe that relationship.

CHAPTER 2
Simple Regression

In the last chapter, we looked at correlation analysis where we measured the linear relationship between pairs of variables. In this chapter we will use simple regression to develop a model describing the relationship between a single dependent variable and a single independent variable. This type of model is not frequently used in business, but understanding simple regression is a good way to begin our discussion of multiple regression.

Recall from high school geometry that the equation for a straight line can be written as follows:

Straight Line

$$Y = mX + b$$

In this equation, *m* is the slope of the line and *b* is the Y-intercept, or just intercept for short. Simple regression is a process for fitting a line to a set of points so the result of that process will be a value for the slope and intercept for that line. Regression uses different symbols from high school geometry for the slope and intercept and writes the intercept first, giving the following equation:

Simple Regression Equation

$$Y = \beta_0 + \beta_1 X$$

where the β_0 represents the intercept and the β_1 represents the slope. These are population symbols; the sample symbols are b_0 and b_1, respectively.

The assumptions of regression are discussed in full in the next chapter, but one of the assumptions of regression is that the data set consists of a random sample of pairs of X and Y variables from a population of all possible pairs of values. Because any sampling involves error, an error term is often added to the equation:

Simple Regression Equation With Error Term

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Although ε simply stands for the error, it cannot be estimated. Also note that it is common to refer to the error term as the *residual.*

Naturally, we wish to estimate the regression equation using sample data, which is written in the following equation:

Sample Simple Regression Equation With Error Term

$$Y = b_0 + b_1 X + e$$

where b_0 estimates β_0, b_1 estimates β_1, and e represents the observed error—the leftover, or residual—from fitting the regression line to a specific set of data. This equation can be written with the subscript "*i*" to represent the specific data points:

Sample Simple Regression Equation for Specific Data Points

$$Y = b_0 + b_1 X_i + e_i$$

where i goes from 1 to n and e_1 is the distance between the line and the first observed point, e_2 is the distance between the line and the second observed point, and so on. When used to estimate values, the equation is written as follows:

Sample Simple Regression Equation for Estimates

$$\hat{Y} = b_0 + b_1 X_1$$

Here, $\hat{Y}$ (pronounced "y-hat") is the value of the dependent variable Y that lies on the fitted regression line at point X. That is, $\hat{Y}_1$ is the result of the equation for $X_{1,1}$, $\hat{Y}_2$ is the value for $X_{1,2}$, and so on.[1]

Figure 2.1 shows a scatterplot for two variables. It also shows three different lines that might be drawn to fit the data. Two have a positive slope and appear to be almost parallel. The other has a negative slope. It is

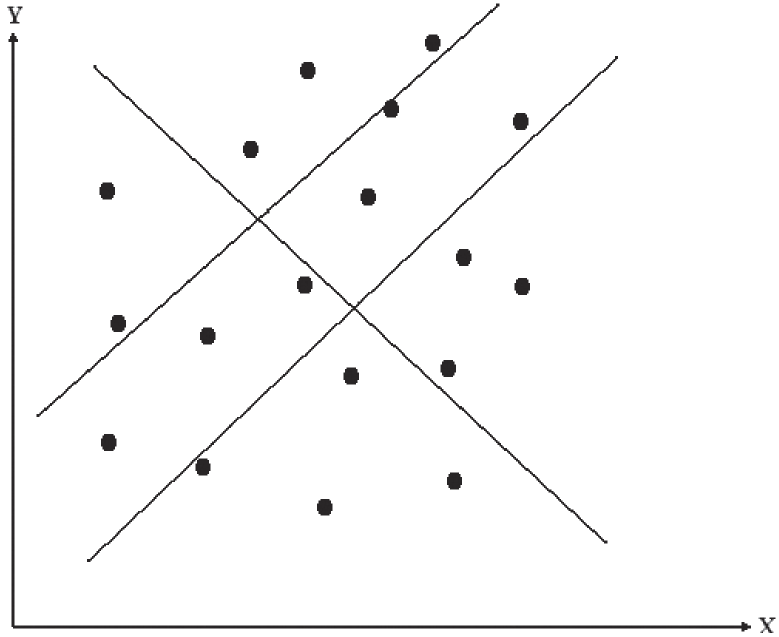


Figure 2.1 More than one line can be used to fit this data

the job of regression to examine all possible lines and to choose the single line that *best fits* the data.

The single line that best fits the data has specific criteria. To illustrate this, Figure 2.2 shows a simplified set of data points with a single line under consideration. For each point, the *vertical* distance between each point and the line under consideration is measured. Distances that go up are positive distances, whereas distances that go down are negative. To avoid having the positive and negative distances cancel out, the distances are squared. This makes them all positive and removes any potential for cancellation. The resulting distances are added. This total is called a *sum of squares*. The line with the smallest sum of squares is the line selected. This procedure is called *least squares regression*.

Naturally, the techniques we will be using do not require you to actually test every possible line. After all, there are an infinite number of potential lines to consider. The mathematical development of the formulas used to calculate the regression coefficients guarantees that the sum of squares will be minimized. However, it is nice to know the background of the formulas.

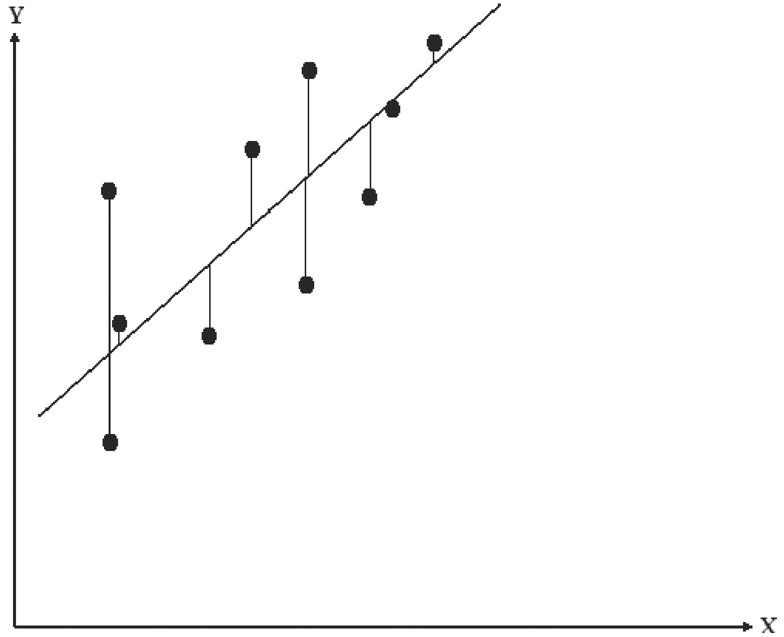


Figure 2.2 Measuring the vertical distance between the point and the line under consideration

We begin by noting that the sums of squares we just mentioned are SS_{errors} (or *SSE*) because they represent the mistake, or error, in the estimate of the line. The following is the formula for *SSE*:

Sum of Squared Errors

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Using calculus, we can then take the partial derivatives of *SSE* with respect to b_0 and b_1 and, because we wish to minimize *SSE*, we set them equal to zero. This yields the normal equations

Normal Regression Equations

$$\sum_{i=1}^{n} Y_i = n \cdot b_0 + b_1 \cdot \sum_{i=1}^{n} X_i$$

and

$$\sum_{i=1}^{n} X_i Y_i = b_0 \cdot \sum_{i=1}^{n} X_i + b_1 \cdot \sum_{i=1}^{n} X_i^2$$

Solving for b_0 and b_1, and rewriting the equations, we obtain the following equations:

Regression Coefficients[2]

Intercept

$$b_0 = \frac{\left(\sum Y\right)\left(\sum X^2\right) - \left(\sum X\right)\left(\sum X \cdot Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

Slope

$$b_1 = \frac{n\left(\sum X \cdot Y\right) - \left(\sum X\right) \cdot \left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

A few notes are in order regarding these formulas:

- As with the correlation coefficient, these formulas only compute sample estimates of population parameters.
- Unlike the correlation coefficient, b_0 and b_1 can take on any value between negative infinity ($-\infty$) and positive infinity ($+\infty$).
- It is important not to read too much into the relative magnitudes of these coefficients. The magnitude is a function of the units used to measure the data. Measure sales in dollars and the coefficients will have one magnitude, measure those same sales in millions of dollars and the coefficients will have a very different magnitude.
- As with the correlation coefficient, except for n, only totals are used in this formula. As a result, the calculations will be very similar to the calculations of the correlation coefficient.
- These are point estimates of β_0 and β_1 and these estimates have variances. This, and the assumption of normalcy, allows us to develop confidence interval estimates and to perform hypothesis testing on them.
- This formula always results in the line going through the point $(\overline{X}, \overline{Y})$.

Normally, you would never perform regression on a data set with an insignificant correlation coefficient. Correlation tests to see if a linear relationship exists and then regression quantifies that relationship. If the hypothesis test of the correlation coefficient indicates that there is no correlation, then there is no correlation for regression to quantify. Nevertheless, we will continue with one of the examples described in chapter 1 because the calculations are fairly straightforward and the sample size was small. Additionally, we will show how to perform the calculations by hand, although these are easily performed using Excel or a statistics package, so there would rarely be an occasion to perform these hand calculations.

Age and Tag Numbers

In Table 1.2, we showed the ages of seven people and the last two digits of their tag numbers. A chart of this data was shown in Figure 1.2. Table 1.6 (repeated here) gave us the data needed to compute the correlation coefficient of 0.4157. Hypothesis testing would then show this correlation coefficient to be insignificant. Table 1.6 also gave us the data we need to compute the slope and intercept of the regression equation. Those regression calculations are shown here as well. Unlike correlation, it does matter which variables are treated as the dependent and independent variables. In this case, it does not make intuitive sense to say that age influences tag numbers or that tag numbers influence age so we will use Age as the independent (*X*) variable and Tag Number as the dependent (*Y*) variable.

Table 1.6 Correlation Coefficient Calculations

Age	Tag no.	Age · Tag	Age^2	Tag no.^{2}
55	2	110	3,025	4
21	28	588	441	784
78	42	3,276	6,084	1,764
61	78	4,758	3,721	6,084
44	66	2,904	1,936	4,356
63	92	5,796	3,969	8,464
32	9	288	1,024	81
354	317	17,720	20,200	21,537

Computing the Intercept

$$b_0 = \frac{\left(\sum Y\right)\left(\sum X^2\right) - \left(\sum X\right)\left(\sum X \cdot Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$b_0 = \frac{(317)(20{,}200) - (354)(17{,}720)}{7(20{,}200) - (354)^2}$$

$$b_0 = \frac{6{,}403{,}400 - 6{,}272{,}880}{141{,}400 - 125{,}316} = \frac{130{,}520}{16{,}084} = 8.1149$$

Computing the Slope

$$b_1 = \frac{n\left(\sum X \cdot Y\right) - \left(\sum X\right) \cdot \left(\sum Y\right)}{n\left(\sum X^2\right) - \left(\sum X\right)^2}$$

$$b_1 = \frac{7(17{,}720) - (354)(317)}{7(20{,}200) - (354)^2}$$

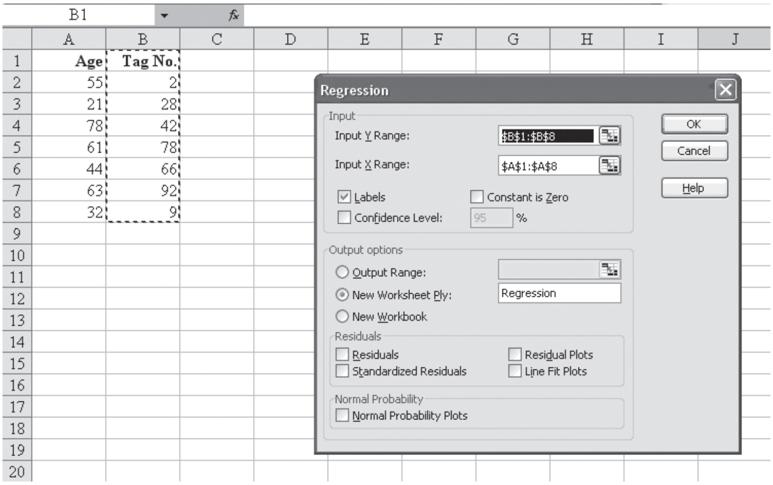
$$b_1 = \frac{124{,}040 - 112{,}218}{141{,}400 - 125{,}316} = \frac{11{,}822}{16{,}084} = 0.7350$$

So the regression equation is given by the line $\hat{Y} = 8.1149 + 0.7350X$.

Using Excel

For the dynamic approach, Excel offers matrix functions that can be used to calculate the regression coefficients and a few other pieces of information, but the information reported by Excel using the static approach is so much more extensive that it makes little sense to calculate regression any other way in Excel. We will use the previous example to illustrate how to use Excel to perform regression calculations.

Regression is performed using the *Analysis Toolpak*. You may need to install this, as described in chapter 1. After doing this, the first step is to load the data file containing the data on which you wish to perform regression, TagNumber.xls in this case. The data will have to be entered in column format. Whereas the data in TagNumber.xls is side by side, it is not necessary to have the dependent variable in a column next to the independent variables, although in practice it is a good idea. In the next chapter, we will be working with multiple independent variables, and Excel does require that all the independent variables be located side by side.



	A	B
1	Age	Tag No.
2	55	2
3	21	28
4	78	42
5	61	78
6	44	66
7	63	92
8	32	9

Regression

Input
Input Y Range: $B$1:$B$8
Input X Range: $A$1:$A$8
☑ Labels □ Constant is Zero
□ Confidence Level: 95 %

Output options
○ Output Range:
● New Worksheet Ply: Regression
○ New Workbook

Residuals
□ Residuals □ Residual Plots
□ Standardized Residuals □ Line Fit Plots

Normal Probability
□ Normal Probability Plots

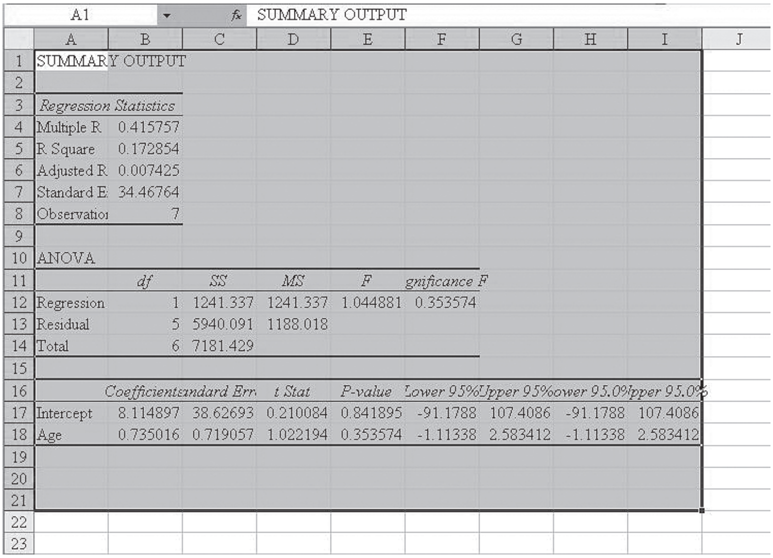
Figure 2.3 The Regression dialog box

Once the worksheet is loaded, you click on the Data tab and then *Data Analysis*. This brings up the *Data Analysis* dialog box shown back in Figure 1.15. This time, select *Regression* from this list. This brings up the *Regression* dialog box shown in Figure 2.3. You use this dialog box to feed the data into Excel, set options, and control what output you get and where the output is placed.

The *Input Y Range* is the range of cells containing the dependent variable, or B1 to B8 in this case. We have included this label, so we will have to include the label for the other variable and we will have to check the *Labels* box. Including labels is a good idea as it makes the printout easier to read. There is no need to remember the range for the data; you can click on the arrow to the right side of the input box and highlight the range manually. The *Input X Range* is the range of cells containing the independent variable, or A1 to A8 in this case.

You must tell Excel where to put the output. Regression generates a lot of output, so it is always best to use a separate worksheet page for the results. In this example, we have given that new page the name "Regression," although you can give it any valid name.

Once everything is entered correctly in the *Regression* dialog box, you click on *OK* to run the regression. Excel performs the calculations and places the results in the new worksheet page, as specified. Those results are shown in Figure 2.4. As you can see, the results are not all that readable. Some of the labels and numbers are large and Excel does not automatically

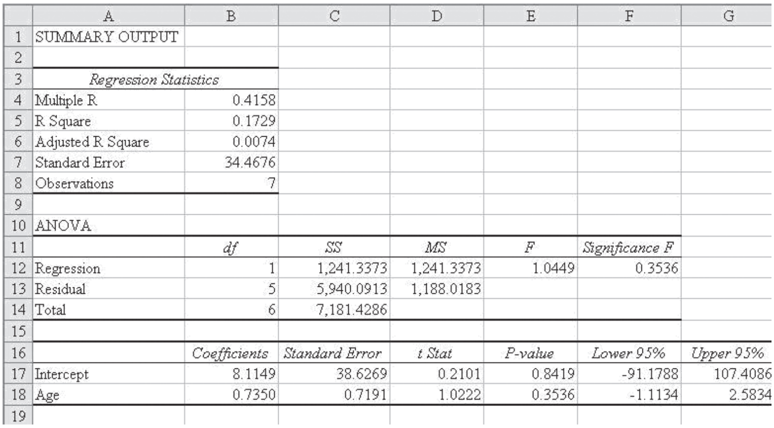


A1 fx SUMMARY OUTPUT

	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.415757								
5	R Square	0.172854								
6	Adjusted R	0.007425								
7	Standard E	34.46764								
8	Observatio	7								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*gnificance F*				
12	Regression	1	1241.337	1241.337	1.044881	0.353574				
13	Residual	5	5940.091	1188.018						
14	Total	6	7181.429							
15										
16		*Coefficient*	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*lpper 95.0%*	
17	Intercept	8.114897	38.62693	0.210084	0.841895	-91.1788	107.4086	-91.1788	107.4086	
18	Age	0.735016	0.719057	1.022194	0.353574	-1.11338	2.583412	-1.11338	2.583412	
19										
20										
21										
22										
23										

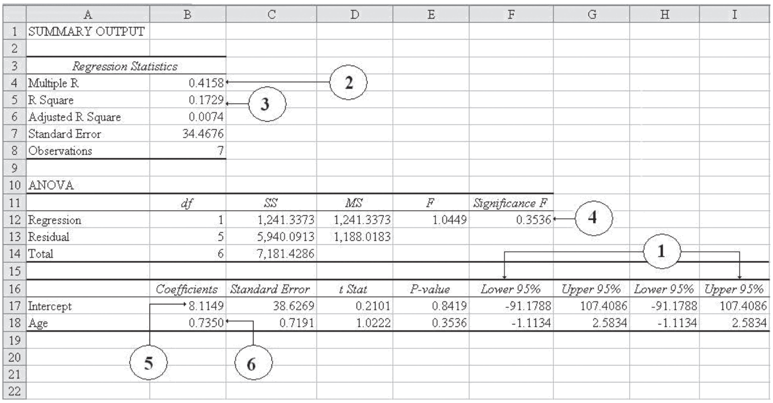
Figure 2.4 Initially, the results of an Excel regression run are jumbled together

change the column width to accommodate this wider information. In addition, Excel does not format the numbers in a readable format. While the data are still highlighted, we can adjust the column widths by clicking on the *Home* tab, in the *Cells* group clicking on *Format*, and under *Cell Size*, clicking on *AutoFit Column Width*. You can also format the numbers to a reasonable number of decimal points. Those results are shown in Figure 2.5. Of course, adjusting the column widths would affect any other



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.4158					
5	R Square	0.1729					
6	Adjusted R Square	0.0074					
7	Standard Error	34.4676					
8	Observations	7					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	1,241.3373	1,241.3373	1.0449	0.3536	
13	Residual	5	5,940.0913	1,188.0183			
14	Total	6	7,181.4286				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	8.1149	38.6269	0.2101	0.8419	-91.1788	107.4086
18	Age	0.7350	0.7191	1.0222	0.3536	-1.1134	2.5834
19							

Figure 2.5 The results of an Excel regression run after some formatting



	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.4158							
5	R Square	0.1729							
6	Adjusted R Square	0.0074							
7	Standard Error	34.4676							
8	Observations	7							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	1,241.3373	1,241.3373	1.0449	0.3536			
13	Residual	5	5,940.0913	1,188.0183					
14	Total	6	7,181.4286						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95%*	*Upper 95%*
17	Intercept	8.1149	38.6269	0.2101	0.8419	-91.1788	107.4086	-91.1788	107.4086
18	Age	0.7350	0.7191	1.0222	0.3536	-1.1134	2.5834	-1.1134	2.5834
19									
20									
21									
22									

Figure 2.6 Reading an Excel printout

data that might be included on this worksheet page. This is yet another reason for placing the regression output on a new page.

Reading an Excel Simple Regression Printout

Figure 2.6 shows Figure 2.5 with reference numbers added. We will be referring to these reference numbers in this discussion. The reference numbers do not, of course, show up in actual Excel results.

The following list explains each of the numbered captions shown in Figure 2.6.

1. Excel shows a 95 percent confidence interval for each coefficient (b_0 and b_1). We will see how to compute these later in this chapter. For now, notice that each interval is given twice. This is somewhat of a bug in Excel. The beginning dialog box allows you to select any given confidence interval you like, and Excel will display that level *along with* the 95 percent level. When you leave the confidence level at the default value of 95 percent, Excel does not compensate and shows the interval only once. For the remainder of this book, we will not show this duplicate set of values, as we usually delete these two extra columns from our worksheets.
2. In simple regression, the *multiple r* is the same as the correlation coefficient. This will *not* be the case with multiple regression.

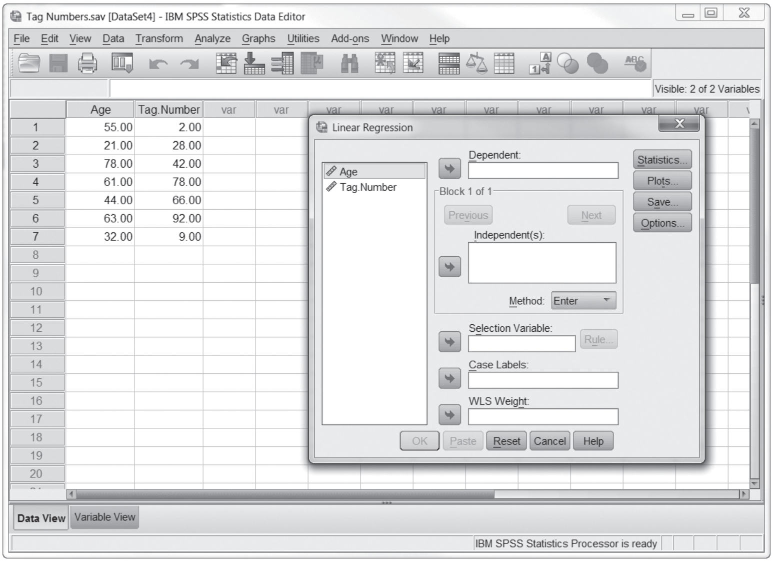
3. *R squared* is the *r* value squared. This is true in both simple and multiple regression. R squared has a very specific meaning. It is the percentage of the variation in the dependent variable that is explained by variations in the independent variables. So in this case, variations in the ages of the respondents in the data set explained only 17.3 percent of the variation in the tag numbers. As expected, that is not a very good showing. Because there is really no relationship between these two variables, even this small value is only due to the small sample size and sampling error.
4. Significant *F* is the *p*-value for testing the overall significance of the model. In simple regression, this will always yield the same results as a two-tailed significance test on the correlation coefficient, so it can be ignored in simple regression. (If the correlation coefficient is significant, then the overall model is significant. If the correlation coefficient is not significant, then the overall model is not significant.) Whereas this can be ignored in simple regression, it will become a very important measure in multiple regression.
5. This is the intercept coefficient.
6. This is the slope coefficient.

More of the values shown on an Excel printout will be discussed later in this chapter and in chapter 3.

Using SPSS

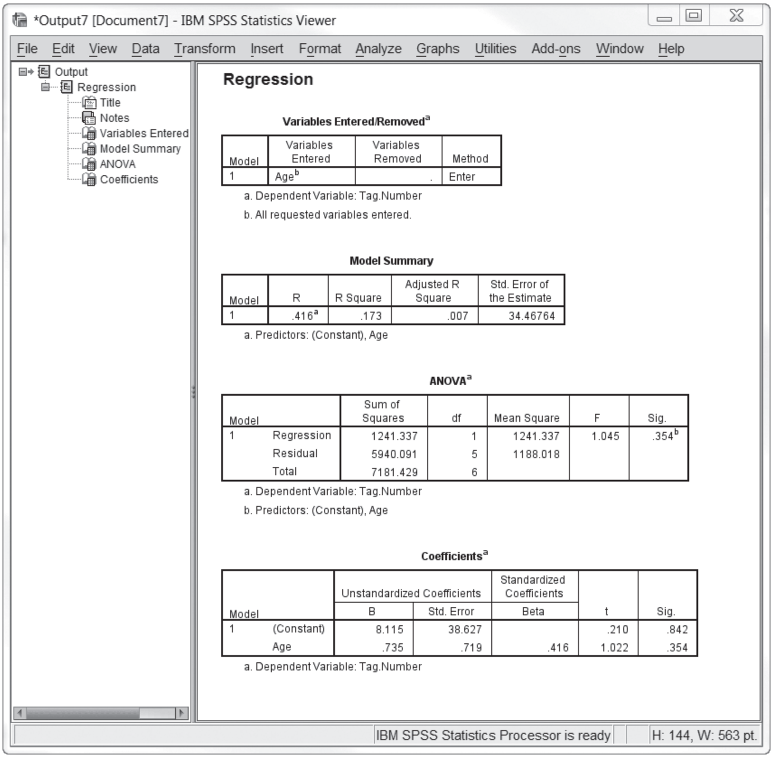
We saw the car tag data in SPSS back in Figure 1.11. To perform simple regression, you click on *Analyze*, *Regression*, and then *Linear*. That brings up the dialog box shown in Figure 2.7. From here, you click on the age variable and the arrow to move it into the *Independent(s)* box and you click on the tag number variable and the arrow to move it into the *Dependent* box. Once you have done this, the *OK* button will change from gray to black and you click on it to run the simple regression. None of the other options needs to be set for this scenario.

Figure 2.8 shows the results of running simple regression on the car tag data in SPSS. The left side of the screen is used to navigate between



	Age	Tag.Number
1	55.00	2.00
2	21.00	28.00
3	78.00	42.00
4	61.00	78.00
5	44.00	66.00
6	63.00	92.00
7	32.00	9.00

Figure 2.7 The Simple Regression dialog box in SPSS


*Output7 [Document7] - IBM SPSS Statistics Viewer
File Edit View Data Transform Insert Format Analyze Graphs Utilities Add-ons Window Help
Output
Regression
Title
Notes
Variables Entered
Model Summary
ANOVA
Coefficients
IBM SPSS Statistics Processor is ready
H: 144, W: 563 pt.

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Age[b]	.	Enter

a. Dependent Variable: Tag.Number

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.416[a]	.173	.007	34.46764

a. Predictors: (Constant), Age

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1241.337	1	1241.337	1.045	.354[b]
	Residual	5940.091	5	1188.018		
	Total	7181.429	6			

a. Dependent Variable: Tag.Number

b. Predictors: (Constant), Age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	8.115	38.627		.210	.842
	Age	.735	.719	.416	1.022	.354

a. Dependent Variable: Tag.Number

Figure 2.8 The SPSS output of running a simple regression

sections of the results. While it is not useful here, it can be very useful when working with large models or multiple scenarios. To move to a new section of the results, just click on that section.

Currently, the *Title* is "Regression," but you can double-click on it and change it to something meaningful like "Car Tag Simple Regression." This simple regression run has no *Notes*. The *Variables Entered/Removed* is not meaningful here but is useful in complex model building. The *Model Summary* area gives us the measures of model quality, r, r^2, adjusted r^2, and the standard error of the estimate. Both r and r^2 have already been discussed. Adjusted r^2 and the standard error of the estimate will be discussed later in this chapter.

The *ANOVA*[3] section is next and measures the statistical significance of the model. The *Coefficients* section gives the slope and intercept for the model along with measures of their statistical significance.

While Excel and SPSS present the results very differently, they both present the same results, at least within rounding differences. That is to be expected. Both tools do an excellent job of simplifying the calculation of regression, both simple regression as we are calculating here and multiple regression as we will calculate in the next chapter. However, what we will find is that when the models get to be more complex, a statistical package like SPSS has some very real advantages over a general-purpose tool like Excel.

More on the Regression Equation

Look back at the regression equation for the car tag example:

$$\hat{Y} = 8.1149 + 0.7350X.$$

What exactly does this mean? We can see this visually represented in Figure 2.9. In this chart, the original data points are shown as dots with the regression overlaid as a line. Notice that the line slopes up. This is expected from the positive slope of 0.7350. Notice that the regression line crosses the Y-axis just above the X-axis. This, too, is to be expected from the Y-intercept value of 8.1149. Finally, notice how the points are spread out widely and are not close to the line at all. This behavior indicates a low r^2, 0.1729 in this case. For a higher r^2, the points would be less scattered.

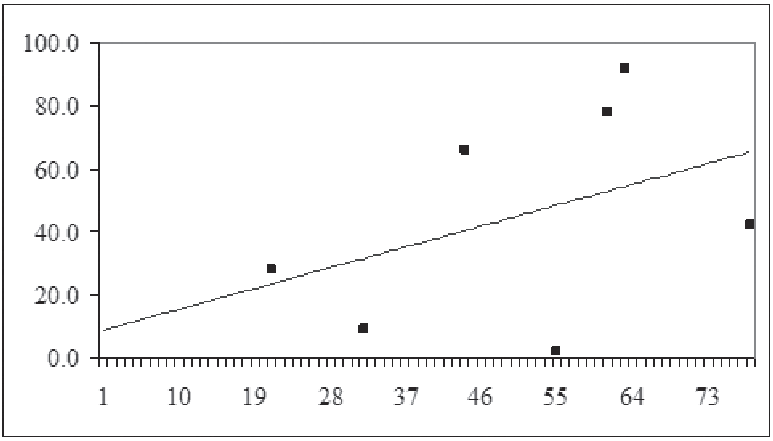


Figure 2.9 A chart of the car tag data and the resulting line

It is important not to read too much into the magnitude of the slope or the Y-intercept, for that matter. Their values are a function of the units used to measure the various variables. Had we measured the ages in months or used the last three digits of the car tags, the magnitude of the slope and Y-intercept would be very different. Furthermore, the r and r^2 would remain the same. Because multiplying one or both variables by a fixed number is a linear transformation, the linear relationship between the variables remains the same. Therefore, the measures of that relationship do not change. Finally, the standard error and ANOVA numbers change because they are measured in the units of the variables. However, the F-value and p-value of ANOVA do not change as they, too, are unaffected by linear transformations.

Federal Civilian Workforce Statistics

We need to explore the simple regression results in more detail, but the car tag example is a poor choice because the results are insignificant. We have only used it so far because the limited number of observations makes it easy to understand and even calculate some of the numbers by hand.

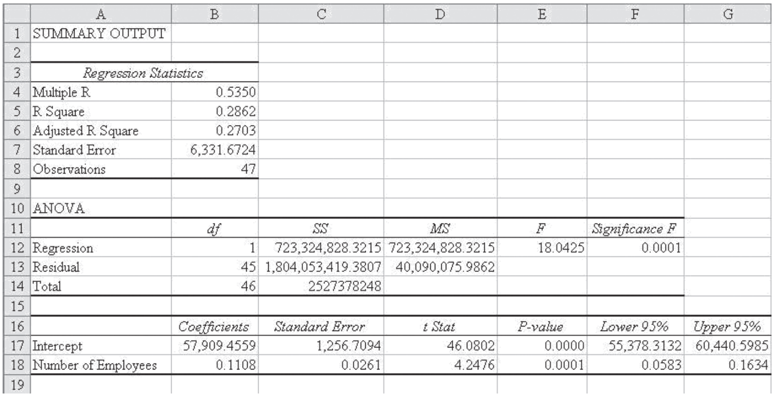
Table 1.4 showed a state-by-state breakdown of the number of federal employees and their average salary for 2007. In chapter 1, we computed the r-value as 0.5350. Whereas that value is fairly low, in testing we found that the correlation was statistically significant. That is important. When the correlation is significant, the simple regression will also be significant. Likewise, when the correlation is insignificant, the simple regression will be insignificant. Note, however, that this does not hold in multiple regression.

We immediately have a problem with this data. For correlation, we did not have to worry about which variable was the independent variable and which was the dependent variable, but it does matter for simple regression. Does salary depend on the number of workers or does the number of workers depend on the salary? Both relationships make sense. A higher salary would attract more workers, whereas having a large number of workers might drive down salaries. For this analysis, we will *assume* the number of workers is the independent variable.

In your own work, you are unlikely to have this problem. You will define the dependent variable you are looking to explain and then search for one or more independent variables that are likely to help explain that already-defined dependent variable.

The Results

Figure 2.10 shows the results of running a simple regression on the federal civilian workforce data with the number of workers as the independent variable. We will look at what many of the numbers in this printout mean:



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.5350					
5	R Square	0.2862					
6	Adjusted R Square	0.2703					
7	Standard Error	6,331.6724					
8	Observations	47					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	723,324,828.3215	723,324,828.3215	18.0425	0.0001	
13	Residual	45	1,804,053,419.3807	40,090,075.9862			
14	Total	46	2527378248				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	57,909.4559	1,256.7094	46.0802	0.0000	55,378.3132	60,440.5985
18	Number of Employees	0.1108	0.0261	4.2476	0.0001	0.0583	0.1634
19							

Figure 2.10 ***The Excel simple regression results on the federal civilian workforce data with the number of workers as the independent variable***

- *Multiple R*. In simple regression, this is the same as the correlation coefficient. It goes from –1 to +1 and measures the strength of the relationship. The closer the value is to –1 or +1, the stronger the relationship, and the closer it is to 0, the weaker the relationship. This relationship is weak.
- *R squared*. In one respect, this is simply the multiple r^2. It goes from 0 to 1. The closer it is to 1, the stronger the relationship, and the closer it is to 0, the weaker the relationship. However, it is also the percentage of the variation in the dependent variable explained by the independent variable. We will see the reason for this later.
- *Adjusted R squared*. This is explained in more detail in chapter 3. Adjusted r^2 is not really an issue for simple regression. With multiple regression, r^2 goes up when you add new variables even if those variables do not help explain the dependent variable. Adjusted r^2 adjusts for this issue so models with different numbers of variables can be compared.
- *Standard error*. This is short for the standard error of Y given X. It measures the variability in the predictions made based on resulting regression model.
- *Observations*. This is simply a count of the number of pairs of observations used in the simple regression calculations.
- *ANOVA*. Most of these values are beyond the scope of this chapter and will not be discussed. Some of these values will be briefly discussed in chapter 3.
- *Significant F*. This is the one critical piece of information in the ANOVA table that we need to discuss. The *Significant F* expresses the significances of the overall model as a *p*-value. Stated very simply, when this value is below 0.05, the overall model is significant.[4] Likewise, when this value is above 0.05, the overall model is insignificant. This is not important for simple regression because the significance of the model mirrors the significance of the correlation coefficient, but that relationship will not hold in multiple regression.
- *Coefficient*. This gives the values for the intercept and slope, or 57,909.4559 and 0.1108, respectively, in this model.

- *Standard error*. This standard error is the standard error associated with either the intercept or the slope.
- *t-stat*. This is the calculated *t*-statistic used to test to see if the intercept and slope are significant.
- *p-value*. When this value is less than 0.05, the corresponding slope or intercept is significant, and when it is greater than 0.05, they are insignificant. As a general rule, we do not test the intercept for significance as it is just an extension of the regression line to the Y-intercept. In simple regression, if the model is significant, the slope will be significant, and if the model is insignificant, the slope will be insignificant.
- *Lower 95 percent and upper 95 percent*. This is a 95 percent confidence interval on the intercept and slope. It is calculated as the coefficient value ±1.96 times the standard error of that value.

Interpretation

So what do these values tell us? The *Significant F* value of 0.0001 tells us the overall model is significant at $\alpha = 0.05$. The r^2 of 0.2862 tells us variation in number of employees explains less than 29 percent of the variation in salary, a very poor showing. The slope of 0.1108 tells us that for every one unit increase in the number of employees, the average salary goes up by 11 cents.

You might be tempted to say that the model is insignificant simply because the slope is only 11 cents; that would be a mistake. When there is little variation in the dependent variable, even a very strong model will have a relatively small slope. Likewise, when there is a large amount of variation in the dependent variable, even a poor model can have a relatively large slope. As a result, you can never judge the strength of the model based on the magnitude of the slope. Additionally, the units used to measure the data will directly affect the magnitude of the slope.

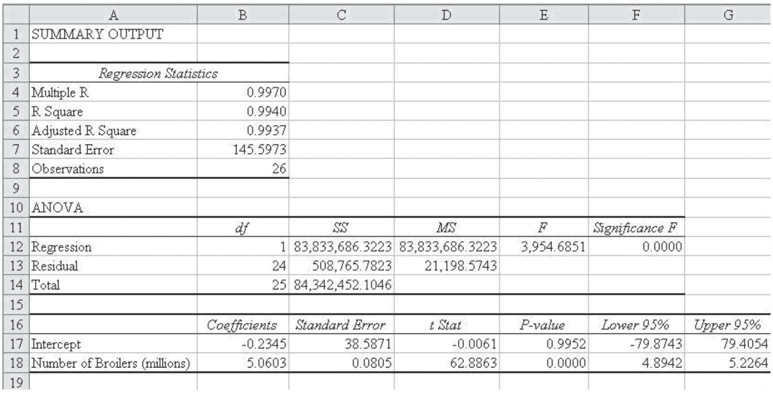
Why does this model do such a poor job? One possible explanation is simply that the number of employees has little or even no impact on salaries, and what correlation we are seeing is being driven by something else. In this case, it is very likely that the cost of living in the individual states is what is driving the salaries, and the size of states is driving the number of employees.

Number of Broilers

Now on to a realistic business example. Whereas realistic simple regression examples in business are few, this next example is actual business data where simple regression works extremely well. Figure 1.1 showed the top 25 broiler-chicken producing states for 2001 by both numbers and pounds, according to the National Chicken Council. The underlying data was shown in Table 1.1. When we explored the correlation in chapter 1, it was very strong at 0.9970. That makes sense: The more broilers a state produces, the higher the weight of those broilers should be. Additionally, broilers likely weigh about the same state to state, so this relationship should be very strong.

Figure 2.11 shows the resulting simple regression. Number of broilers (in millions) is the independent variable and pounds liveweight (in millions) is the dependent variable. The Significant *F* value of 0.0000 tells us the model is significant. The $r^2 = 0.9940$ tells us that variation in the independent variable explains over 99 percent of the variation in the dependent variable.

The intercept is –0.2345 or almost 0. The intercept does not always make sense because many times it is nothing more than an extension of the regression line to a Y-axis that may be far away from the actual data. However, in this case you would expect 0 broilers to weigh 0 pounds, so an intercept very near 0 makes perfect sense. The slope of 5.0603 means



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9970					
5	R Square	0.9940					
6	Adjusted R Square	0.9937					
7	Standard Error	145.5973					
8	Observations	26					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	83,833,686.3223	83,833,686.3223	3,954.6851	0.0000	
13	Residual	24	508,765.7823	21,198.5743			
14	Total	25	84,342,452.1046				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-0.2345	38.5871	-0.0061	0.9952	-79.8743	79.4054
18	Number of Broilers (millions)	5.0603	0.0805	62.8863	0.0000	4.8942	5.2264
19							

Figure 2.11 Simple regression on the data for number of broiler chickens

that an increase of 1 million broilers will increase the pounds liveweight by 5.0603 million. In other words, 1 broiler chicken weighs, on average, about 5 pounds—again exactly what you would expect.

Exploring the Broiler Model Further

This is a great model. The model is significant, it explains most of the variation in the dependent variable, and all the coefficients make perfect sense theoretically. You could not ask anything more of the model. That makes this model a good foundation for some additional discussions. Some of the material in this section is based on confidence intervals. You may want to review material on confidence intervals from a prior statistics course before continuing.

The slope in this model is 5.0603. This is b_1, a sample statistic and an estimate of the population parameter β_1. That is, we estimate the population slope to be 5.0603 based on this sample data. Had a different sample been used—say, a selection of different states or the same data from another year—then our estimate of the population slope would be different. But how different would it have been? A confidence interval can give us an indication. Recall that you calculate a 95 percent confidence interval using the following formula:

Formula for a Confidence Interval on the Slope

$$b_1 \pm t_{0.05,n-2} \cdot S_{\overline{b_1}}$$

The b_1 is, of course, the sample value for the slope, or 5.0603 in this example. The t is the Student t-value with $\alpha = 0.05$ and $n - 2$ degrees of freedom. Because $n = 26$ in this example, the degrees of freedom are $26 - 2 = 24$, giving us a t-value of 2.0639. The s-value is the standard error, which the printout tells us is 0.0805. The confidence interval is calculated as follows:

Confidence Interval Calculations

$$\begin{aligned}
&b_1 \pm t_{0.05,n-2} \cdot s_{\overline{b_1}} \\
&5.0603 \pm 2.0639 \cdot 0.0805 \\
&\left[4.8942, \quad 5.2264\right]
\end{aligned}$$

This is, of course, the same interval shown in the Excel printout. When this interval contains zero, the slope is insignificant. Because the interval above does not contain zero, the interval is significant. In simple regression, this test of model significance will always match the other two tests (i.e., the hypothesis test on the correlation coefficient and the *F*-test on the regression model) we have already discussed.

Recall our regression equation.

Regression Equation

$$\hat{Y} = -0.2345 + 5.0603X.$$

This is the equation we use to produce a forecast. If you look back at the data, you will see that Georgia produced 1,247.3 million broilers in 2001 with pounds liveweight of 6,236.5. Suppose we wished to estimate how much pounds liveweight Georgia would have produced had the state produced 1,300 million broilers. We simply plug 1,300 in for *X* in the previous equation, and we see their pounds liveweight would have increased to 6,578.2.

Forecast for 1,300 Million Broilers

$$\begin{aligned} \hat{Y} &= -0.2345 + 5.0603X. \\ &= -0.2345 + 5.0603(1{,}300) \\ &= 6{,}578.2 \end{aligned}$$

But how good a forecast is that? If it ranged from 3,000 to 10,000, then it is not very useful. On the other hand, if it ranged from only 6,500 to 6,656, then it is a very useful estimate. The 6,578.2 is a point estimate of a population value. Once again, we can compute a confidence interval to find the 95 percent range.

The formula used for the confidence interval depends on the type of interval you are constructing. You use the first formula shown when you are computing the confidence interval for the average fitted value. That is, the resulting interval would be the interval for the average of all states that produce 1,300 broilers.

You use the second formula when computing the confidence interval for a single prediction for a new value. Because this confidence interval for the second formula is for a single observation, there is no opportunity for observations to average out, so it results in a wider interval.

Confidence Interval for Average Fitted Value

$$\hat{y}_i \pm t_{0.05,n-2}\sqrt{MSE \cdot \left[\frac{1}{n}+\frac{\left(X_i-\overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right]}$$

Confidence Interval for Predicted Value

$$\hat{y}_i \pm t_{0.05,n-2}\sqrt{MSE \cdot \left[1+\frac{1}{n}+\frac{\left(X_i-\overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right]}$$

Either equation uses several values Excel (or SPSS) does not give us plus one value (mean squared error [*MSE*]) that we have not yet discussed. The value $(X_i - \overline{X})^2$ is different for each confidence interval because the *X*-value is included in the formula. The value $\sum\left(X_i - \overline{X}\right)^2$ is not provided by Excel or SPSS either, but it is easy to compute in Excel. For this problem $\overline{X} = 322.5$ and $\sum\left(X_i - \overline{X}\right)^2 = 3{,}273{,}902.4$. It was computed by subtracting the 322.5 from each observation, squaring the result, and computing the total.

If you look at the ANOVA table in an Excel or SPSS printout, there is a column labeled "MS." MSE is the bottom number, the one on the residual row. For this example, it is 21,198.57.

This gives us the information needed to compute both intervals. Because the only difference for the second interval is the additional "1+," we will take a few shortcuts in its calculation.

Confidence Interval for Average Fitted Value

$$\hat{y}_i \pm t_{0.05,n-2}\sqrt{MSE \cdot \left[\frac{1}{n}+\frac{\left(X_i-\overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right]}$$

$$6{,}578.2 \pm 2.0639\sqrt{21{,}198.57 \cdot \left[\frac{1}{26} + \frac{(1{,}300 - 322.5)^2}{3{,}273{,}902.4}\right]}$$

$$6{,}578.2 \pm 2.0639\sqrt{21{,}198.57 \cdot [0.03846 + 0.2919]}$$

$$6{,}578.2 \pm 2.0639\sqrt{7{,}002.22}$$

$$6{,}578.2 \pm 172.7$$

$$[6{,}405.5, 6{,}750.9]$$

Confidence Interval for Predicted Value

$$\hat{y}_i \pm t_{0.05,n-2}\sqrt{MSE \cdot \left[1 + \frac{1}{n} + \frac{\left(X_i - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right]}$$

$$6{,}578.2 \pm 2.0639\sqrt{21{,}198.57 \cdot [1 + 0.03846 + 0.2919]}$$

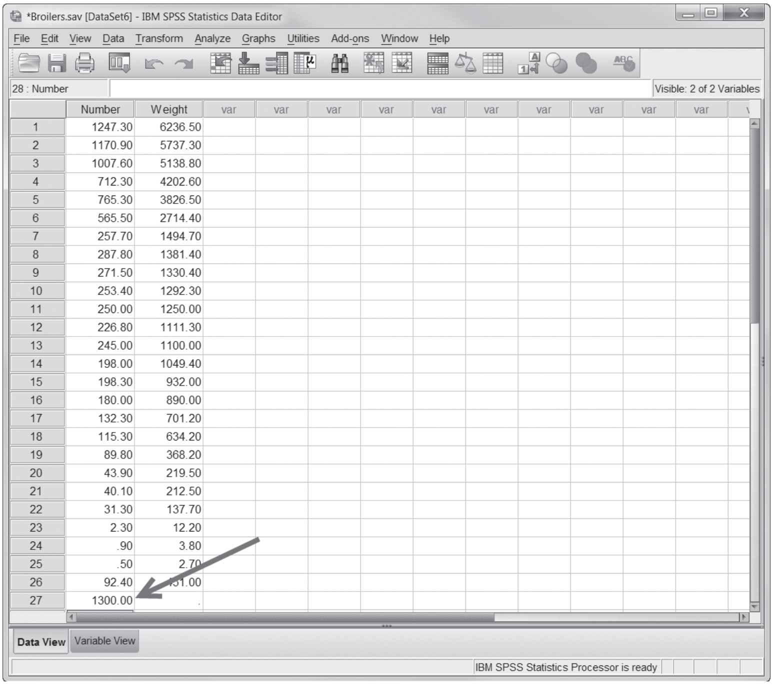
$$6{,}578.2 \pm 346.6$$

$$[6{,}231.6, 6{,}924.8]$$

Whereas SPSS does not give you the values needed to compute these confidence intervals, it will compute them for you. To do this, begin by entering the value of the independent variable you wish to forecast at the bottom of the data set without entering a dependent variable. You can see the 1,300 entered in Figure 2.12. While we are getting only a single prediction here, SPSS can handle as many predictions as you like.

Now, begin as before and click on *Analyze*, *Regression*, and then *Linear*, which brings up the dialog box shown in Figure 2.7. From there, click on the *Save* button. That brings up the dialog box shown in Figure 2.13. As shown in Figure 2.13, we wish to save the *Mean* and *Individual Confidence Intervals*, and as always, we will use a 95 percent confidence interval, although SPSS allows you to specify any value you like. You click on *Continue* to the *Linear Regression* dialog box and continue your regression as before.

In addition to producing the regression results, and adding more data to the output display, SPSS adds four variables to the data file, as shown in Figure 2.14. The variables LMCL_1 and UMCL_1 are the confidence


*Broilers.sav [DataSet6] - IBM SPSS Statistics Data Editor
File
Edit
View
Data
Transform
Analyze
Graphs
Utilities
Add-ons
Window
Help
Data View
Variable View
IBM SPSS Statistics Processor is ready

28 : Number

Visible: 2 of 2 Variables

	Number	Weight	var	var	var	var	var	var	var	var	var	var
1	1247.30	6236.50										
2	1170.90	5737.30										
3	1007.60	5138.80										
4	712.30	4202.60										
5	765.30	3826.50										
6	565.50	2714.40										
7	257.70	1494.70										
8	287.80	1381.40										
9	271.50	1330.40										
10	253.40	1292.30										
11	250.00	1250.00										
12	226.80	1111.30										
13	245.00	1100.00										
14	198.00	1049.40										
15	198.30	932.00										
16	180.00	890.00										
17	132.30	701.20										
18	115.30	634.20										
19	89.80	368.20										
20	43.90	219.50										
21	40.10	212.50										
22	31.30	137.70										
23	2.30	12.20										
24	.90	3.80										
25	.50	2.70										
26	92.40	451.00										
27	1300.00	.										

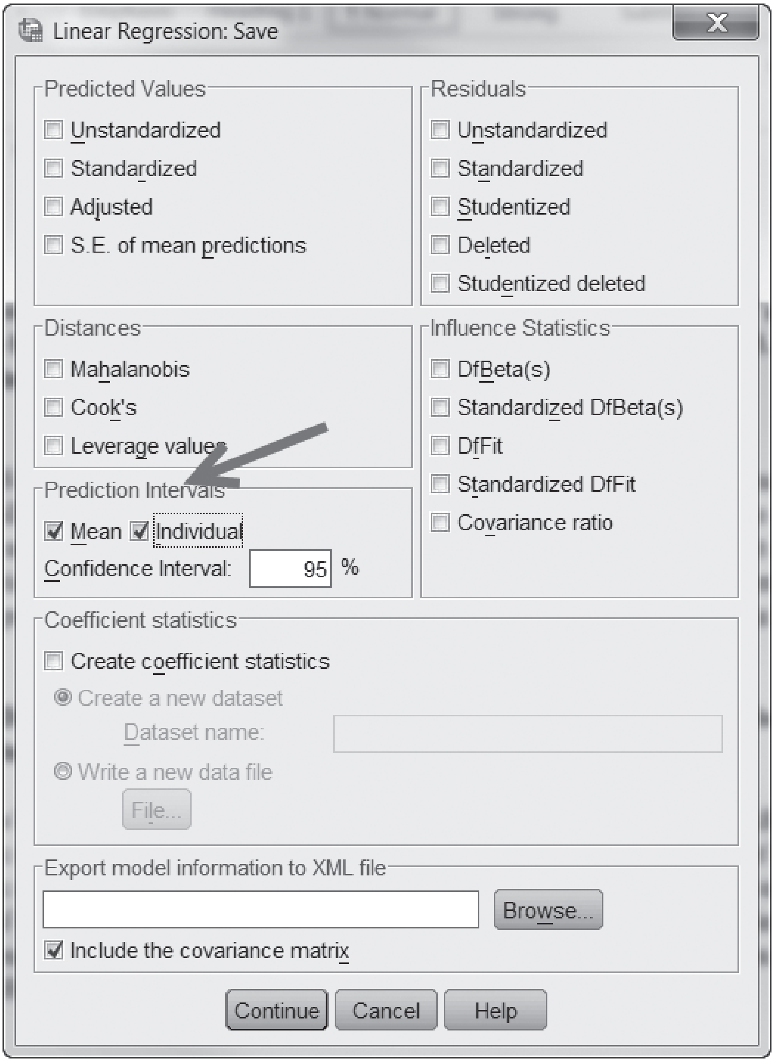
Figure 2.12 Data set up for having SPSS calculate a confidence interval for a predicted value and an average fitted value

interval for the mean prediction confidence interval, and LICI_1 and UICI_1 are the confidence interval for the single-value confidence interval.

The widths of the confidence intervals are not constant because the values of X_i is included in the equation. In this example, the widths for the average value confidence interval range from a low of 59.2 to a high of 164.5. The confidence interval is narrowest near $\bar{X}$ and widest at the extreme values. Note that the last line of Figure 2.14 shows the two confidence intervals for 1,300 and the values are the same as we computed earlier.

Some Final Thoughts on Simple Regression

The widespread use of spreadsheets and inexpensive statistical software has made the use of regression analysis both easy and common. This has both positive and negative consequences. On the good side, more people now have access to a very powerful tool for analyzing relationships and


Linear Regression: Save

Predicted Values
☐ Unstandardized
☐ Standardized
☐ Adjusted
☐ S.E. of mean predictions

Residuals
☐ Unstandardized
☐ Standardized
☐ Studentized
☐ Deleted
☐ Studentized deleted

Distances
☐ Mahalanobis
☐ Cook's
☐ Leverage values

Influence Statistics
☐ DfBeta(s)
☐ Standardized DfBeta(s)
☐ DfFit
☐ Standardized DfFit
☐ Covariance ratio

Prediction Intervals
☑ Mean ☑ Individual
Confidence Interval: 95 %

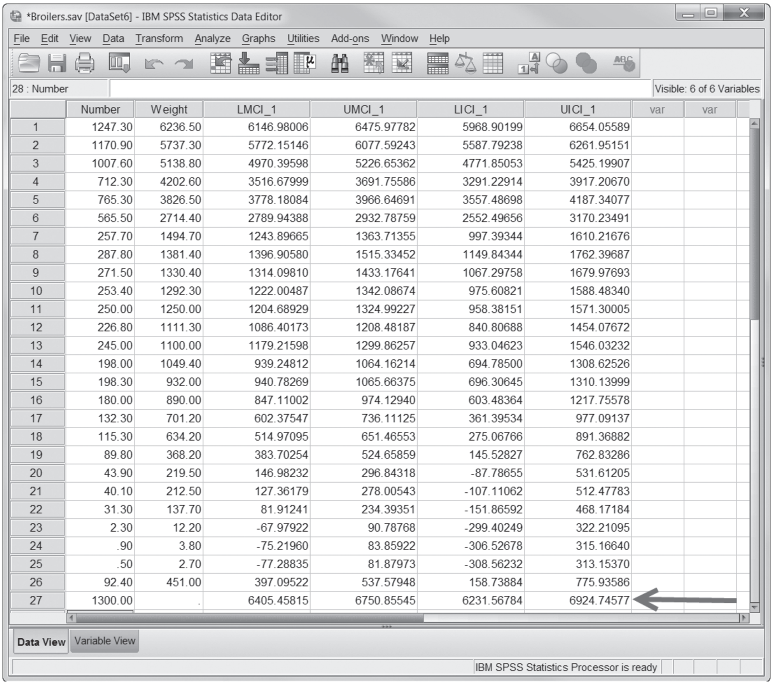
Coefficient statistics
☐ Create coefficient statistics
● Create a new dataset
Dataset name:
○ Write a new data file
File...

Export model information to XML file
Browse...
☑ Include the covariance matrix

Figure 2.13 Telling it to save the confidence intervals in SPSS

performing forecasts. However, it has caused some people to use regression analysis without understanding it and in situations for which it was not appropriate. To help the reader avoid the downfalls of using regression inappropriately, we now offer a few suggestions:

1. Never use regression, or for that matter any statistical tool, without understanding the underlying data. As we saw in the discussion of causality


*Broilers.sav [DataSet6] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
28 : Number
Visible: 6 of 6 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready

	Number	Weight	LMCI_1	UMCI_1	LICI_1	UICI_1	var	var
1	1247.30	6236.50	6146.98006	6475.97782	5968.90199	6654.05589		
2	1170.90	5737.30	5772.15146	6077.59243	5587.79238	6261.95151		
3	1007.60	5138.80	4970.39598	5226.65362	4771.85053	5425.19907		
4	712.30	4202.60	3516.67999	3691.75586	3291.22914	3917.20670		
5	765.30	3826.50	3778.18084	3966.64691	3557.48698	4187.34077		
6	565.50	2714.40	2789.94388	2932.78759	2552.49656	3170.23491		
7	257.70	1494.70	1243.89665	1363.71355	997.39344	1610.21676		
8	287.80	1381.40	1396.90580	1515.33452	1149.84344	1762.39687		
9	271.50	1330.40	1314.09810	1433.17641	1067.29758	1679.97693		
10	253.40	1292.30	1222.00487	1342.08674	975.60821	1588.48340		
11	250.00	1250.00	1204.68929	1324.99227	958.38151	1571.30005		
12	226.80	1111.30	1086.40173	1208.48187	840.80688	1454.07672		
13	245.00	1100.00	1179.21598	1299.86257	933.04623	1546.03232		
14	198.00	1049.40	939.24812	1064.16214	694.78500	1308.62526		
15	198.30	932.00	940.78269	1065.66375	696.30645	1310.13999		
16	180.00	890.00	847.11002	974.12940	603.48364	1217.75578		
17	132.30	701.20	602.37547	736.11125	361.39534	977.09137		
18	115.30	634.20	514.97095	651.46553	275.06766	891.36882		
19	89.80	368.20	383.70254	524.65859	145.52827	762.83286		
20	43.90	219.50	146.98232	296.84318	-87.78655	531.61205		
21	40.10	212.50	127.36179	278.00543	-107.11062	512.47783		
22	31.30	137.70	81.91241	234.39351	-151.86592	468.17184		
23	2.30	12.20	-67.97922	90.78768	-299.40249	322.21095		
24	.90	3.80	-75.21960	83.85922	-306.52678	315.16640		
25	.50	2.70	-77.28835	81.87973	-308.56232	313.15370		
26	92.40	451.00	397.09522	537.57948	158.73884	775.93586		
27	1300.00	.	6405.45815	6750.85545	6231.56784	6924.74577		

Figure 2.14 The data file after running regression and saving the intervals

in chapter 1, it takes a deep understanding of the data and the theory behind it to establish causality. Without either a cause-and-effect relationship or some other theoretical reason for the two variables to move in common, it makes no sense to try to model the variables statistically.

2. Never assume a cause-and-effect relationship exists simply because correlation—even a strong correlation—exists between the variables.
3. Start off with a scatterplot and then correlation analysis before performing simple regression. There is no point trying to model a weak or nonexistent relationship.
4. When using regression to forecast, remember that the further away you go from the range of data used to construct the model, the less reliable the forecast will be.

We will return to these issues in multiple regression. As we will see, all these items will not only still be issues but also be even more complex in multiple regression.

CHAPTER 3

Multiple Regression

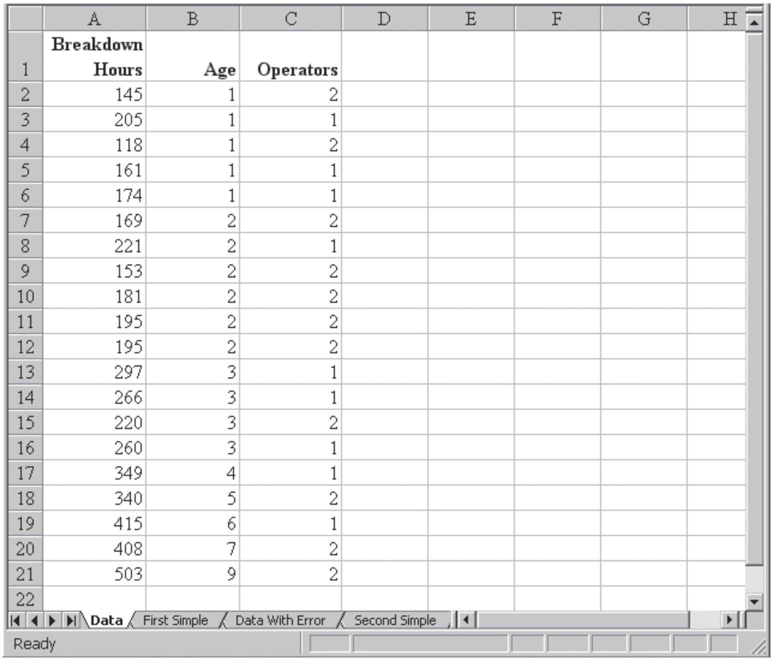
In the last chapter, we saw how to construct a simple regression model. Simple regression described the linear relationship between one dependent variable and a single independent variable. However, in most business situations it takes more than a single independent variable to explain the behavior of the dependent variable. For example, a model to explain company sales might need to include advertising levels, prices, competitor actions, and perhaps many more factors. This example of using various independent variables—like advertising, price, and others—is a good mental model to use when thinking about multiple regression.

When we wish to use more than one independent variable in our regression model, it is called *multiple regression*. Multiple regression can handle as many independent variables as is called for by your theory—at least as long as you have an adequate sample size. However, like simple regression, it, too, is limited to one dependent, or explained, variable.

As we will see, multiple regression is nothing more than simple regression with more independent variables. Most business situations are complex enough that using more than one independent variable does a much better job of either describing how the independent variables impact the dependent variable or producing a forecast of the future behavior of the dependent variable.

Multiple Regression as Several Simple Regression Runs

In addition to the name change, the procedure for calculating the regression model itself changes, although that is not immediately obvious when performing those calculations using Excel. Before we get into that, we will illustrate multiple regression using simple regression.



	A	B	C	D	E	F	G	H
1	Breakdown Hours	Age	Operators					
2	145	1	2					
3	205	1	1					
4	118	1	2					
5	161	1	1					
6	174	1	1					
7	169	2	2					
8	221	2	1					
9	153	2	2					
10	181	2	2					
11	195	2	2					
12	195	2	2					
13	297	3	1					
14	266	3	1					
15	220	3	2					
16	260	3	1					
17	349	4	1					
18	340	5	2					
19	415	6	1					
20	408	7	2					
21	503	9	2					
22								

Figure 3.1 An Excel worksheet with one dependent variable and two independent variables

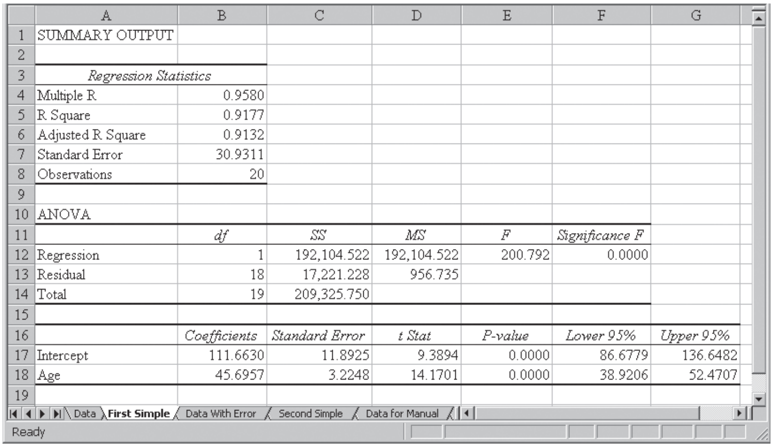
Simple Regression Example

Figure 3.1 shows machine maintenance data for 20 machines in a medium-sized factory. The first column shows the number of hours between its last breakdowns, the second column shows the age of the machine in years, and the third column shows the number of operators. Because we have not yet seen how to use more than one independent variable in regression, we will perform simple regression with breakdown hours as the dependent variable and age as the independent variable. The results of that simple regression are shown in Figure 3.2.

As you can see from Figure 3.2, the overall model is significant and the variation in the age of the machine explains almost 92 percent (0.9177) of the variation in the breakdown hours, giving the resulting regression equation:

The Resulting Regression Equation

$$\text{Breakdown Hours} = 111.6630 + 45.6957(\text{Age})$$



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9580					
5	R Square	0.9177					
6	Adjusted R Square	0.9132					
7	Standard Error	30.9311					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	192,104.522	192,104.522	200.792	0.0000	
13	Residual	18	17,221.228	956.735			
14	Total	19	209,325.750				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	111.6630	11.8925	9.3894	0.0000	86.6779	136.6482
18	Age	45.6957	3.2248	14.1701	0.0000	38.9206	52.4707
19							

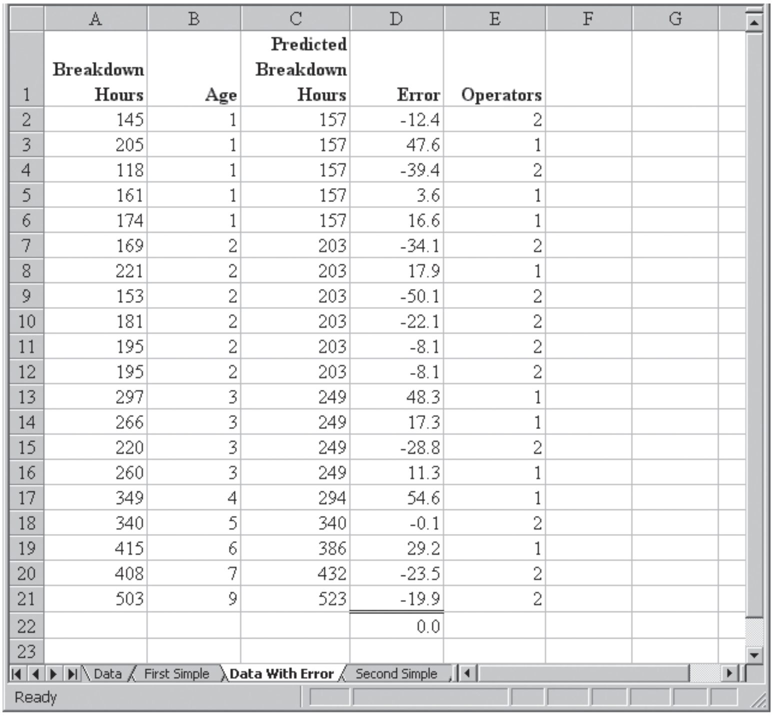
Figure 3.2 Using simple regression and one of the two independent variables

As good as these results are, perhaps the addition of the number of operators as a variable can improve it. In order to see this, we will begin by computing the breakdown hours suggested by the model by plugging the age of the machine into the previous equation. We will then subtract this value from the actual value to obtain the error term for each machine. Those results are shown in Figure 3.3.

Notice that the sum of the error terms is zero. This is a result of how regression works and will always be the case. The error values represent the portion of the variation in breakdown hours that are unexplained by age, because, if age was a perfect explanation, all the error values would be zero. Some of this variation is, naturally, random variation that is unexplainable. However, some of it might be due to the number of operators because that varies among the machines. A good reason for calling this a residual, rather than error, is that parts of this error might, in fact, be explained by another variable—the number of operators in this case.

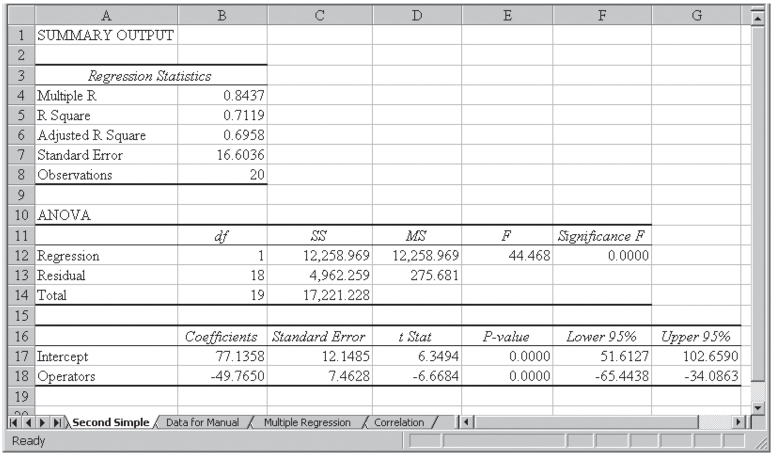
We can check by performing a second simple regression, this time with residual as the dependent variable and number of operators as the independent variable. Those results are shown in Figure 3.4.

Notice that this regression is also significant. The variable for the number of operators explains 71 percent (0.7119) of the variation in the residual term or 71 percent of the remaining 8 percent unexplained variation of the original simple regression model, giving the resulting equation:



	A	B	C	D	E	F	G
1	Breakdown Hours	Age	Predicted Breakdown Hours	Error	Operators		
2	145	1	157	-12.4	2		
3	205	1	157	47.6	1		
4	118	1	157	-39.4	2		
5	161	1	157	3.6	1		
6	174	1	157	16.6	1		
7	169	2	203	-34.1	2		
8	221	2	203	17.9	1		
9	153	2	203	-50.1	2		
10	181	2	203	-22.1	2		
11	195	2	203	-8.1	2		
12	195	2	203	-8.1	2		
13	297	3	249	48.3	1		
14	266	3	249	17.3	1		
15	220	3	249	-28.8	2		
16	260	3	249	11.3	1		
17	349	4	294	54.6	1		
18	340	5	340	-0.1	2		
19	415	6	386	29.2	1		
20	408	7	432	-23.5	2		
21	503	9	523	-19.9	2		
22				0.0			
23							

Figure 3.3 Calculating the part of the breakdown hours not explained by the age of the machine



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.8437					
5	R Square	0.7119					
6	Adjusted R Square	0.6958					
7	Standard Error	16.6036					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	12,258.969	12,258.969	44.468	0.0000	
13	Residual	18	4,962.259	275.681			
14	Total	19	17,221.228				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	77.1358	12.1485	6.3494	0.0000	51.6127	102.6590
18	Operators	-49.7650	7.4628	-6.6684	0.0000	-65.4438	-34.0863
19							

Figure 3.4 Performing simple regression between the residual and the number of operators

Resulting Regression Equation

$$\text{Residual} = 77.1358 - 49.7650(\text{Age})$$

Although this approach of using multiple simple regression runs seems to have worked well in this simplified example, we will see a better approach in the next section. Additionally, with more complex regression applications with many more variables, this approach would quickly become unworkable.

Multiple Regression

In the previous example, we performed simple regression twice: once on the dependent variable and the first independent variable and then on the leftover variation and the second independent variable. With multiple regression, we simultaneously regress all the independent variables against a single dependent variable. Stated another way, the population regression model for a single dependent variable Y and a set of k independent variables $X_1, X_2, \ldots, X_k$ gives the following:

Population Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

where β_0 is the Y-intercept, each of the β_i's for i = 1 to k is the slope of the regression surface with respect to the variable X_i, and e is the error term. This error term is also commonly referred to as the *residual.*

Of course, we rarely work with population data, so we are usually interested in calculating the sample regression model as an estimate of the population regression model:

Sample Regression Model

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + \varepsilon$$

where b_i is the sample statistic that estimates the population parameter β_i.

As you may recall from the last chapter, the graph of the results of simple regression is a line. That is because there are two variables, one independent variable and one dependent variable. With only two variables, the data are defined in two-dimensional space and thus a line. With multiple regression, we have at least three dimensions and possibly many more. When there are two independent variables and one dependent variable, we have three-dimensional space and the results of the regression are a plane in this space. When we exceed two independent variables, we exceed three-dimensional space and, therefore, we exceed our ability to graph the results as well as our ability to visualize those results. When this happens, the results of the regression are said to be a *hyperplane* that exists in *hyperspace*.

Assumptions for Multiple Regression

As you would expect, the assumptions for multiple regression are very similar to the assumptions for simple regression:

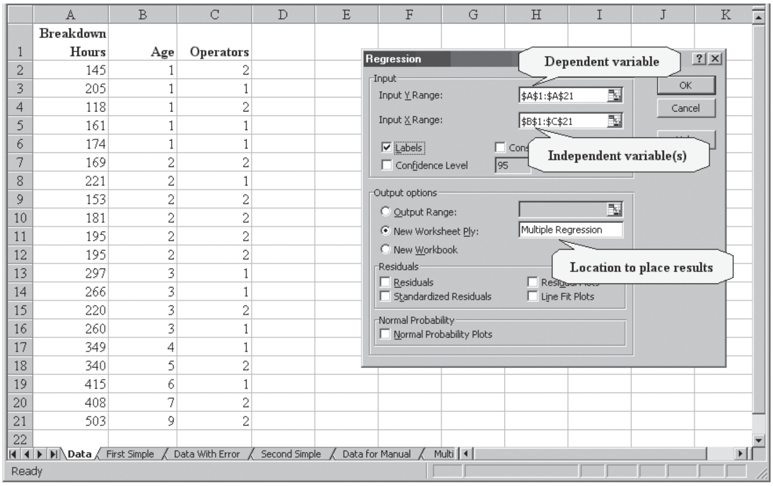
1. For any specific value of any of the independent variables, the values of the dependent variable Y are normally distributed. This is called the assumption of *normality*. As a result of the dependent variable being normally distributed, the error terms will also be normally distributed.
2. The variance for the normal distribution of possible values for the dependent variable is the same for each value of each independent variable. This is called the *equal variance* assumption and is sometimes referred to as *homoscedasticity*.[1]
3. There is a linear relationship between the dependent variable and each of the independent variables. This is called the *linearity* assumption. Because the technique of regression (simple or multiple) only works on linear relationships, when this assumption is violated, that independent variable is usually found to be insignificant. That is, it is found not to make any important contribution to the model. As a result, this assumption is self-enforcing.
4. None of the independent variables are correlated with each other. Although this assumption does not have a name, we will refer to its violation as multicollinearity in a later section, so we will refer to this assumption as the *nonmulticollinearity* assumption.

5. The observations of the dependent variable are independent of each other. That is, there is no correlation between successive error terms, they do not move around together, and there is no trend.[2] This is naturally called the *independence* assumption. In a later section, we will refer to the violation of this assumption as autocorrelation.

Using Excel to Perform Multiple Regression

Excel is able to perform the multiple regression calculations for us. The steps are the same as with simple regression. You begin by clicking on the Data tab, then *Data Analysis*, and then selecting *Regression* from the list of techniques. Excel makes no distinction between simple and multiple regression. Fill in the resulting dialog box just as before, only this time enter the two columns that contain the two independent variables. This is shown in Figure 3.5.

The steps for running multiple regression in Excel are the exact same steps we performed to run simple regression in Excel. In fact, making a distinction between simple and multiple regression is somewhat artificial. As it turns out, some of the complexities that occur when you have two or more independent variables are avoided when there is only one independent variable, so it makes sense to discuss this simpler form first. Nevertheless, both simple and multiple regression are really just regression.



	A	B	C
1	Breakdown Hours	Age	Operators
2	145	1	2
3	205	1	1
4	118	1	2
5	161	1	1
6	174	1	1
7	169	2	2
8	221	2	1
9	153	2	2
10	181	2	2
11	195	2	2
12	195	2	2
13	297	3	1
14	266	3	1
15	220	3	2
16	260	3	1
17	349	4	1
18	340	5	2
19	415	6	1
20	408	7	2
21	503	9	2
22			

Regression

Input

Input Y Range: $A$1:$A$21

Input X Range: $B$1:$C$21

☑ Labels

□ Cons

□ Confidence Level 95

Output options

○ Output Range:

● New Worksheet Ply: Multiple Regression

○ New Workbook

Residuals

□ Residuals

□ Resi

□ Standardized Residuals

□ Line Fit Plots

Normal Probability

□ Normal Probability Plots

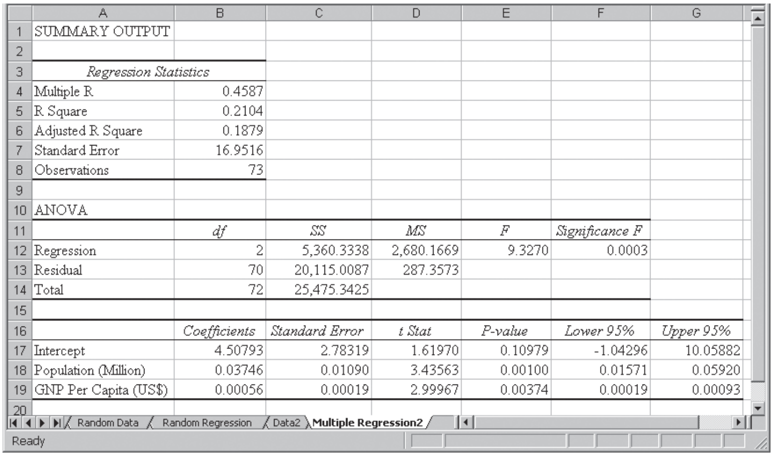
Figure 3.5 Performing multiple regression with Excel

One note is critical regarding the way Excel handles independent variables. Excel *requires* that all the independent variables be located in contiguous columns. That is, there can be nothing in between any of the independent variables, not even hidden columns. This is true regardless of whether you have just two independent variables or if you have dozens. (Of course, this is not an issue in simple regression.) It also requires that all the observations be side by side. That is, all of observation 1 must be on the same row, all of observation 2 must be on the same row, and so on. It is not, however, required that the single dependent variable be beside the independent variables or that the dependent variable observations be on the same rows as their counterparts in the set of independent variables. From a data management perspective, this is nevertheless a very good idea, and this is the way that we will present all the data sets used as examples.

Additionally, Excel's regression procedure sometimes becomes confused by merged cells, even when those merged cells are not within the data set or its labels. If you get an error message saying Excel cannot complete the regression, the first thing you should check for is merged cells.

Example

Figure 3.6 shows the results of running multiple regression on the machine data we have been discussing. Notice that the results match the



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.4587					
5	R Square	0.2104					
6	Adjusted R Square	0.1879					
7	Standard Error	16.9516					
8	Observations	73					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	5,360.3338	2,680.1669	9.3270	0.0003	
13	Residual	70	20,115.0087	287.3573			
14	Total	72	25,475.3425				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	4.50793	2.78319	1.61970	0.10979	-1.04296	10.05882
18	Population (Million)	0.03746	0.01090	3.43563	0.00100	0.01571	0.05920
19	GNP Per Capita (US$)	0.00056	0.00019	2.99967	0.00374	0.00019	0.00093
20							

Figure 3.6 The result of running multiple regression on the machine data discussed earlier in this chapter

appearance of the simple regression results with the exception of having one additional row of information at the bottom to support the additional independent variable. This is, of course, to be expected.

In the machine example, we had the first, simple regression equation:

First Simple Regression Equation

$$\hat{Y} = 111.6630 + 45.6957X$$

and the second regression equation was

Second Regression Equation

$$\hat{Y} = \ 77.1358 - 9.7650b_2$$

Adding these together, we obtain the following equation:

Combined Regression Equation

$$\hat{Y} = 183.7988 + 45.6957b_1 - 49.7650b_2$$

Note that this is similar to, but not exactly equal to, the multiple regression equation we just obtained:

Multiple Regression Equation

$$\hat{Y} = 185.3875 + 47.3511b_1 - 50.7684b_2$$

One of the assumptions of multiple regression is the nonmulticollinearity assumption: There is no correlation between the independent variables. In this example, there is slight (0.1406) correlation between the two independent variables. It is this slight correlation that prevents the total of the individual simple regression equations from totaling to the multiple regression equation. The higher the correlation, the greater the difference there will be between the equations derived using these two approaches. Because there is virtually always some degree of correlation between independent variables, in practice we would never approach a multiple regression equation as a series of simple regression equations.

Another Example

Oftentimes, businesses have a lot of data but their data are not in the right format for statistical analysis. In fact, this so-called dirty data is one of the leading problems that businesses face when trying to analyze the data they have collected. We will explore this problem with an example.

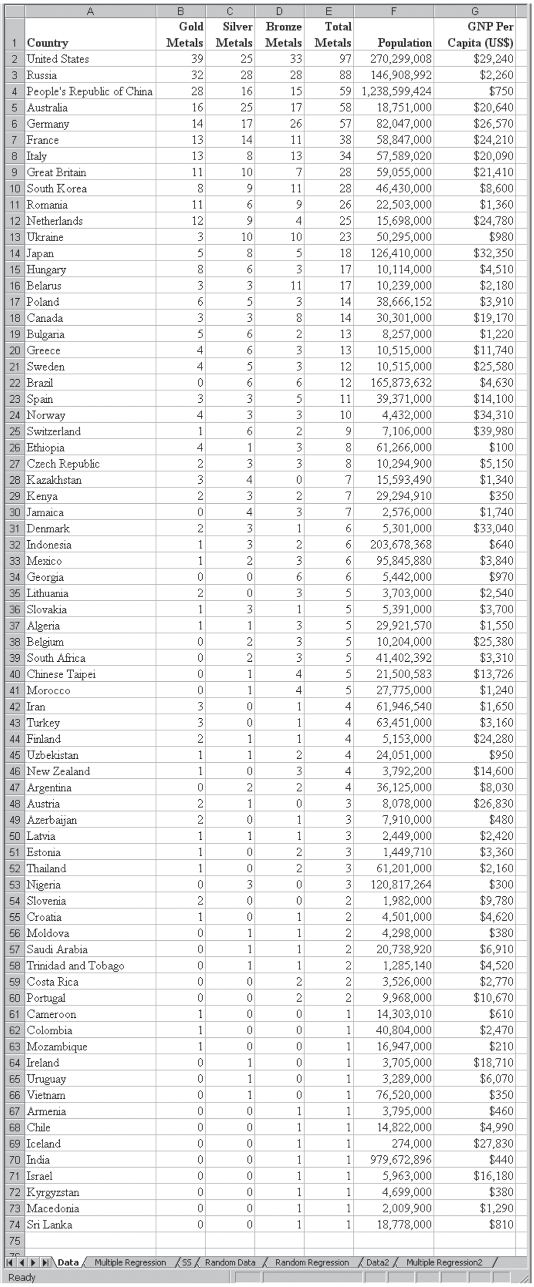
Figure 3.7 shows the medal count by country for the 2000 Olympic Games in Sydney, Australia. It shows the number of gold, silver, and bronze medals along with the total medal count. Also shown are the population and per capita gross national product (GNP) for each country. We will use this information for another multiple regression example, but we must address the dirty data first.

Data Cleaning

Before we continue with the multiple regression portion of this example, we will take this opportunity to discuss data cleaning. Data cleaning is when we resolve any problems with the data that keep us from using it for statistical analysis. These problems must be resolved before these data can be used for regression, or anything else for that matter. Otherwise, any results you obtain will not be meaningful.

When this Sydney Olympics data set was first put together, it had a couple of significant problems. The first problem is that population size and GNP were not available for all the countries. For example, an estimate of GNP was not available for Cuba. The number of countries for which a full set of data was not available was very small, and their medal count was minor so these countries were dropped from the data set.

The second problem was that the per capita GNP was naturally measured in their own currency and a standard scale was needed because it makes sense to use a standard to compare values of a variable where each observation was measured using a different scale. This was handled by converting all the currencies to U.S. dollars, but this raised a third problem: Because the value of currencies constantly fluctuates against the U.S. dollar, which conversion value should be used? We decided to use the conversion value in effect around the time of the Olympics, and the data shown is based on that conversion.



	A	B	C	D	E	F	G
1	Country	Gold Metals	Silver Metals	Bronze Metals	Total Metals	Population	GNP Per Capita (US$)
2	United States	39	25	33	97	270,299,008	$29,240
3	Russia	32	28	28	88	146,908,992	$2,260
4	People's Republic of China	28	16	15	59	1,238,599,424	$750
5	Australia	16	25	17	58	18,751,000	$20,640
6	Germany	14	17	26	57	82,047,000	$26,570
7	France	13	14	11	38	58,847,000	$24,210
8	Italy	13	8	13	34	57,589,020	$20,090
9	Great Britain	11	10	7	28	59,055,000	$21,410
10	South Korea	8	9	11	28	46,430,000	$8,600
11	Romania	11	6	9	26	22,503,000	$1,360
12	Netherlands	12	9	4	25	15,698,000	$24,780
13	Ukraine	3	10	10	23	50,295,000	$980
14	Japan	5	8	5	18	126,410,000	$32,350
15	Hungary	8	6	3	17	10,114,000	$4,510
16	Belarus	3	3	11	17	10,239,000	$2,180
17	Poland	6	5	3	14	38,666,152	$3,910
18	Canada	3	3	8	14	30,301,000	$19,170
19	Bulgaria	5	6	2	13	8,257,000	$1,220
20	Greece	4	6	3	13	10,515,000	$11,740
21	Sweden	4	5	3	12	10,515,000	$25,580
22	Brazil	0	6	6	12	165,873,632	$4,630
23	Spain	3	3	5	11	39,371,000	$14,100
24	Norway	4	3	3	10	4,432,000	$34,310
25	Switzerland	1	6	2	9	7,106,000	$39,980
26	Ethiopia	4	1	3	8	61,266,000	$100
27	Czech Republic	2	3	3	8	10,294,900	$5,150
28	Kazakhstan	3	4	0	7	15,593,490	$1,340
29	Kenya	2	3	2	7	29,294,910	$350
30	Jamaica	0	4	3	7	2,576,000	$1,740
31	Denmark	2	3	1	6	5,301,000	$33,040
32	Indonesia	1	3	2	6	203,678,368	$640
33	Mexico	1	2	3	6	95,845,880	$3,840
34	Georgia	0	0	6	6	5,442,000	$970
35	Lithuania	2	0	3	5	3,703,000	$2,540
36	Slovakia	1	3	1	5	5,391,000	$3,700
37	Algeria	1	1	3	5	29,921,570	$1,550
38	Belgium	0	2	3	5	10,204,000	$25,380
39	South Africa	0	2	3	5	41,402,392	$3,310
40	Chinese Taipei	0	1	4	5	21,500,583	$13,726
41	Morocco	0	1	4	5	27,775,000	$1,240
42	Iran	3	0	1	4	61,946,540	$1,650
43	Turkey	3	0	1	4	63,451,000	$3,160
44	Finland	2	1	1	4	5,153,000	$24,280
45	Uzbekistan	1	1	2	4	24,051,000	$950
46	New Zealand	1	0	3	4	3,792,200	$14,600
47	Argentina	0	2	2	4	36,125,000	$8,030
48	Austria	2	1	0	3	8,078,000	$26,830
49	Azerbaijan	2	0	1	3	7,910,000	$480
50	Latvia	1	1	1	3	2,449,000	$2,420
51	Estonia	1	0	2	3	1,449,710	$3,360
52	Thailand	1	0	2	3	61,201,000	$2,160
53	Nigeria	0	3	0	3	120,817,264	$300
54	Slovenia	2	0	0	2	1,982,000	$9,780
55	Croatia	1	0	1	2	4,501,000	$4,620
56	Moldova	0	1	1	2	4,298,000	$380
57	Saudi Arabia	0	1	1	2	20,738,920	$6,910
58	Trinidad and Tobago	0	1	1	2	1,285,140	$4,520
59	Costa Rica	0	0	2	2	3,526,000	$2,770
60	Portugal	0	0	2	2	9,968,000	$10,670
61	Cameroon	1	0	0	1	14,303,010	$610
62	Colombia	1	0	0	1	40,804,000	$2,470
63	Mozambique	1	0	0	1	16,947,000	$210
64	Ireland	0	1	0	1	3,705,000	$18,710
65	Uruguay	0	1	0	1	3,289,000	$6,070
66	Vietnam	0	1	0	1	76,520,000	$350
67	Armenia	0	0	1	1	3,795,000	$460
68	Chile	0	0	1	1	14,822,000	$4,990
69	Iceland	0	0	1	1	274,000	$27,830
70	India	0	0	1	1	979,672,896	$440
71	Israel	0	0	1	1	5,963,000	$16,180
72	Kyrgyzstan	0	0	1	1	4,699,000	$380
73	Macedonia	0	0	1	1	2,009,900	$1,290
74	Sri Lanka	0	0	1	1	18,778,000	$810
75							

Figure 3.7 Data from the 2000 Olympic Games in Sydney, Australia

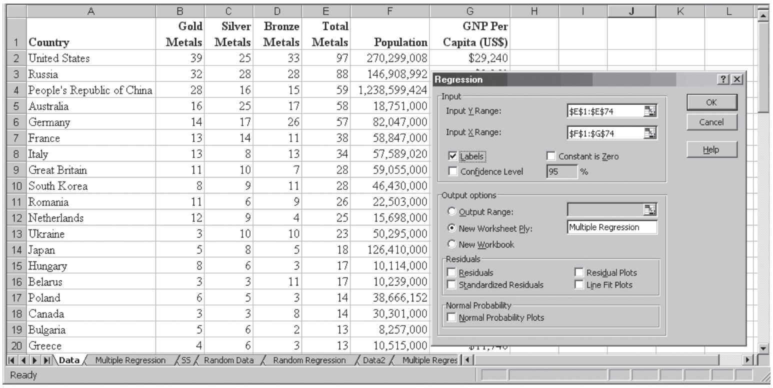
It is not uncommon for business data to require cleaning before being ready for use in statistical analysis. In this context, we use statistical analysis in a very broad sense and not just in reference to multiple regression. As long as only a few data points are discarded and as long as any data conversions are reasonable, the data cleaning should not have too much of an impact on the results of any statistical analysis performed on the data. In any case, we do not have much of a choice. Without cleaning, the data would not be in a useable form.

Olympics Example Continued

Returning to our data set from the 2000 Sydney Olympics, we will use total medal count[3] as the dependent variable and per capita GNP and population as the independent variables. Figure 3.8 shows the dialog box filled out to perform the multiple regression and Figure 3.9 shows the results.

r^2 Can Be Low but Significant

Notice that in the Sydney Olympics example the variations in population and per capita GNP explain only 21 percent (0.2104) of the variation in total medals, yet the overall model is significant and the individual variables are all significant as well. This is an important point. It is not



	A	B	C	D	E	F	G
1	Country	Gold Metals	Silver Metals	Bronze Metals	Total Metals	Population	GNP Per Capita (US$)
2	United States	39	25	33	97	270,299,008	$29,240
3	Russia	32	28	28	88	146,908,992	
4	People's Republic of China	28	16	15	59	1,238,599,424	
5	Australia	16	25	17	58	18,751,000	
6	Germany	14	17	26	57	82,047,000	
7	France	13	14	11	38	58,847,000	
8	Italy	13	8	13	34	57,589,020	
9	Great Britain	11	10	7	28	59,055,000	
10	South Korea	8	9	11	28	46,430,000	
11	Romania	11	6	9	26	22,503,000	
12	Netherlands	12	9	4	25	15,698,000	
13	Ukraine	3	10	10	23	50,295,000	
14	Japan	5	8	5	18	126,410,000	
15	Hungary	8	6	3	17	10,114,000	
16	Belarus	3	3	11	17	10,239,000	
17	Poland	6	5	3	14	38,666,152	
18	Canada	3	3	8	14	30,301,000	
19	Bulgaria	5	6	2	13	8,257,000	
20	Greece	4	6	3	13	10,515,000	$11,740

Regression

Input

Input Y Range: $E$1:$E$74

Input X Range: $F$1:$G$74

☑ Labels ☐ Constant is Zero

☐ Confidence Level 95 %

Output options

○ Output Range:

● New Worksheet Ply: Multiple Regression

○ New Workbook

Residuals

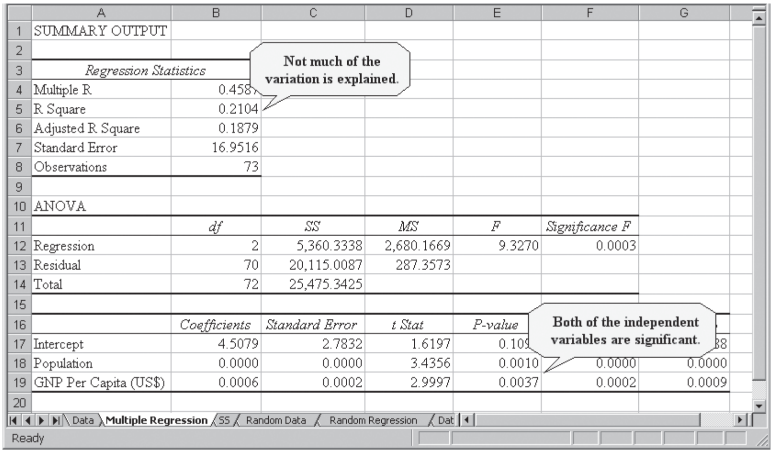
☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Figure 3.8 The dialog box used to perform multiple regression on the 2000 Olympic Games data set



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.458					
5	R Square	0.2104					
6	Adjusted R Square	0.1879					
7	Standard Error	16.9516					
8	Observations	73					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	5,360.3338	2,680.1669	9.3270	0.0003	
13	Residual	70	20,115.0087	287.3573			
14	Total	72	25,475.3425				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*		
17	Intercept	4.5079	2.7832	1.6197	0.109		38
18	Population	0.0000	0.0000	3.4356	0.0010	0.0000	0.0000
19	GNP Per Capita (US$)	0.0006	0.0002	2.9997	0.0037	0.0002	0.0009
20							

Not much of the variation is explained.
Both of the independent variables are significant.
Data
Multiple Regression
SS
Random Data
Random Regression
Dat
Ready

Figure 3.9 The results of the multiple regression analysis on the 2000 Olympic Games data set

necessary for r^2 to be high in order for the overall model to be significant. This is especially true with larger data sets—that is, with higher numbers of observations. When businesses analyze massive data sets, it is not uncommon for even unrelated variables to show significant correlations for this very reason.

Students sometimes also make the mistake of thinking that only models with high r^2 values are useful. It is easy to see why students might believe this. Because r^2 represents the percentage of the variation in the dependent variable that is explained by variation in the independent variables, one might conclude that a model that explains only a small percentage of the variation is not all that useful.

In a business situation, this is, in fact, a reasonable assumption. A forecasting model that explained only 21 percent of the variation in demand would not be very useful in helping to plan production. Likewise, a market analysis that ends up explaining only 21 percent of the variation in demand would likely have missed the more important explainer variables.

However, in other areas, explaining even a small percentage of the variation might be useful. For example, doctors might find a model that explained only 21 percent of the variation in the onset of Alzheimer's disease to be very useful. Thus the decision about how useful a model is, at least once it has been found to be statistically significant, should be theory based and not statistically based. This requires knowledge of the field under study rather than statistics. This is the main reason that statisticians usually need help from knowledgeable outsiders when developing models.

Example

We will now look at an example involving celebrities. Although this is clearly a nonbusiness example (unless, of course, you are in the business of making movies), the procedures and considerations are exactly the same as performing a marketing analysis to try to understand why your product is (or is not) popular.

Forbes collected the following information on the top 100 celebrities for 2000:

- Earnings rank, or simply a rank ordering of 1999 earnings
- Earnings for 1999
- Web hits across the Internet
- Press clips from Lexis-Nexis
- Magazine covers
- Number of mentions on radio and television

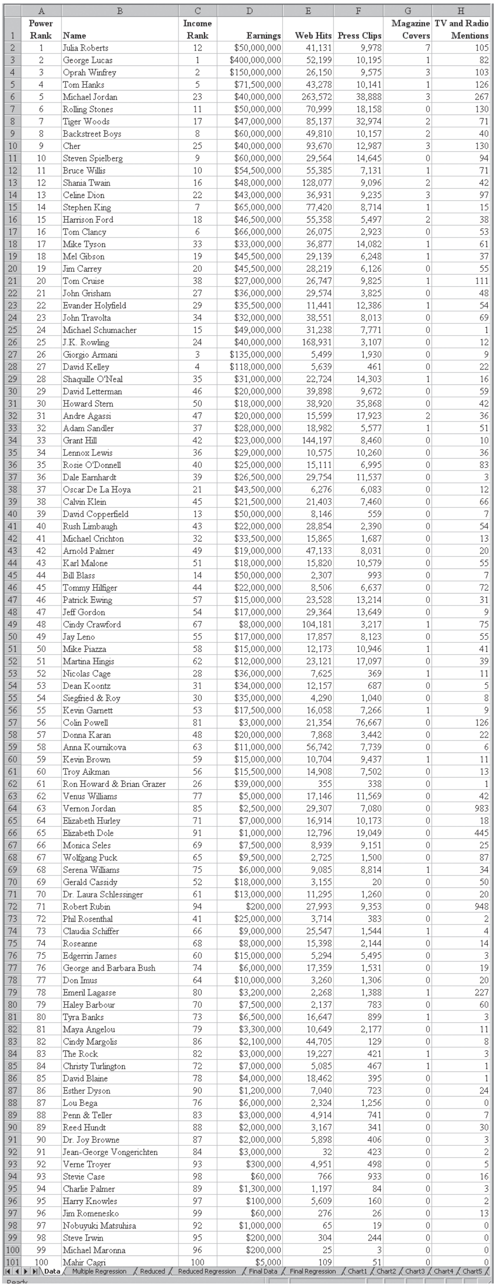
The data are collected in the worksheet Celebrities.xls, which is shown in Figure 3.10. *Forbes* used this information to decide on a "power rank" for each celebrity. We will use multiple regression to try to discover the rationale behind the power rank. That regression is shown in Figure 3.11.

The r^2 value is 0.9245, so over 92 percent of the variation in the power rank is explained by this data, giving the resulting equation:

Power Rank Equation

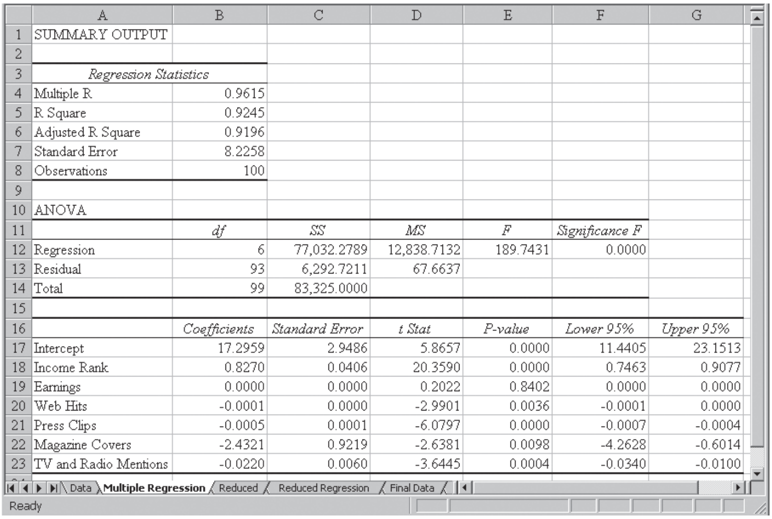
$$\begin{aligned}\text{Power Rank} = {} & 17.2959 + 0.8270(\text{Income Rank}) \\ & + 0.00004[4](\text{Earnings}) - 0.0001[5](\text{Web Hits}) - 0.0005(\text{Press Clips}) \\ & - 2.4321(\text{Magazine Covers}) - 0.0220(\text{TV and Radio Mentions})\end{aligned}$$

Some of the signs in this equation seem unusual. We will have more to say about this later. But before we get into this, we need to discuss how to



	A	B	C	D	E	F	G	H
1	Power Rank	Name	Income Rank	Earnings	Web Hits	Press Clips	Magazine Covers	TV and Radio Mentions
2	1	Julia Roberts	12	$50,000,000	41,131	9,978	7	105
3	2	George Lucas	1	$400,000,000	52,199	10,195	1	82
4	3	Oprah Winfrey	2	$150,000,000	26,150	9,575	3	103
5	4	Tom Hanks	5	$71,500,000	43,278	10,141	1	126
6	5	Michael Jordan	23	$40,000,000	263,572	38,888	3	267
7	6	Rolling Stones	11	$50,000,000	70,999	18,158	0	130
8	7	Tiger Woods	17	$47,000,000	85,137	32,974	2	71
9	8	Backstreet Boys	8	$60,000,000	49,810	10,157	2	40
10	9	Cher	25	$40,000,000	93,670	12,987	3	130
11	10	Steven Spielberg	9	$60,000,000	29,564	14,645	0	94
12	11	Bruce Willis	10	$54,500,000	55,385	7,131	1	71
13	12	Shania Twain	16	$48,000,000	128,077	9,096	2	42
14	13	Celine Dion	22	$43,000,000	36,931	9,235	3	97
15	14	Stephen King	7	$65,000,000	77,420	8,714	1	15
16	15	Harrison Ford	18	$46,500,000	55,358	5,497	2	38
17	16	Tom Clancy	6	$66,000,000	26,075	2,923	0	53
18	17	Mike Tyson	33	$33,000,000	36,877	14,082	1	61
19	18	Mel Gibson	19	$45,500,000	29,139	6,248	1	37
20	19	Jim Carrey	20	$45,500,000	28,219	6,126	0	55
21	20	Tom Cruise	38	$27,000,000	26,747	9,825	1	111
22	21	John Grisham	27	$36,000,000	29,574	3,825	0	48
23	22	Evander Holyfield	29	$35,500,000	11,441	12,386	1	54
24	23	John Travolta	34	$32,000,000	38,551	8,013	0	69
25	24	Michael Schumacher	15	$49,000,000	31,238	7,771	0	1
26	25	J.K. Rowling	24	$40,000,000	168,931	3,107	0	12
27	26	Giorgio Armani	3	$135,000,000	5,499	1,930	0	9
28	27	David Kelley	4	$118,000,000	5,639	461	0	22
29	28	Shaquille O'Neal	35	$31,000,000	22,724	14,303	1	16
30	29	David Letterman	46	$20,000,000	39,898	9,672	0	59
31	30	Howard Stern	50	$18,000,000	38,920	35,868	0	42
32	31	Andre Agassi	47	$20,000,000	15,599	17,923	2	36
33	32	Adam Sandler	37	$28,000,000	18,982	5,577	1	51
34	33	Grant Hill	42	$23,000,000	144,197	8,460	0	10
35	34	Lennox Lewis	36	$29,000,000	10,575	10,260	0	36
36	35	Rosie O'Donnell	40	$25,000,000	15,111	6,995	0	83
37	36	Dale Earnhardt	39	$26,500,000	29,754	11,537	0	3
38	37	Oscar De La Hoya	21	$43,500,000	6,276	6,083	0	12
39	38	Calvin Klein	45	$21,500,000	21,403	7,460	0	66
40	39	David Copperfield	13	$50,000,000	8,146	559	0	7
41	40	Rush Limbaugh	43	$22,000,000	28,854	2,390	0	54
42	41	Michael Crichton	32	$33,500,000	15,865	1,687	0	13
43	42	Arnold Palmer	49	$19,000,000	47,133	8,031	0	20
44	43	Karl Malone	51	$18,000,000	15,820	10,579	0	55
45	44	Bill Blass	14	$50,000,000	2,307	993	0	7
46	45	Tommy Hilfiger	44	$22,000,000	8,506	6,637	0	72
47	46	Patrick Ewing	57	$15,000,000	23,528	13,214	0	31
48	47	Jeff Gordon	54	$17,000,000	29,364	13,649	0	9
49	48	Cindy Crawford	67	$8,000,000	104,181	3,217	1	75
50	49	Jay Leno	55	$17,000,000	17,857	8,123	0	55
51	50	Mike Piazza	58	$15,000,000	12,173	10,946	1	41
52	51	Martina Hingis	62	$12,000,000	23,121	17,097	0	39
53	52	Nicolas Cage	28	$36,000,000	7,625	369	1	11
54	53	Dean Koontz	31	$34,000,000	12,157	687	0	5
55	54	Siegfried & Roy	30	$35,000,000	4,290	1,040	0	8
56	55	Kevin Garnett	53	$17,500,000	16,058	7,266	1	9
57	56	Colin Powell	81	$3,000,000	21,354	76,667	0	126
58	57	Donna Karan	48	$20,000,000	7,868	3,442	0	22
59	58	Anna Kournikova	63	$11,000,000	56,742	7,739	0	6
60	59	Kevin Brown	59	$15,000,000	10,704	9,437	1	11
61	60	Troy Aikman	56	$15,500,000	14,908	7,502	0	13
62	61	Ron Howard & Brian Grazer	26	$39,000,000	355	338	0	1
63	62	Venus Williams	77	$5,000,000	17,146	11,569	0	42
64	63	Vernon Jordan	85	$2,500,000	29,307	7,080	0	983
65	64	Elizabeth Hurley	71	$7,000,000	16,914	10,173	0	18
66	65	Elizabeth Dole	91	$1,000,000	12,796	19,049	0	445
67	66	Monica Seles	69	$7,500,000	8,939	9,151	0	25
68	67	Wolfgang Puck	65	$9,500,000	2,725	1,500	0	87
69	68	Serena Williams	75	$6,000,000	9,085	8,814	1	34
70	69	Gerald Cassidy	52	$18,000,000	3,155	20	0	50
71	70	Dr. Laura Schlessinger	61	$13,000,000	11,295	1,260	0	20
72	71	Robert Rubin	94	$200,000	27,993	9,353	0	948
73	72	Phil Rosenthal	41	$25,000,000	3,714	383	0	2
74	73	Claudia Schiffer	66	$9,000,000	25,547	1,544	1	4
75	74	Roseanne	68	$8,000,000	15,398	2,144	0	14
76	75	Edgerrin James	60	$15,000,000	5,294	5,495	0	3
77	76	George and Barbara Bush	74	$6,000,000	17,359	1,531	0	19
78	77	Don Imus	64	$10,000,000	3,260	1,306	0	20
79	78	Emeril Lagasse	80	$3,200,000	2,268	1,388	1	227
80	79	Haley Barbour	70	$7,500,000	2,137	783	0	60
81	80	Tyra Banks	73	$6,500,000	16,647	899	1	3
82	81	Maya Angelou	79	$3,300,000	10,649	2,177	0	11
83	82	Cindy Margolis	86	$2,100,000	44,705	129	0	8
84	83	The Rock	82	$3,000,000	19,227	421	1	3
85	84	Christy Turlington	72	$7,000,000	5,085	467	1	1
86	85	David Blaine	78	$4,000,000	18,462	395	0	1
87	86	Esther Dyson	90	$1,200,000	7,040	723	0	24
88	87	Lou Bega	76	$6,000,000	2,324	1,256	0	0
89	88	Penn & Teller	83	$3,000,000	4,914	741	0	7
90	89	Reed Hundt	88	$2,000,000	3,167	341	0	30
91	90	Dr. Joy Browne	87	$2,000,000	5,898	406	0	3
92	91	Jean-George Vongerichten	84	$3,000,000	32	423	0	2
93	92	Verne Troyer	93	$300,000	4,951	498	0	5
94	93	Stevie Case	98	$60,000	766	933	0	16
95	94	Charlie Palmer	89	$1,300,000	1,197	84	0	3
96	95	Harry Knowles	97	$100,000	5,609	160	0	2
97	96	Jim Romenesko	99	$60,000	276	26	0	13
98	97	Nobuyuki Matsuhisa	92	$1,000,000	65	19	0	0
99	98	Steve Irwin	95	$200,000	304	244	0	0
100	99	Michael Maronna	96	$200,000	25	3	0	0
101	100	Mahir Cagri	100	$5,000	109	51	0	0

Figure 3.10 Forbes 2000 data on 100 celebrities



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9615					
5	R Square	0.9245					
6	Adjusted R Square	0.9196					
7	Standard Error	8.2258					
8	Observations	100					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	6	77,032.2789	12,838.7132	189.7431	0.0000	
13	Residual	93	6,292.7211	67.6637			
14	Total	99	83,325.0000				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	17.2959	2.9486	5.8657	0.0000	11.4405	23.1513
18	Income Rank	0.8270	0.0406	20.3590	0.0000	0.7463	0.9077
19	Earnings	0.0000	0.0000	0.2022	0.8402	0.0000	0.0000
20	Web Hits	-0.0001	0.0000	-2.9901	0.0036	-0.0001	0.0000
21	Press Clips	-0.0005	0.0001	-6.0797	0.0000	-0.0007	-0.0004
22	Magazine Covers	-2.4321	0.9219	-2.6381	0.0098	-4.2628	-0.6014
23	TV and Radio Mentions	-0.0220	0.0060	-3.6445	0.0004	-0.0340	-0.0100

Figure 3.11 The multiple regression results on the year 2000 Forbes data on 100 celebrities with Power Rank as the dependent variable

evaluate the significance of the overall model, as well as the significances of the individual components.

The *F*-Test on a Multiple Regression Model

The first statistical test we need to perform is to test and see if the overall multiple regression model is significant. After all, if the overall model is insignificant then there is no point is looking to see what parts of the model might be significant. Back with simple regression, we performed the following test on the correlation coefficient between the single dependent variable and the single independent variable and said that if the correlation was significant then the overall model would also be significant:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

With multiple regression, we can no longer use that approach. The reason is simple: If there are six independent variables, then there are six different correlations between the single dependent variable and each of the independent variables.[6]

With simple regression, a significant correlation between the single dependent and the single dependent variable indicated a significant model. With multiple regression, we may end up with data where some of the independent variables are significant whereas others are insignificant. For this reason, we need a test that will test all the independent variables at once. That is, we want to test the following:

$$\begin{aligned}&H_0 : \beta_1 = \beta_2 = \quad = \beta_k = 0\\ &\text{H1: } \beta_i \neq \text{for some } i\end{aligned}$$

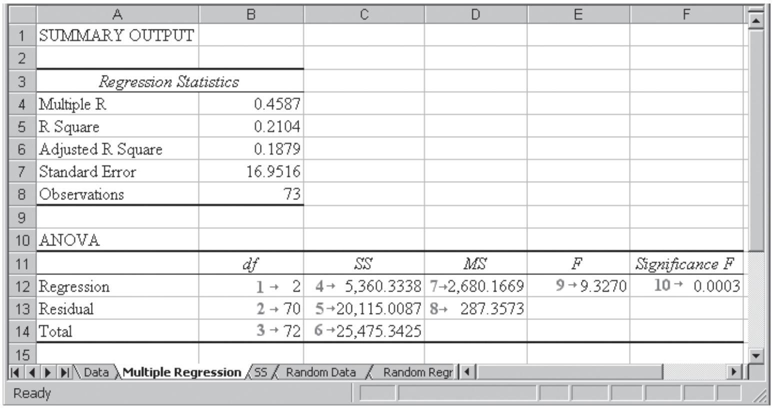
In other words, we are testing to see that at least one β_i is not equal to zero.

Think about the logic for a minute. If all the β_i's in the model equal zero, then the data we have collected is of no value in explaining the dependent variable. So, basically, in specifying this null hypothesis we are saying that the model is useless. On the other hand, if at least one of the β_i's is not zero, then at least some part of the model helps us explain the dependent variable. This is the logic behind the alternative hypothesis. Of course, we will still need to delve into the model and figure out which part is really helping us. That is a topic for a later section.

We will use analysis of variance (ANOVA) to perform the test on the hypotheses shown previously. Before looking at the hypothesis test, a couple of notes are in order regarding the aforementioned hypotheses. First, notice that the null hypothesis does *not* specify the intercept, β_0. As with simple regression, we are rarely interested in the significance of the intercept. As discussed previously, it is possible that some of the β_i's will be significant whereas others will be insignificant. As long as any one of them is significant, the model will pass the ANOVA test.

We will briefly review ANOVA as it relates to multiple regression hypothesis testing. Interested students should refer to a statistics textbook for more details. Figure 3.12 shows just the ANOVA section from the 2000 Sydney Olympics results regression shown previously. In this discussion, the numbers with arrows after them are shown in the figure as references to the items under discussion. They are not, of course, normally a part of the ANOVA results.

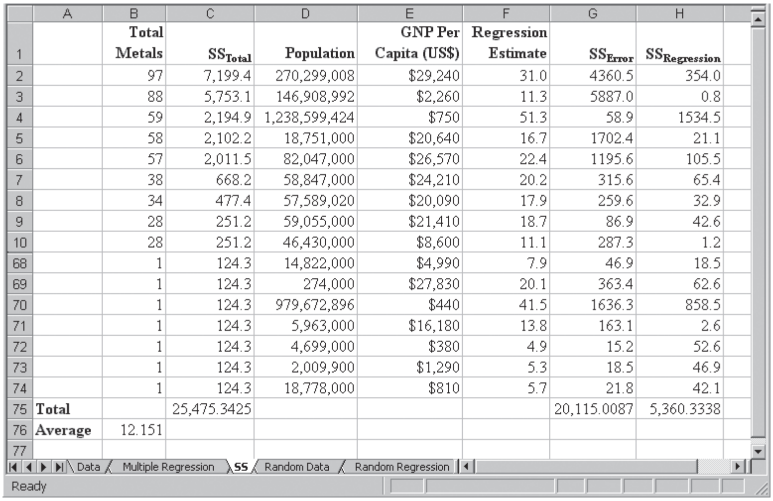
We will now discuss each of the notes shown in Figure 3.12.



	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0.4587				
5	R Square	0.2104				
6	Adjusted R Square	0.1879				
7	Standard Error	16.9516				
8	Observations	73				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1 → 2	4 → 5,360.3338	7 → 2,680.1669	9 → 9.3270	10 → 0.0003
13	Residual	2 → 70	5 → 20,115.0087	8 → 287.3573		
14	Total	3 → 72	6 → 25,475.3425			
15						

Figure 3.12 ANOVA section for the 2000 Sydney Olympics results regression

1. The regression degrees of freedom is *k*, the number of independent variables (two in this case).
2. The residual (or error) degrees of freedom is *n* – *k*–1, or 73 – 2 – 1 = 70 in this case.
3. The total degrees of freedom is *n* –1, or 72 in this case.
4. This is the sums of squares (*SS*) regression. We will abbreviate this as *SSR*. Its calculation will be discussed in more detail shortly.
5. This is *SS* residual. In order to avoid confusion with *SS* regression, we will abbreviate this as *SSE*. Its calculation will be discussed in more detail shortly.
6. This is *SS* total. We will abbreviate this as *SST*. Its calculation will be discussed in more detail shortly.
7. This is mean square regression, or *MS* regression. We will abbreviate this as *MSR*. It is calculated as *SSR* / *k*.
8. This is mean square residual/error or *MS* residual. We will abbreviate this as *MSE*. It is calculated as *SSE* / (*n* – *k* – 1).
9. The *F* ratio is calculated as *MSR/MSE*. This value is chi square distributed with *k*,*n* – *k* – 1 degrees of freedom.
10. This is the *p*-value for the *F*-test. When it is less than 0.05, the overall model is significant, and when it is greater than 0.05, the overall model is not significant. That is, you reject the null hypothesis that all the β_k coefficients are zero. If the overall model is not significant,



	A	B	C	D	E	F	G	H
1		Total Metals	SS_{Total}	Population	GNP Per Capita (US$)	Regression Estimate	SS_{Error}	$SS_{Regression}$
2		97	7,199.4	270,299,008	$29,240	31.0	4360.5	354.0
3		88	5,753.1	146,908,992	$2,260	11.3	5887.0	0.8
4		59	2,194.9	1,238,599,424	$750	51.3	58.9	1534.5
5		58	2,102.2	18,751,000	$20,640	16.7	1702.4	21.1
6		57	2,011.5	82,047,000	$26,570	22.4	1195.6	105.5
7		38	668.2	58,847,000	$24,210	20.2	315.6	65.4
8		34	477.4	57,589,020	$20,090	17.9	259.6	32.9
9		28	251.2	59,055,000	$21,410	18.7	86.9	42.6
10		28	251.2	46,430,000	$8,600	11.1	287.3	1.2
68		1	124.3	14,822,000	$4,990	7.9	46.9	18.5
69		1	124.3	274,000	$27,830	20.1	363.4	62.6
70		1	124.3	979,672,896	$440	41.5	1636.3	858.5
71		1	124.3	5,963,000	$16,180	13.8	163.1	2.6
72		1	124.3	4,699,000	$380	4.9	15.2	52.6
73		1	124.3	2,009,900	$1,290	5.3	18.5	46.9
74		1	124.3	18,778,000	$810	5.7	21.8	42.1
75	**Total**		25,475.3425				20,115.0087	5,360.3338
76	**Average**	12.151						
77								

Figure 3.13 The values from the data from the 2000 Sydney Olympics that will be used to calculate the sums of squares values

that is, if all the β_k coefficients are equal to zero, then there is no point in continuing with the regression analysis.

We will briefly review the calculation of *SSR*, *SSE*, and *SST*. However, because Excel and every other multiple regression program report these values, no additional emphasis will be placed on their manual calculation. Figure 3.13 shows the data from the 2000 Sydney Olympics with the information required to calculate all three sums of squares values. Note that some of the rows are hidden in this figure. That was done to reduce the size of the figure.

From the data, the average of the number of medals won (*Y*) is 12.151, giving the following equation for *SST*:

SST

$$SST = \sum\left(Y - \bar{Y}\right)^2$$

For row 1, 97 – 12.151 = 84.8 and 84.8^2 = 7,199.4.[7] For row 2, 88 – 12.151 = 75.8 and 75.8^2 = 5,753.1. These calculations are carried out for each of the data points, and the total of these squared values is 25,475.3425. This is the *SST* value shown back in Figure 3.12.

The formula for *SSE* is the following:

SSE

$$SSE = \sum \left(Y - \hat{Y}\right)^2$$

For row 1, the predicted *Y* value is 31.0 and $(97 – 31.0)^2$ = 4,360.5. For row 2, the predicted *Y* value is 11.3 and $(88 – 11.3)^2$ = 5,887.0. These calculations are carried out for all the data, and the total of these squared values is 20,115.0087. This is the *SSE* value shown back in Figure 3.12.

At this point, there is no need to compute *SSR* because the formula

SST

$$SSR + SSE = SST$$

allows us to compute it based on *SST* and *SSE*.[8] Nevertheless, *SSR* is represented by the following equation:

SSR

$$SSR = \sum \left(\hat{Y} - \bar{Y}\right)^2$$

For row 1, that gives us 31.0 – 12.151 = 18.8 and 18.8^2 = 354.0. For row 2, 11.3 – 12.151 = –0.9 and $–0.9^2$ = 0.8. These calculations are carried out for all the data, and the total of these squared values is 5,360.3338. This is the *SSE* value shown back in Figure 3.12.

How Good Is the Fit?

In the last chapter on simple regression, we saw that r^2 represents the percentage of the variation in the dependent variable that is explained by variations in the independent variable. That relationship holds in multiple regression, only now more than one independent variable is varying. In the last chapter, we simply accepted this definition of r^2. Now that we have discussed ANOVA, we are ready to see how r^2 is calculated:

Calculation of r^2

$$r^2 = \frac{SSR}{SST}$$

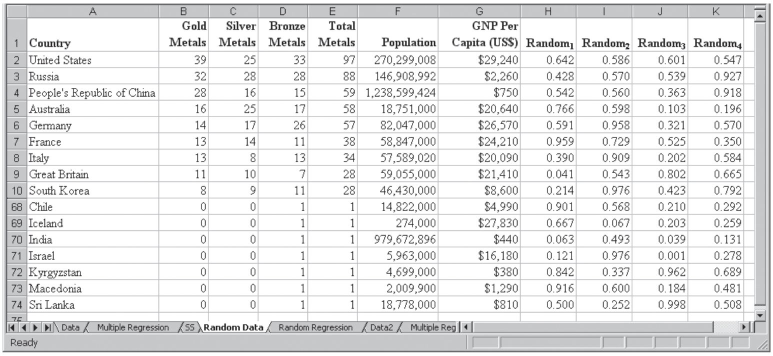
From this calculation and Figure 3.12, we have SSR = 5,360.3338 and SST = 25,475.3425. That gives us 5,360.3338/25,475.3425 = 0.2104. This is exactly the value shown in Figure 3.12 for r^2.

Notice that the numerator of this formula for r^2 is the variation explained by regression and the denominator is the total variation. Thus this formula is the ratio of explained variation to total variation. In other words, it calculates the percentage of explained variation to total variation.

The value of r^2 suffers from a problem when variables are added. We will illustrate this problem with an example.

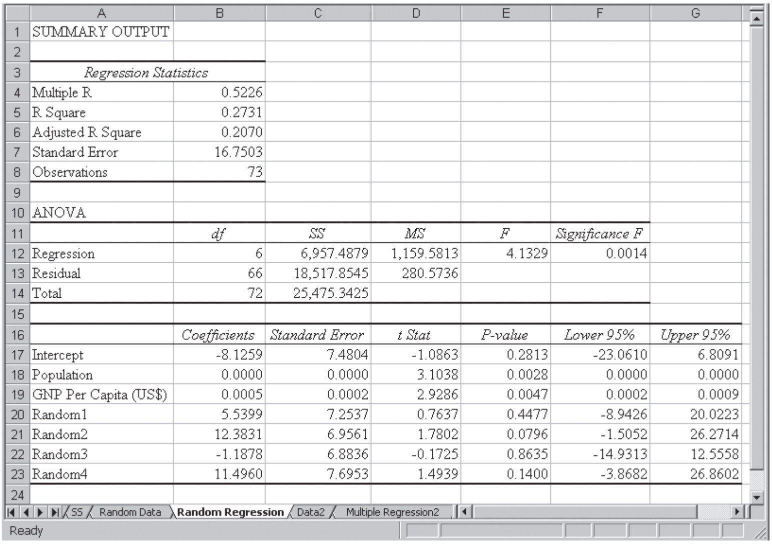
Example

Figure 3.14 shows the data from the 2000 Sydney Olympics with four variables added. Each of these variables was added by using the Excel random number generator. These random numbers were then converted from a formula (=RAND()) to a hardwired number[9] so their value would not change while the regression was being calculated and so you could experiment with the same set of random numbers. Although the numbers are still random, they can just no longer vary. Because these numbers were randomly generated, they should not help to explain the results at



	A	B	C	D	E	F	G	H	I	J	K
		Gold	Silver	Bronze	Total		GNP Per				
1	Country	Metals	Metals	Metals	Metals	Population	Capita (US$)	Random_1	Random_2	Random_3	Random_4
2	United States	39	25	33	97	270,299,008	$29,240	0.642	0.586	0.601	0.547
3	Russia	32	28	28	88	146,908,992	$2,260	0.428	0.570	0.539	0.927
4	People's Republic of China	28	16	15	59	1,238,599,424	$750	0.542	0.560	0.363	0.918
5	Australia	16	25	17	58	18,751,000	$20,640	0.766	0.598	0.103	0.196
6	Germany	14	17	26	57	82,047,000	$26,570	0.591	0.958	0.321	0.570
7	France	13	14	11	38	58,847,000	$24,210	0.959	0.729	0.525	0.350
8	Italy	13	8	13	34	57,589,020	$20,090	0.390	0.909	0.202	0.584
9	Great Britain	11	10	7	28	59,055,000	$21,410	0.041	0.543	0.802	0.665
10	South Korea	8	9	11	28	46,430,000	$8,600	0.214	0.976	0.423	0.792
68	Chile	0	0	1	1	14,822,000	$4,990	0.901	0.568	0.210	0.292
69	Iceland	0	0	1	1	274,000	$27,830	0.667	0.067	0.203	0.259
70	India	0	0	1	1	979,672,896	$440	0.063	0.493	0.039	0.131
71	Israel	0	0	1	1	5,963,000	$16,180	0.121	0.976	0.001	0.278
72	Kyrgyzstan	0	0	1	1	4,699,000	$380	0.842	0.337	0.962	0.689
73	Macedonia	0	0	1	1	2,009,900	$1,290	0.916	0.600	0.184	0.481
74	Sri Lanka	0	0	1	1	18,778,000	$810	0.500	0.252	0.998	0.508

Figure 3.14 The data for the 2000 Sydney Olympics with four completely random variables added



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.5226					
5	R Square	0.2731					
6	Adjusted R Square	0.2070					
7	Standard Error	16.7503					
8	Observations	73					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	6	6,957.4879	1,159.5813	4.1329	0.0014	
13	Residual	66	18,517.8545	280.5736			
14	Total	72	25,475.3425				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-8.1259	7.4804	-1.0863	0.2813	-23.0610	6.8091
18	Population	0.0000	0.0000	3.1038	0.0028	0.0000	0.0000
19	GNP Per Capita (US$)	0.0005	0.0002	2.9286	0.0047	0.0002	0.0009
20	Random1	5.5399	7.2537	0.7637	0.4477	-8.9426	20.0223
21	Random2	12.3831	6.9561	1.7802	0.0796	-1.5052	26.2714
22	Random3	-1.1878	6.8836	-0.1725	0.8635	-14.9313	12.5558
23	Random4	11.4960	7.6953	1.4939	0.1400	-3.8682	26.8602
24							

Figure 3.15 The results of running multiple regression on the 2000 Sydney Olympics data with four completely random variables added

all. That is, they have absolutely no relationship to the number of medals won. As a result, you would expect that the percentage of variation explained (r^2) would also not change.

Figure 3.15 shows the results of the new multiple regression including the four random variables. As expected, none of these variables is significant. However, the r^2 value goes up from 0.2104 when the regression was run without the random variables to 0.2731 when the random variables are included. Why?

With simple regression, the model is a line. A line is uniquely defined by two points, so simple regression on two points will always perfectly define a line. This is called a *sample-specific* solution. That is, r^2 will always be 1.00 even if the two variables have nothing to do with each other. For any regression with k independent variables, a model with k + 1 observations will be uniquely defined with r^2 of 1.00. Additionally, as the number of variables increases, even when those variables have no useful information, the value of r^2 will always increase. That is why r^2 increased in the previous example when the four random, and useless, variables were added. If we were to add more variables containing purely random data, r^2 would go up again. The opposite is also true. If you drop a variable from a model, even a variable like these random numbers that have no relationship to the dependent variable, r^2 will go down.

To account for r^2 increasing every time a new variable is added, an alternative measure of fit, called the adjusted multiple coefficient of determination, or adjusted r^2, is also computed using the following formula:

Multiple Coefficient of Determination, or Adjusted r^2

$$\bar{R}^2 = 1 - \frac{SSE/(n-(k+1))}{SST/(n-1)}$$

As with r^2, the symbol $\bar{R}^2$ is for the population value and $\bar{r}^2$ is for the sample. Because this symbol is not used in computer printouts, we will just use adjusted r^2. Although r^2 always increases when another variable is added, the adjusted r^2 does not *always* increase because it takes into account changes in the degrees of freedom. However, the adjusted r^2 may also increase when unrelated variables are added, as it did in the previous example, increasing from 0.1879 to 0.2070. When the adjusted r^2 does increase with the addition of unrelated variables, its level of increase will generally be much less than r^2, as was the case here.

In certain rare cases, it is possible for the adjusted r^2 to be negative. The reason for this is explained in the sidebar in more detail, but it happens only when there is little or no relationship between the dependent variable and the independent variables.

Box 3.1

Negative Adjusted r^2

The value of r^2 must always be positive. There is no surprise there. In simple regression, r^2 is simply the correlation coefficient (r) squared, and any value squared must be positive. In multiple regression, r^2 is calculated with the following formula:

Calculation of r^2

$$r^2 = \frac{SSR}{SST}$$

Because both *SSR* and *SST* are always positive, it follows that r^2 is also always positive.

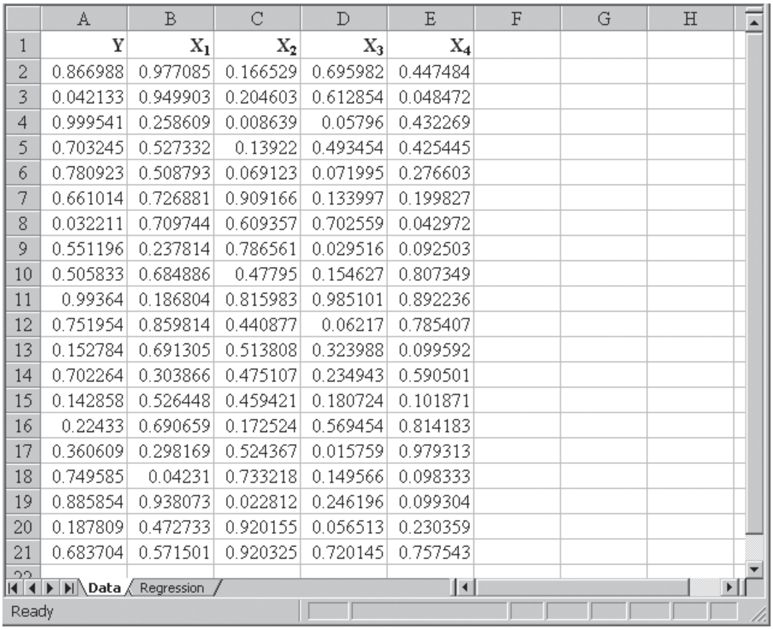
As we saw earlier, r^2 can be adjusted for the number of variables, producing what was called adjusted r^2, using the following formula:

Multiple Coefficient of Determination, or Adjusted r^2

$$\bar{R}^2 = 1 - \frac{SSE \big/ (n-(k+1))}{SST \big/ (n-1)}$$

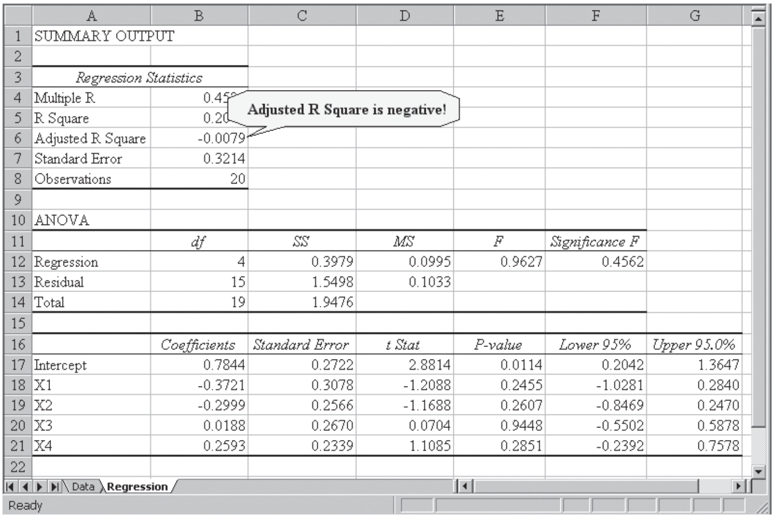
One would expect that adjusted r^2 would also be positive and, for the most part, that is the case. However, almost by accident, the author noticed that adjusted r^2 can sometimes be negative.

In developing an example of what would happen to r^2 when there was no relationship between the variables (r^2 is low but not zero), the author put together a spreadsheet where the data were generated using the Excel random number function. It looked much like the data set shown in Figure 3.16. Although the data were generated with the random number function, they were converted to fixed values. Otherwise,



	A	B	C	D	E	F	G	H
1	Y	X_1	X_2	X_3	X_4			
2	0.866988	0.977085	0.166529	0.695982	0.447484			
3	0.042133	0.949903	0.204603	0.612854	0.048472			
4	0.999541	0.258609	0.008639	0.05796	0.432269			
5	0.703245	0.527332	0.13922	0.493454	0.425445			
6	0.780923	0.508793	0.069123	0.071995	0.276603			
7	0.661014	0.726881	0.909166	0.133997	0.199827			
8	0.032211	0.709744	0.609357	0.702559	0.042972			
9	0.551196	0.237814	0.786561	0.029516	0.092503			
10	0.505833	0.684886	0.47795	0.154627	0.807349			
11	0.99364	0.186804	0.815983	0.985101	0.892236			
12	0.751954	0.859814	0.440877	0.06217	0.785407			
13	0.152784	0.691305	0.513808	0.323988	0.099592			
14	0.702264	0.303866	0.475107	0.234943	0.590501			
15	0.142858	0.526448	0.459421	0.180724	0.101871			
16	0.22433	0.690659	0.172524	0.569454	0.814183			
17	0.360609	0.298169	0.524367	0.015759	0.979313			
18	0.749585	0.04231	0.733218	0.149566	0.098333			
19	0.885854	0.938073	0.022812	0.246196	0.099304			
20	0.187809	0.472733	0.920155	0.056513	0.230359			
21	0.683704	0.571501	0.920325	0.720145	0.757543			

Figure 3.16 Random data



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.45					
5	R Square	0.20					
6	Adjusted R Square	-0.0079					
7	Standard Error	0.3214					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	0.3979	0.0995	0.9627	0.4562	
13	Residual	15	1.5498	0.1033			
14	Total	19	1.9476				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95.0%*
17	Intercept	0.7844	0.2722	2.8814	0.0114	0.2042	1.3647
18	X1	-0.3721	0.3078	-1.2088	0.2455	-1.0281	0.2840
19	X2	-0.2999	0.2566	-1.1688	0.2607	-0.8469	0.2470
20	X3	0.0188	0.2670	0.0704	0.9448	-0.5502	0.5878
21	X4	0.2593	0.2339	1.1085	0.2851	-0.2392	0.7578
22							

Adjusted R Square is negative!

Figure 3.17 The results of running multiple regression on the random data

the data would change each time the worksheet was loaded and so the results would not match the data.

When these data were used in regression, the r^2 value was low (0.2043), as expected, and the regression was insignificant, again as expected. These results are shown in Figure 3.17. What was unexpected was the adjusted r^2 value of –0.0079. At first, the author suspected a bug in Excel, but after doing some research, it became clear that Excel was working properly.

To see this, we will rewrite the equation for adjusted r^2 shown previously:

Rewriting Adjusted r^2 Formula

$$\overline{R}^2 = 1 - \left(\frac{SSE}{SST} \cdot \frac{n-1}{n-k-1} \right)$$

Because $SSR = 1 - SSE$ and $r^2 = SST/SSR$, we can rewrite this equation in terms of r^2:

Simplifying Equation

$$\overline{R}^2 = 1 - A \cdot \left(1 - r^2\right)$$

where

$$A = \frac{n-1}{n-k-1}$$

Notice that A is greater than one any time k (number of independent variables) is greater than zero. If r^2 were zero, then our equation would reduce to the following:

Equation When r^2 Is Zero

$$\text{Adjusted } r^2 = 1 - A(1 - 0) = 1 - A$$

Because A is greater than one for any regression run, adjusted r^2 must be negative for any regression run with $r^2 = 0$. Additionally, as r^2 increases, the chance of $A(1 - r^2)$ being greater than one slowly decreases. Thus adjusted r^2 can be negative only for very low values of r^2.

We can see this in the previous example. Here, $n = 20$, $k = 4$, and $r^2 = 0.2043$. That results in $A = 19/15 = 1.2667$ and $1.2667(1 - 0.2043) = 1.0079$. When we subtract 1.0079 from 1, we obtain the negative adjusted r^2 of $-\ 0.0079$.

When there are a large number of observations relative to the number of variables, the values of r^2 and adjusted r^2 will be close to one another. As a general rule of thumb, we recommend a bare minimum of 5 observations for each independent variable in a multiple regression model with 10 per independent variable being much better.

Testing the Significance of the Individual Variables in the Model

So far, we have discussed our regression models in general terms, and we have only been concerned with whether the overall model is significant—that is, if it passes the F-test. As was discussed previously, passing the F-test only tells us that the overall model is significant, and if there are collinear variables, then one of the variables is significant, though *not* that all of them are significant. In other words, once one of these collinear variables is dropped out, at least one variable will be significant. We now need to explore how to test the individual b_i coefficients.

This test was not required with simple regression because there was only one independent variable, so if the overall model was significant, that one independent variable must be significant. However, multiple regression has two or more independent variables and the overall model will be significant if only one of them is significant.[10] Because we only want significant variables in our final model, we need a way to identify insignificant variables so they can be discarded from the model.

The test of significance will need to be carried out for each independent variable using the following hypotheses:

$$\begin{aligned} H_0 &: \beta_k = 0 \\ H_1 &: \beta_k \neq 0 \end{aligned}$$

Of course, when there is reason to believe that the coefficient should behave in a predetermined fashion, a one-tailed test can be used rather than a two-tailed test. As always, the selection of a one-tailed or two-tailed test is theory based. Before discussing how to perform this hypothesis test on the slope coefficient for each variable, we need to discuss several problems with the test.

Interdependence

All the regression slope estimates come from a common data set. For this reason, the estimates of the individual slope coefficients are interdependent. Each individual test is then carried out at a common alpha value, say $\alpha = 0.05$. However, due to interdependence, the overall alpha value for the individual tests, as a set, cannot be determined.

Multicollinearity

In multiple regression, we want—in fact we need—each independent variable to have a strong and significant correlation with the dependent variable. However, one of the assumptions of multiple regression is that the independent variables are not correlated *with each other*. When this assumption is violated, the condition is called multicollinearity. Multicollinearity will be discussed in more detail later. For now, it is enough to know that the presence of multicollinearity causes the independent variables to rob one another of their explanatory ability. When a significant variable has its explanatory ability reduced by another variable, it may test as insignificant even though it may well have significant explanatory abilities and even if it is an important variable from a theoretical perspective.

Autocorrelation

One of the assumptions of multiple regression is that the error or residual terms are not correlated with themselves. When this assumption is violated, the condition is called autocorrelation.[11] Autocorrelation can only occur when the data are time-series data—that is, measurements of the same variables at varying points in time. More will be said about autocorrelation later. For now, autocorrelation causes some variables to appear to be more significant than they really are, raising the chance of rejecting the null hypothesis.

Repeated-Measures Test

The problem of repeated measures is only a major concern when a large number of variables need to be tested. Alpha represents the percentage of times that a true null hypothesis (that the variable is insignificant when used with regression) will be rejected just due to sampling error. At $\alpha = 0.05$, we have a 5 percent chance of rejecting a true null hypothesis ($\beta_k = 0$) just due to sampling error. When there are only a few variables, we need not be overly concerned, but it is not uncommon for regression models to have a large number of variables. The author constructed a regression model for his dissertation with over 200 variables. At $\alpha = 0.05$, on average, this model could be expected to reject as many as 10 true null hypotheses just due to sampling error.[12]

Recall the following hypotheses:

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

However, a one-tailed test could be performed if desired. The test statistic is distributed according to the Student *t*-distribution with $n - (k + 1)$ degrees of freedom. The test statistic is represented as follows:

Test Statistic for Individual Multiple Regression Slope Parameters

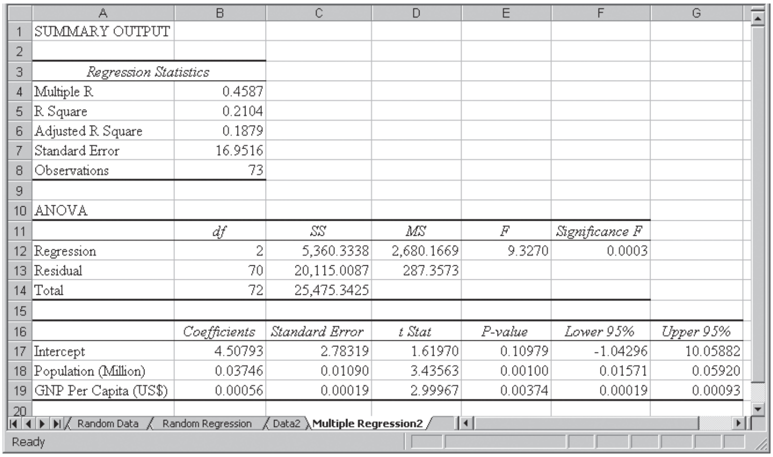
$$t = \frac{b_i}{s(b_i)}$$

where $s(b_i)$ is an estimate of the population standard deviation of the estimator $s(b_i)$. The population parameter is unknown to us, naturally, and the calculation of the sample value is beyond the scope of this textbook. However, the value $s(b_i)$ is calculated by Excel and shown in the printout.

2000 Sydney Olympics Example

Figure 3.18 shows the ANOVA from the 2000 Sydney Olympics data set without the random numbers. This time, the population values have been divided by 1,000,000 to lower the units shown in the results. This is a linear transformation so only the slope coefficient and the values that build off of it are affected. You can see this by comparing Figure 3.18 to Figure 3.9. In general, any linear transformation will leave the impact of the variables intact. Specifically, the linear transformation will not change the significance of the variable or the estimate of Y made by the resulting model.

For the Population variable, the coefficient is 0.03746 and $s(b_1)$ is 0.01090 so the test statistic is 0.03746/0.01090 = 3.43563. This value is shown as t Stat in Figure 3.18. For the per capita GNP variable, the coefficient is 0.00056 and $s(b_2)$ is 0.00019 so the test statistic is 0.00056/0.00019 = 2.99967, which is also shown in Figure 3.18. For a



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.4587					
5	R Square	0.2104					
6	Adjusted R Square	0.1879					
7	Standard Error	16.9516					
8	Observations	73					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	5,360.3338	2,680.1669	9.3270	0.0003	
13	Residual	70	20,115.0087	287.3573			
14	Total	72	25,475.3425				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	4.50793	2.78319	1.61970	0.10979	-1.04296	10.05882
18	Population (Million)	0.03746	0.01090	3.43563	0.00100	0.01571	0.05920
19	GNP Per Capita (US$)	0.00056	0.00019	2.99967	0.00374	0.00019	0.00093
20							

Figure 3.18 The ANOVA from the 2000 Sydney Olympics data set without the random numbers

two-tailed test, the *p*-value is shown in Figure 3.18, and the null hypothesis can be accepted or rejected based solely on this value.

For a one-tailed hypothesis test, you can double the *p*-value and compare that value to alpha.[13] However, you must be careful when you do this as it is possible to reach the wrong conclusion. To be certain, for a one-tailed test, the calculated test statistic should be compared with the critical value.

In this example, we would expect that having a larger population would give you a larger pool of athletes from which to select Olympic athletes and so would increase the number of medals won. In other words, we would use the following hypotheses:

$$H_0 : \beta_1 \leq 0$$
$$H_1 : \beta_1 > 0$$

Likewise, having a higher per capita GNP should lead to more money to spend on athlete training. In other words, we could use these hypotheses:

$$H_0 : \beta_2 \leq 0$$
$$H_1 : \beta_2 > 0$$

So both of these hypothesis tests should be performed as a one-tailed right test. For a one-tailed right test with $\alpha = 0.05$ and 70 degrees of freedom, the critical Student *t*-value is 1.66692. Both the test statistic of 3.43563 for population and 2.99967 for per capita GNP exceed this value, so both slope coefficients are significant.

Forbes Celebrity Example

Looking back at Figure 3.10, we see the regression results for multiple regression on the top 100 celebrities in 2000 from *Forbes*. The dependent variable is *Forbes*'s power rank. The following are the independent variables:

- Income rank (β_1)
- Earnings for 1999 (β_2)

- Web hits across the Internet (β_3)
- Press clips from Lexis-Nexis (β_4)
- Magazine covers (β_5)
- Number of mentions on radio and television (β_6)

For income rank, 1 is highest and 100 is lowest, so we would expect that the *lower* the number, the higher the power ranking. For earnings, web hits, press clips, and magazine covers, we would expect that a higher value would indicate a higher power ranking. For the number of times mentioned on radio and television, the relationship is not clear. Being mentioned more could mean they are accomplishing good things and raising their power ranking, or it could mean they are involved in a scandal, which probably lowers their power ranking. Because the direction is unclear, we will use a two-tailed test for this variable. The six sets of hypotheses we need to test are therefore represented in Table 3.1.

With 93 degrees of freedom, the Student *t* critical value for a one-tailed left test is –1.66140, for a one-tailed right it is +1.66140, and for a two-tailed test it is ±1.9858. Given that, the critical values and decisions for the six variables are given in Table 3.2.

Note that had we used a two-tailed test for all the variables, only β_2 would have been insignificant. This is most likely due to our poor understanding of the relationship between these variables and the power of these celebrities. It is not uncommon for the results of regression to cause researchers to rethink their theory. In any case, hypotheses should never be redone just to make a variable significant. Rather, they should only be changed when there is good reason to believe that the original theory is flawed.

Because it is unlikely that *Forbes* went to all the trouble to collect and report this data without then including it in its power ranking, we are willing to believe that our original theory regarding the hypotheses was wrong. Given that, we conclude that we do not know enough to set a direction for the hypotheses and will use a two-tailed test for all variables. Those results are shown in Table 3.3.

Note that income is still insignificant, but all the other variables are significant. This is most likely due to income ranking and income explaining the same variation and so income ranking is robbing the explanatory

Table 3.1 Hypotheses Used

Income Rank	Earnings for 1999	Web Hits
$H_0: \beta_1 \geq 0$	$H_0: \beta_2 \leq 0$	$H_0: \beta_3 \leq 0$
$H_1: \beta_1 < 0$	$H_1: \beta_2 > 0$	$H_1: \beta_3 > 0$
Press Clips	**Magazine Covers**	**Radio and TV Mentions**
$H_0: \beta_4 \leq 0$	$H_0: \beta_5 \leq 0$	$H_0: \beta_6 = 0$
$H_1: \beta_4 > 0$	$H_1: \beta_5 > 0$	$H_1: \beta_6 \neq 0$

Table 3.2. Hypothesis Test Results

Income Rank	Earnings for 1999	Web Hits
$H_0: \beta_1 \geq 0$	$H_0: \beta_2 \leq 0$	$H_0: \beta_3 \leq 0$
$H_1: \beta_1 < 0$	$H_1: \beta_2 > 0$	$H_1: \beta_3 > 0$
Critical value: –1.66140	Critical value: +1.66140	Critical value: +1.66140
Test statistics: 20.3590	Test Statistics: 0.2022	Test statistics: –2.9901
Decision: Accept	Decision: Accept	Decision: Accept
Press Clips	**Magazine Covers**	**Radio and TV Mentions**
$H_0: \beta_4 \leq 0$	$H_0: \beta_5 \leq 0$	$H_0: \beta_6 = 0$
$H_1: \beta_4 > 0$	$H_1: \beta_5 > 0$	$H_1: \beta_6 \neq 0$
Critical Value: +1.66140	Critical Value: +1.66140	Critical Value: ±1.9858
Test Statistics: –6.0797	Test Statistics: –2.6381	Test Statistics: –3.6445
Decision: Accept	Decision: Accept	Decision: Reject

Table 3.3 Hypothesis Test Results Using All Two-Tailed Tests

Income Rank	Earnings for 1999	Web Hits
$H_0: \beta_1 = 0$	$H_0: \beta_2 = 0$	$H_0: \beta_3 = 0$
$H_1: \beta_1 \neq 0$	$H_1: \beta_2 \neq 0$	$H_1: \beta_3 \neq 0$
Critical Value: ±1.9858	Critical Value: ±1.9858	Critical Value: ±1.9858
Test Statistics: 20.3590	Test Statistics: 0.2022	Test Statistics: –2.9901
Decision: Reject	Decision: Accept	Decision: Reject
Press Clips	**Magazine Covers**	**Radio and TV Mentions**
$H_0: \beta_4 = 0$	$H_0: \beta_5 = 0$	$H_0: \beta_6 = 0$
$H_1: \beta_4 \neq 0$	$H_1: \beta_5 \neq 0$	$H_1: \beta_6 \neq 0$
Critical Value: ±1.9858	Critical Value: ±1.9858	Critical Value: ±1.9858
Test Statistics: –6.0797	Test Statistics: –2.6381	Test Statistics: –3.6445
Decision: Reject	Decision: Reject	Decision: Reject

power from income. This must be treated before we have a final model. We will revisit this again later in the chapter.

Automating Hypothesis Testing on the Individual Variables

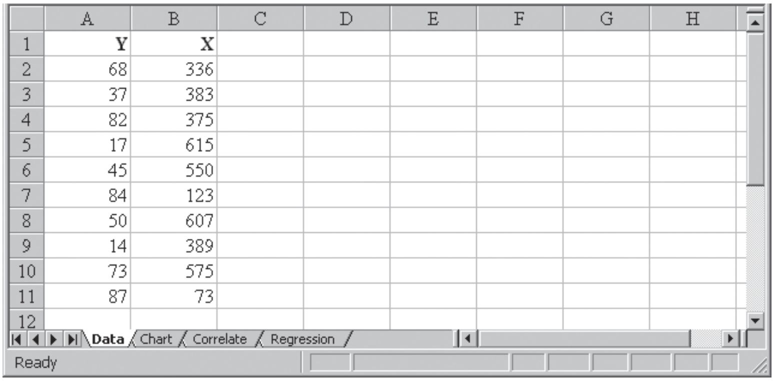
Excel provides a *p*-value for each regression coefficient that can be used to perform variable hypothesis testing, as long as it is used with care. When used carelessly, it can cause you to make the wrong decision. We will illustrate this using an example.

Example

Figure 3.19 shows a set of fictitious simple regression data. Figure 3.20 shows a chart of this data. As you can see from Figure 3.20, the data have a negative relationship. Figure 3.21 shows the resulting simple regression run.

As you can see in Figure 3.21, the overall model is not significant because the *p*-value for the *F*-test is only 0.0844. Notice that the *p*-value for the Student *t*-test on β_1 is also 0.0844. This will always be the case in simple regression, but not, however, in multiple regression.

In the last chapter, we tested the correlation coefficient to see if it was significant, so we will do the same here. Given the chart shown in Figure 3.20, we will assume that the relationship is negative. That is, we will make the following hypotheses:



	A	B	C	D	E	F	G	H
1	Y	X						
2	68	336						
3	37	383						
4	82	375						
5	17	615						
6	45	550						
7	84	123						
8	50	607						
9	14	389						
10	73	575						
11	87	73						
12								

Figure 3.19 Fictitious simple regression data

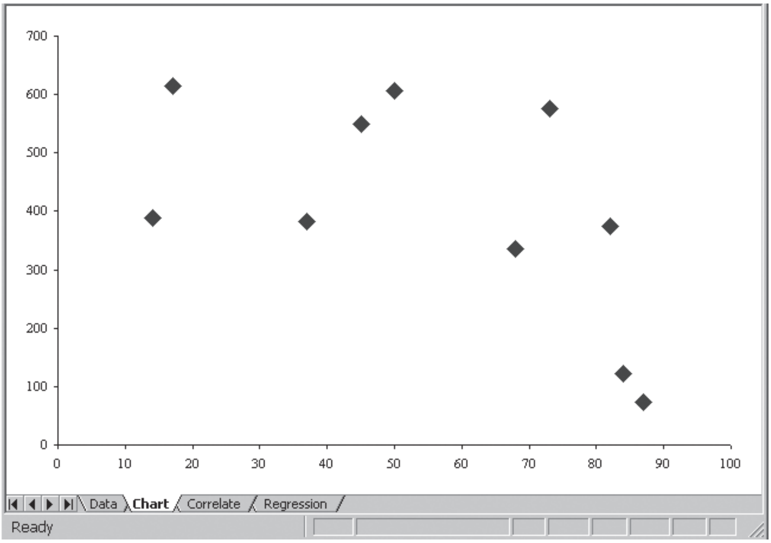
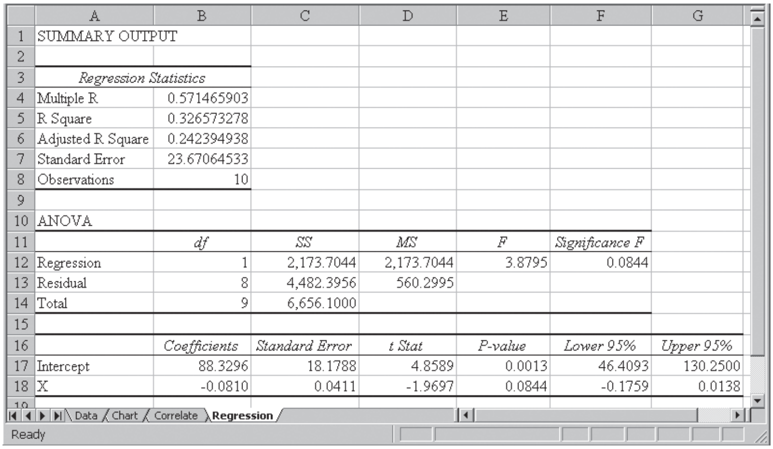


Figure 3.20 A chart of the data



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.571465903					
5	R Square	0.326573278					
6	Adjusted R Square	0.242394938					
7	Standard Error	23.67064533					
8	Observations	10					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	2,173.7044	2,173.7044	3.8795	0.0844	
13	Residual	8	4,482.3956	560.2995			
14	Total	9	6,656.1000				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	88.3296	18.1788	4.8589	0.0013	46.4093	130.2500
18	X	-0.0810	0.0411	-1.9697	0.0844	-0.1759	0.0138

Figure 3.21 The resulting simple regression run

$$H_0 : \rho \geq 0$$
$$H_1 : \rho < 0$$

The correlation coefficient (not shown) is –0.5715. This correlation coefficient is significant.

At first, you may be concerned regarding the apparent discrepancy between concluding the model was insignificant using the *F*-test and concluding it was significant by testing the correlation coefficient. However, there is no real discrepancy. The *F*-test tests the entire model at once. When the model contains many variables, it would not be uncommon for some of them to be tested as one-tailed right, others as one-tailed left, and still others as two tailed. We assume a two-tailed *F*-test in multiple regression to avoid this issue. In simple regression, we only have one variable, so we can tailor the test to fit that single variable. Note also that the *F*-test can be converted to a one-tailed test by dividing the two-tailed *p*-value by two, obtaining 0.0422 and making this model significant using the *F*-test and matching our results under the hypothesis test on the correlation coefficient.

Excel gives two-tailed *p*-values for both the *F*-test and the Student *t*-test for the individual coefficients. These can be converted to one-tailed *p*-values by dividing them by two. We generally do not do this for the *F*-test in multiple regression because all the coefficients are rarely in agreement regarding their number of tails and direction of testing. However, this is perfectly acceptable for the individual coefficients because we are testing the variables one at a time. However, we must be careful not to allow this shortcut to cause us to reach the wrong conclusion. We will continue with our example to see how this might happen.

We already know from the previous example that this model is significant when we assume a negative relationship. Now, what happens if we assume a positive relationship? That is, we assume the following:

$$H_0 : \rho \le 0$$
$$H_1 : \rho > 0$$

Therefore,

$$H_0 : \beta_1 \le 0$$
$$H_1 : \beta_1 > 0$$

Just by looking at the chart back in Figure 3.20, we know that this assumption will lead us to conclude that the relationship is insignificant. However, if we simply take the Student *t*-value for the β_1 coefficient of 0.0844 from the regression in Figure 3.21 and divide by two, we obtain 0.0422 and we therefore reject the null hypothesis and conclude the model is significant. This time, we truly have a contradiction.

Briefly, the problem is that Excel computes the test statistic and then finds the area on both sides. It takes the smaller of these two values and doubles it to compute the two-tailed *p*-value. Simply dividing by two to obtain the one-tailed *p*-value gives no consideration to which side of the mean the rejection region falls.

The way to avoid this is to compare the sign of β_1 as assumed in H_0 and b_1 as calculated by regression. When these are the same, you cannot reject the null hypothesis and conclude the variable is significant regardless of the *p*-value. After all, if the null hypothesis assumes the coefficient is negative and the sample coefficient is negative, we would never want to conclude that this assumption was false. In the previous example, the null hypothesis was that the β_1 was negative and b_1 was –0.0810. Because these have the same sign, we must assume that this variable is not significant.

Conclusion

In this chapter, you have seen how to perform multiple regression, how to test the overall model for significance, and how to avoid problems when testing for significance. In the next chapter, we will see how to pull this together and construct meaningful multiple regression models.

CHAPTER 4
Model Building

In business, we build regression models to accomplish something. Typically, either we wish to explain the behavior of the dependent variable or we wish to forecast the future behavior of the dependent variable based on the expected future behavior of the independent variables. Often, we wish to do both. In order to accomplish this, we select variables that we believe will help us explain or forecast the dependent variable. If the dependent variable were sales, for example, then we would select independent variables like advertising, pricing, competitor actions, the economy, and so on. We would not select independent variables like supplier lead time or corporate income tax rates because these variables are unlikely to help us explain or forecast sales. As was discussed in chapter 3, we want the overall model to be significant and we want the individual independent variables to also all be significant. In summary, our three criteria for a multiple regression model are the following:

1. Variables should make sense from a theoretical standpoint. That is, in business, it makes sense from a business perspective to include each variable.
2. The overall model is significant. That is, it passes the *F*-test.
3. Every independent variable in the model is significant. That is, each variable passes its individual Student *t*-test.

In chapter 3, we saw how to test an overall model for significance, as well as how to test the individual slope coefficients. We now need to investigate how to deal with the situation where the overall model is significant but some or all of the individual slope coefficients are insignificant.

Before we go on, you may have noticed that the previous sentence states that “where the overall model is significant but *some or all* of the individual slope coefficients are *insignificant*." Although rare, it is possible for the overall model to be significant although none of the individual slope coefficients is significant as we begin to work with the model. This can only happen when the model has a great deal of multicollinearity. When this happens, it will always be the case that the following procedures will result in at least one variable becoming significant. Stated another way, when the overall model is significant, we must end up with at least one significant slope coefficient, regardless of how they appear in the original results.

We call the process of moving from including every variable to only including those variables that provide statistical value *model building*. In business, we use these models to do something, such as produce a forecast. Oftentimes, these models are used over and over. For example, a forecasting model might be used every month. Building a model with only variables that provide statistical value has the added benefit of minimizing the cost of the data collection associated with maintaining that model.

Partial *F*-Test

One way to test the impact of dropping one or more variables from a model is with the partial *F*-test. With this test, the *F* statistic for the model with and without the variables under consideration for dropping is computed and the two values are compared. We will illustrate this using the celebrities worksheet.

Example

When we first looked at this model in chapter 3, the earnings variable (β_2) appeared to be insignificant. The Student *t*-value for magazine covers (β_5) was also low, so we will include it with earnings to see if both should be dropped. The model with all the variables included is called the *full model.* It is of the following form:

Full Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

We already have the results for this model. They were shown back in Figure 3.11. We will call the model with β_1 and b_5 dropped the *reduced model*. It is of the following form:

Reduced Model

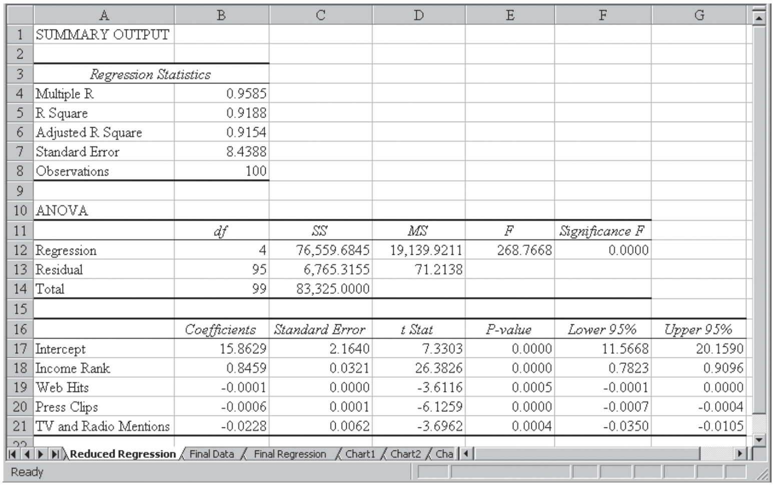
$$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_6 x_6$$

This multiple regression run is shown in Figure 4.1. In this case, the hypotheses for the partial *F*-test are the following:

$$H_0 : \beta_2 = \beta_5 = 0$$
$$H_1 : \beta_2 \neq 0 \text{ and/or } \beta_5 \neq 0$$

The partial *F* statistic is *F* distributed with $r, n - (k + 1)$ degrees of freedom where *r* is the number of variables that were dropped to create the reduced model (two in this case), *n* is the number of observations, and *k* is the number of independent variables in the full mode. With $\alpha = 0.05$ and 2,93 degrees of freedom, the value of the *F* statistic is 3.0943.

The partial *F*-test statistic is based on the sums of squares (*SSE*) for the reduced and full model and the mean square error (*MSE*) for the full model. It is computed as follows:



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9585					
5	R Square	0.9188					
6	Adjusted R Square	0.9154					
7	Standard Error	8.4388					
8	Observations	100					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	76,559.6845	19,139.9211	268.7668	0.0000	
13	Residual	95	6,765.3155	71.2138			
14	Total	99	83,325.0000				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	15.8629	2.1640	7.3303	0.0000	11.5668	20.1590
18	Income Rank	0.8459	0.0321	26.3826	0.0000	0.7823	0.9096
19	Web Hits	-0.0001	0.0000	-3.6116	0.0005	-0.0001	0.0000
20	Press Clips	-0.0006	0.0001	-6.1259	0.0000	-0.0007	-0.0004
21	TV and Radio Mentions	-0.0228	0.0062	-3.6962	0.0004	-0.0350	-0.0105

Figure 4.1 The multiple regression run for the reduced celebrity model

Partial *F*-Test Statistic

$$F = \frac{\left(SSE_R - SSE_F\right)\big/ r}{MSE_F}$$

Where SSE_R is the *SSE* for the reduced model, SSE_F is the *SSE* for the full model, *r* is the number of variables that were dropped to create the reduced model, and MSE_F is *MSE* for the full model. For the previous examples, the partial *F*-test statistic is the following:

$$F = \frac{\left(SSE_R - SSE_F\right)\big/ r}{MSE_F} = \frac{(6{,}765.3155 - 6{,}292.7211)\big/ 2}{67.6637} = \frac{236.2972}{67.66367} \approx 3.49$$

Because 3.49 is greater than the critical value of 3.0943, we reject the null hypothesis and conclude that either β_2 or β_5 or both are not zero. This was to be expected because we had already concluded that β_5 was not equal to zero.

Partial F *Approaches*

The partial *F*-test can be carried out over any combination of variables in order to arrive at a final model. As you can imagine, this would be very tedious and is usually automated by a statistical package. There are four overall approaches that a statistical package can take to select the variables to include in the final model:[1]

1. *All possible combinations of variables*. The computer simply tries all possible combinations of *k* independent variables and then picks the best one. If we are considering *k* independent variables, then there are 2^k–1 possible sets of variables. Once all the possible models are computed, the best one is selected according to some criteria, such as the highest adjusted r^2 or the lowest *MSE*.
2. *Forward selection*. The computer begins with the model containing no variables. It then adds the single variable with the highest significant *F* statistic. Once a variable has been added, the computer looks at the partial *F* statistic of adding one more variable. It then adds the

one with the highest *F* value to the model, as long as that variable meets the significance requirement (e.g., $\alpha = 0.05$). Once added, this, and all the variables that are added later, remains in the model. That is, once in, a variable is never discarded. The process continues until no more variables are available that meet the significance requirement.

3. *Backward elimination*. The computer begins with the model containing all the variables. It then computes the partial *F* statistic for dropping each single variable. It then drops the variable that has the lowest partial *F* statistic. This continues until all the variables remaining in the model meet the significance requirement. Once a variable is dropped from the model, it is never considered for reentry into the model.
4. *Stepwise regression*. This is a combination of forward selection and backward elimination. The weakness of these two approaches is that they never reevaluate a variable. Stepwise regression begins as forward selection, finding the single variable to put into the model. It then goes on to find the second variable to enter, as always, assuming it meets the significance requirement. Once a second variable enters the model, it uses backward elimination to make sure that the first variable meets the criteria to stay in the model. If not, it is dropped. Next, it uses forward selection to select the next variable to enter and then uses backward elimination to make sure that all the variables should remain. This two-step approach assures us that any interaction (multicollinearity) between variables is accounted for. The process continues until no more variables will enter or leave the model. Stepwise is the most common approach to deciding on the variables that are to remain in the model.

Forward, backward, and stepwise selection can be seen in operation in Box 4.1.

When working with a large number of variables, model building can be difficult and time-consuming with Excel. This is the situation when there is a real business case for investing in a powerful statistical package like SPSS.

Box 4.1

Forward, Backward, and Stepwise Regression in SPSS

The following shows the results of using a statistical package called SPSS to perform forward, backward, and stepwise regression on the data in the SellingPrice.xls worksheet. This worksheet is discussed in more detail a little later in the chapter.

Forward Selection

With this approach, SPSS begins with the model containing no variables. It then adds the single variable with the highest significant *F* statistic.

Variables Entered and Removed

Its first report, shown below, shows the variables that have entered the model and the order in which they entered. In this case, the first variable to enter was *Asking Price* and the second to enter was *Time on Market*.

Model	Variables Entered	Method
1	Asking Price	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	Time on Market	Forward (Criterion: Probability-of-F-to-enter < = .050)

Model Summary

Its next report, shown here, summarizes each model as it is being built. The report shows the *r*, r^2, adjusted r^2, and standard error of the estimate for each model.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.992	.985	.984	$4,600.24
2	.995	.990	.989	$3,841.44

ANOVA

Its next report, shown here, shows the ANOVA table for each model as it is being built.

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37,637,951,771	1	37,637,951,771	1,779	.000
	Residual	592,542,895	28	21,162,246		
	Total	38,230,494,666	29			
2	Regression	37,832,064,683	2	18,916,032,342	1,282	.000
	Residual	398,429,983	27	14,756,666		
	Total	38,230,494,666	29			

Coefficients

The next report shows the coefficients for each model as the model is being built.

		Unstandardized Coefficients		*Standardized Coefficients*		
Model		**B**	**Std. Error**	**Beta**	**t**	**Sig.**
1	(Constant)	8,832.825	3160.942		2.794	.009
	Asking Price	.876	.021	.992	42.173	.000
2	(Constant)	11,536.825	2742.819		4.206	.000
	Asking Price	.888	.018	1.006	50.244	.001
	Time on Market	–273.687	75.461	–.073	–3.627	.001

The final model is represented in the following equation:

$$\text{Price} = \$11{,}536.825 + (0.888 \times \text{Asking Price}) \\ -(273.687 \times \text{Time on Market})$$

Excluded Variables

The final report shows the variables that were excluded from each model.

						Collinearity Statistics
Model		**Beta In**	**t**	**Sig.**	**Partial Correlation**	**Tolerance**
1	Bedrooms	–.088	–1.617	.117	–.297	.178
	Bathrooms	–.081	–2.035	.052	–.365	.316
	Square Feet	–.088	–1.282	.211	–.239	.116
	Age	–.039	–1.576	.127	–.290	.863
	Time on Market	–.073	–3.627	.001	–.572	.962

						Collinearity Statistics
Model		**Beta In**	**t**	**Sig.**	**Partial Correlation**	**Tolerance**
2	Bedrooms	–.009	–.166	.869	–.033	.137
	Bathrooms	–.028	–.712	.483	–.138	.249
	Square Feet	–.019	–.306	.762	–.060	.103
	Age	–.024	–1.126	.270	–.216	.826

Backward Selection

With this approach, SPSS begins with the model containing all the variables. It then computes the partial *F*-statistic for dropping each single variable. It then drops the variable that has the lowest partial *F*-statistic. This continues until all the variables remaining in the model meet the significance requirement. The reports on this method, using the same data, are shown on the pages that follow.

Variables Entered/Removed

Model	**Variables Entered**	**Variables Removed**	**Method**
1	Asking Price, Time on Market, Age, Bathrooms, Bedrooms, Square Feet		Enter
2		Asking Price	Backward (Criterion: Probability of F-to- remove > = .100).
3		Bathrooms	Backward (Criterion: Probability of F-to- remove > = .100).

Model Summary

Model	**R**	**R Square**	**Adjusted R Square**	**Std. Error of the Estimate**
1	.997	.993	.991	$3,401.22
2	.996	.993	.991	$3,416.73
3	.996	.992	.991	$3,509.24

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37,964,423,478	6	6,327,403,913	546.960	.000
	Residual	266,071,189	23	11,568,313		
	Total	38,230,494,667	29			
2	Regression	37,950,317,903	5	7,590,063,581	650.166	.000
	Residual	280,176,764	24	11,674,032		
	Total	38,230,494,667	29			
3	Regression	37,922,626,237	4	9,480,656,559	769.863	.000
	Residual	307,868,430	25	12,314,737		
	Total	38,230,494,667	29			

Coefficients

		Unstandardized Coefficients		*Standardized Coefficients*		
Model		**B**	**Std. Error**	**Beta**	**t**	**Sig.**
1	(Constant)	49,579.707	12046.503		4.116	.000
	Bedrooms	13,488.252	4515.791	.448	2.987	.007
	Bathrooms	–6,202.340	3956.130	–.135	–1.568	.131
	Square Feet	16.722	6.990	.386	2.392	.025
	Age	–2,104.376	727.050	–.191	–2.894	.008
	Time on Market	–517.636	128.547	–.137	–4.027	.001
	Asking Price	.257	.232	.291	1.104	.281
2	(Constant)	61,929.842	4495.695		13.775	.000
	Bedrooms	17,001.038	3219.658	.565	5.280	.000
	Bathrooms	–6,119.732	3973.455	–.133	–1.540	.137
	Square Feet	24.197	1.748	.559	13.844	.000
	Age	–2,877.174	197.901	–.261	–14.538	.000
	Time on Market	–633.994	73.960	–.168	–8.572	.000
3	(Constant)	56,174.704	2567.304		21.881	.000
	Bedrooms	12,322.397	1095.710	.409	11.246	.000
	Square Feet	25.617	1.525	.592	16.798	.000
	Age	–2,906.610	202.309	–.264	–14.367	.000
	Time on Market	–6,58.871	74.129	–.175	–8.888	.000

In this case, the final model is represented in the following equation:

$$\begin{aligned}\text{Selling Price} = {} & 56{,}174.704 + 12{,}322.397(\text{Bedrooms}) \\ & + 25.617(\text{Square Feet}) - 2{,}906.610(\text{Age}) \\ & - 6{,}58.871(\text{Time on Market})\end{aligned}$$

This is the same model we will end up developing when we approach this problem using Excel. This is to be expected, as the approach we will be using closely mirrors backward selection. Note that this model is very different from the model that was built using forward selection.

Excluded Variables

						Collinearity Statistics
Model		**Beta In**	**t**	**Sig.**	**Partial Correlation**	**Tolerance**
2	Asking Price	.291	1.104	.281	.224	4.362E-03
3	Asking Price	.283	1.044	.307	.208	4.364E-03
	Bathrooms	–.133	–1.540	.137	–.300	4.070E-02

Stepwise Regression

Using this approach, SPSS finds the single variable to put into the model. It then goes on to find the second variable to enter, as always, assuming it meets the significance requirement. Once a second variable enters the model, it uses backward elimination to make sure that the first variable meets the criteria to stay in the model. If not, it is dropped. Next, it uses forward selection to select the next variable to enter and then uses backward elimination to make sure that all the variables should remain.

Variables Entered/Removed

Model	**Variables Entered**	**Variables Removed**	**Method**
1	Asking Price		Stepwise (Criteria: Probability-of-F-to-enter < = .050, Probability-of-F-to-remove < = .100).
2	Time on Market		Stepwise (Criteria: Probability-of-F-to-enter < = .050, Probability-of-F-to-remove < = .100).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.992	.985	.984	$4,600.24
2	.995	.990	.989	$3,841.44

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	37,637,951,771	1	37,637,951,771	1,778.542	.000
	Residual	592,542,895	28	21,162,246		
	Total	38,230,494,667	29			
2	Regression	37,832,064,683	2	18,916,032,342	1,281.864	.000
	Residual	398,429,983	27	14,756,666		
	Total	38,230,494,667	29			

Coefficients

		Unstandardized Coefficients		*Standardized Coefficients*		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	8,832.825	3160.942		2.794	.009
	Asking Price	.876	.021	.992	42.173	.000
2	(Constant)	11,536.825	2742.819		4.206	.000
	Asking Price	.888	.018	1.006	50.244	.000
	Time on Market	–273.687	75.461	–.073	–3.627	.001

Using stepwise regression, the model developed matches the forward selection method. This is, of course, not always the case.

Excluded Variables

						Collinearity Statistics
Model		Beta In	t	Sig.	Partial Correlation	Tolerance
1	Bedrooms	–.088	–1.617	.117	–.297	.178
	Bathrooms	–.081	–2.035	.052	–.365	.316
	Square Feet	–.088	–1.282	.211	–.239	.116

						Collinearity Statistics
Model		**Beta In**	**t**	**Sig.**	**Partial Correlation**	**Tolerance**
	Age	–.039	–1.576	.127	–.290	.863
	Time on Market	–.073	–3.627	.001	–.572	.962
2	Bedrooms	–.009	–.166	.869	–.033	.137
	Bathrooms	–.028	–.712	.483	–.138	.249
	Square Feet	–.019	–.306	.762	–.060	.103
	Age	–.024	–1.126	.270	–.216	.826

Summary

The advantage of using any of these approaches in SPSS, rather than building the model manually in Excel, is that SPSS completely automates the process. You simply select the approach to use and SPSS does everything else for you.

None of these four methods for building a model guarantees that we will find the one best model. Because the order of testing can make a difference, it is possible, though not likely, that changing the order in which the variables are entered into the computer will change the results.

Model Building Using Excel

Unfortunately, Excel is not able to automate, or even easily perform, any of the four model-building approaches discussed previously. For this reason, it is always best to work with a dedicated statistics package when trying to develop complex multiple regression models. However, not everyone has access to a dedicated statistics package, as they can be expensive and difficult to use.

To compensate for Excel's inability to automate building regression models, the author has developed an approach to model building in Excel that approximates backward elimination only using the r^2 value that Excel displays rather than the partial F that we would otherwise have to manually compute for each regression run. For models with fewer numbers of observations, the adjusted r^2 can be used in place of r^2 in making the decisions. This approach has been found to work well on actual data, but its application can be long and tedious when a model has more than a couple of insignificant variables. To make matters worse, none of the steps can be automated. To make matters worse still, the Excel requirement of having the independent variables in contiguous columns causes a good deal of data manipulation problems. For these reasons, readers with complex regression problems are again encouraged to use a statistical program like SPSS or SAS to tackle these more complex problems.

The steps for model building in Excel are as follows:

1. Run the regression with all the variables in the model. If the overall model is insignificant, then stop. When this happens, none of the individual variables will be significant, so there is no point continuing. This is not caused by any violation of regression assumptions. This is usually caused by an error in the theory used to select the variables or, less likely, by a problem with the data, such as having outliers in the data.
2. Test each slope coefficient using the Student t-test. If all the slope coefficients are significant, then stop; you have the final version of the model.
3. If some of the variables are insignificant, make a list of all the insignificant slope coefficients.
4. One at a time, drop a single variable from this list and rerun the multiple regression. Record the resulting r^2 and then reinsert the variable into the data set.[2]
5. Once you have dropped all the insignificant variables one at a time, look at all the r^2 values you have recorded. For the run with the highest r^2, permanently drop that variable from the data set. That is, drop the variable whose absence causes the least reduction in explanatory power. This variable will be dropped forever and will not be reconsidered for reentry into the model. To do otherwise would be to greatly overwork the problem.

6. Rerun the multiple regression without the variable permanently dropped in step 5. If all the variables are significant, then stop; you are finished. If not, then return to step 2 and continue until all the slope coefficients are significant. The steps for model building in Excel are illustrated visually by the flowchart on page 117.

Selling Price Example

Understanding the real estate market is a common use of regression, a use we will explore in the following example. This example also illustrates using Excel to perform model building. The worksheet SellingPrice.xls contains fictitious data where the dependent variable is the selling price of a house and the independent variables are the following:

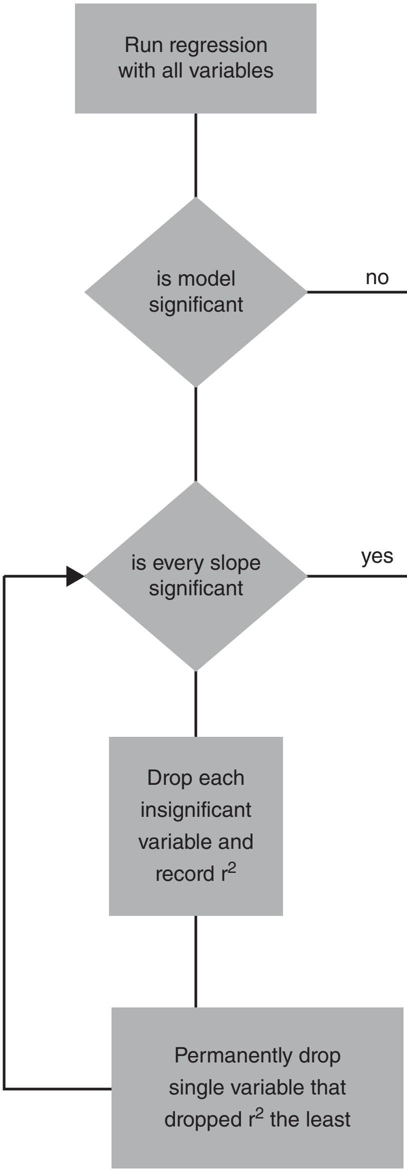
- The number of bedrooms
- The number of bathrooms
- The size of the house in square feet
- The age of the house in years
- The time the house was on the market before being sold, in months
- The initial asking price

This is the same data set that was used to illustrate various approaches to multiple regression using SPSS in Box 4.1. The data are shown in Figure 4.2 and the initial regression run is shown in Figure 4.3.

To simplify matters and allow us to use the *p*-value for all the slope coefficients, we will assume that all the slope coefficient hypothesis tests are two-tailed tests. Using these criteria, Figure 4.3 shows us that the following variables are not significant:

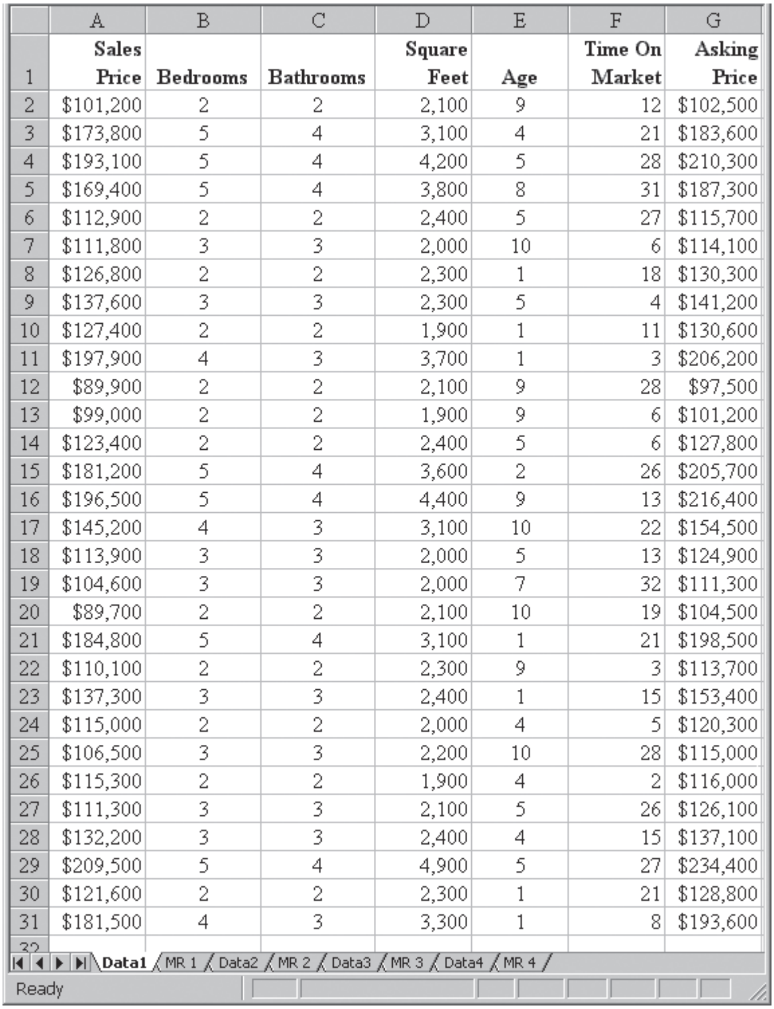
- The number of bathrooms
- The initial asking price

So we must drop each variable, in turn, and record the resulting value of r^2. Rerunning the regression and dropping number of bathrooms shows us the problems of using Excel to perform more than one multiple


Run regression with all variables
is model significant
no
stop
is every slope significant
yes
stop model done
Drop each insignificant variable and record r^2
Permanently drop single variable that dropped r^2 the least

regression run on the same data. Specifically, there are two problems: one more serious and one less so.

The more serious problem is that Excel requires independent variables to be in contiguous columns and the number of bathrooms is in the

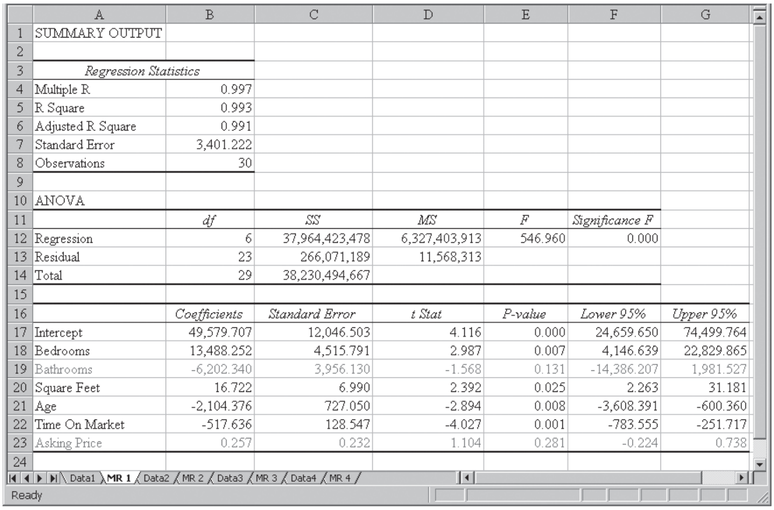


	A	B	C	D	E	F	G
1	Sales Price	Bedrooms	Bathrooms	Square Feet	Age	Time On Market	Asking Price
2	$101,200	2	2	2,100	9	12	$102,500
3	$173,800	5	4	3,100	4	21	$183,600
4	$193,100	5	4	4,200	5	28	$210,300
5	$169,400	5	4	3,800	8	31	$187,300
6	$112,900	2	2	2,400	5	27	$115,700
7	$111,800	3	3	2,000	10	6	$114,100
8	$126,800	2	2	2,300	1	18	$130,300
9	$137,600	3	3	2,300	5	4	$141,200
10	$127,400	2	2	1,900	1	11	$130,600
11	$197,900	4	3	3,700	1	3	$206,200
12	$89,900	2	2	2,100	9	28	$97,500
13	$99,000	2	2	1,900	9	6	$101,200
14	$123,400	2	2	2,400	5	6	$127,800
15	$181,200	5	4	3,600	2	26	$205,700
16	$196,500	5	4	4,400	9	13	$216,400
17	$145,200	4	3	3,100	10	22	$154,500
18	$113,900	3	3	2,000	5	13	$124,900
19	$104,600	3	3	2,000	7	32	$111,300
20	$89,700	2	2	2,100	10	19	$104,500
21	$184,800	5	4	3,100	1	21	$198,500
22	$110,100	2	2	2,300	9	3	$113,700
23	$137,300	3	3	2,400	1	15	$153,400
24	$115,000	2	2	2,000	4	5	$120,300
25	$106,500	3	3	2,200	10	28	$115,000
26	$115,300	2	2	1,900	4	2	$116,000
27	$111,300	3	3	2,100	5	26	$126,100
28	$132,200	3	3	2,400	4	15	$137,100
29	$209,500	5	4	4,900	5	27	$234,400
30	$121,600	2	2	2,300	1	21	$128,800
31	$181,500	4	3	3,300	1	8	$193,600

Figure 4.2 The selling price data

middle of the data set. In order to meet the contiguous columns requirement, we must move the number of bathrooms data out of the way and then either delete the now blank column (if there is nothing else above or below it) or move the remaining data over to remove the blank column.

Due to the chance of deleting other parts of the worksheet while doing this or accidentally deleting part of your data, the author recommends that you take two steps to protect yourself. First, move the regression



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.997					
5	R Square	0.993					
6	Adjusted R Square	0.991					
7	Standard Error	3,401.222					
8	Observations	30					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	6	37,964,423,478	6,327,403,913	546.960	0.000	
13	Residual	23	266,071,189	11,568,313			
14	Total	29	38,230,494,667				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	49,579.707	12,046.503	4.116	0.000	24,659.650	74,499.764
18	Bedrooms	13,488.252	4,515.791	2.987	0.007	4,146.639	22,829.865
19	Bathrooms	-6,202.340	3,956.130	-1.568	0.131	-14,386.207	1,981.527
20	Square Feet	16.722	6.990	2.392	0.025	2.263	31.181
21	Age	-2,104.376	727.050	-2.894	0.008	-3,608.391	-600.360
22	Time On Market	-517.636	128.547	-4.027	0.001	-783.555	-251.717
23	Asking Price	0.257	0.232	1.104	0.281	-0.224	0.738
24							

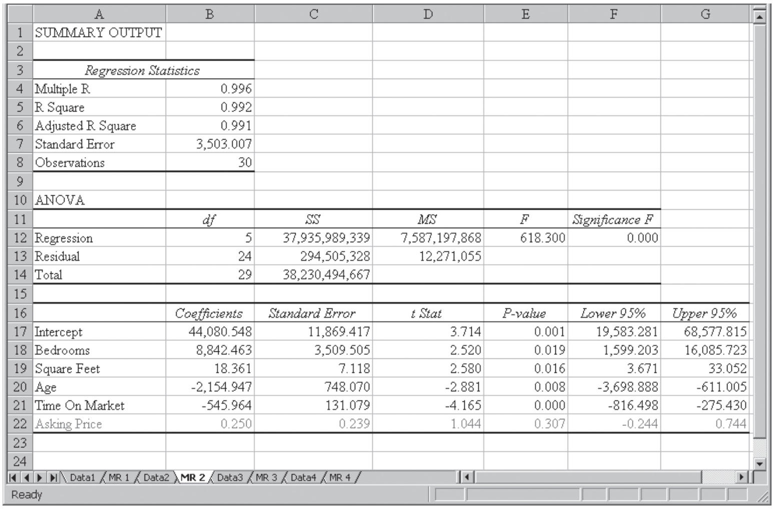
Figure 4.3. The initial regression run on the selling price data

data to its own worksheet tab so there is no chance of damaging other data with while you move data and delete columns. Second, before you begin, make a backup copy of your data on a second worksheet tab. That way, if you accidentally delete data, you can go to this backup sheet and recover it.

The less serious problem is that the regression function in Excel remembers your last set of inputs and has no reset button. That means you must remember to manually change the setting for the independent variables and output sheet each time you run regression.

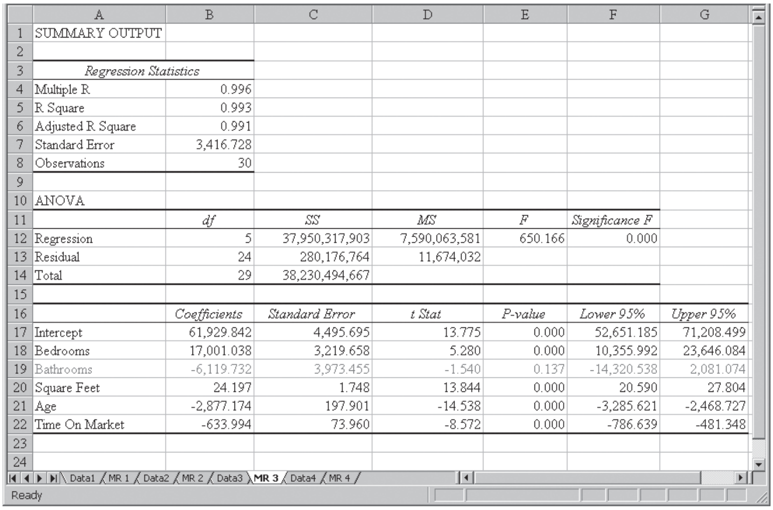
Dropping Number of Bathrooms results in an r^2 value of 0.992. That regression run is shown in Figure 4.4. We now put back in the Number of Bathrooms and drop the Initial Asking Price. Dropping the Initial Asking Price results in an r^2 value of 0.993. That regression run is shown in Figure 4.5.

Dropping the Initial Asking Price results in a higher value for r^2. (0.993 versus 0.992), so Initial Asking Price is permanently dropped from the model. This run is shown in Figure 4.5. Number of Bathrooms remains insignificant in this model, but now it is the only insignificant variable so it is dropped from the model without any testing. Those results are shown in Figure 4.6.



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.996					
5	R Square	0.992					
6	Adjusted R Square	0.991					
7	Standard Error	3,503.007					
8	Observations	30					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	5	37,935,989,339	7,587,197,868	618.300	0.000	
13	Residual	24	294,505,328	12,271,055			
14	Total	29	38,230,494,667				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	44,080.548	11,869.417	3.714	0.001	19,583.281	68,577.815
18	Bedrooms	8,842.463	3,509.505	2.520	0.019	1,599.203	16,085.723
19	Square Feet	18.361	7.118	2.580	0.016	3.671	33.052
20	Age	-2,154.947	748.070	-2.881	0.008	-3,698.888	-611.005
21	Time On Market	-545.964	131.079	-4.165	0.000	-816.498	-275.430
22	Asking Price	0.250	0.239	1.044	0.307	-0.244	0.744
23							
24							

Figure 4.4 Dropping number of bathrooms from the model

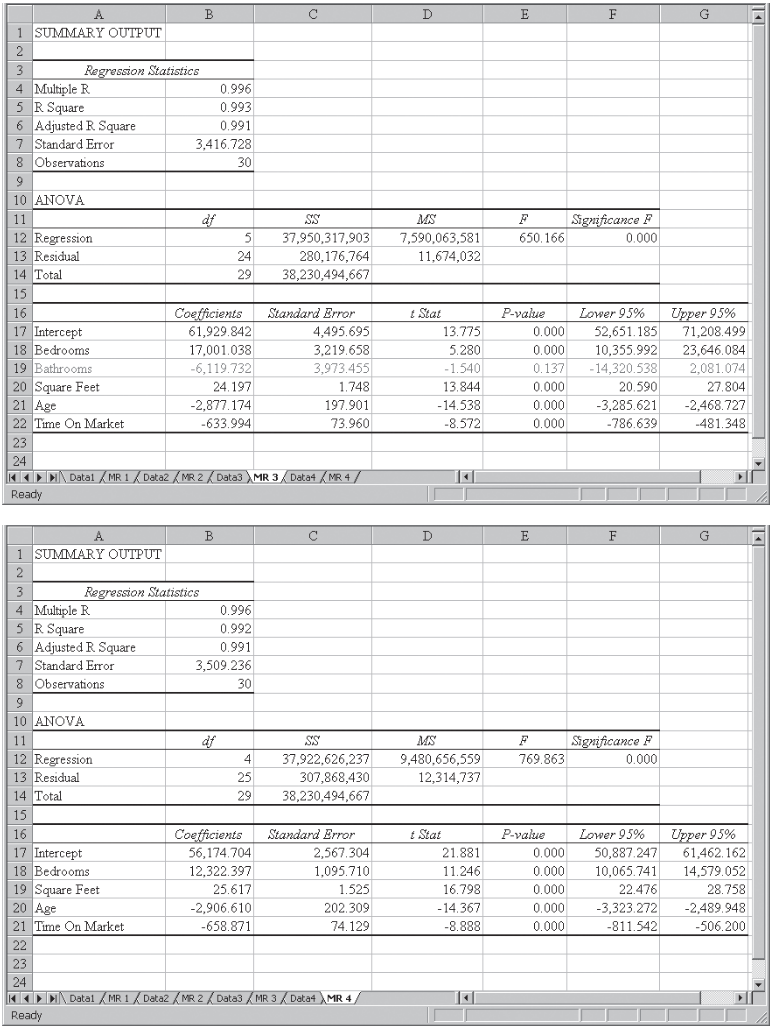


	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.996					
5	R Square	0.993					
6	Adjusted R Square	0.991					
7	Standard Error	3,416.728					
8	Observations	30					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	5	37,950,317,903	7,590,063,581	650.166	0.000	
13	Residual	24	280,176,764	11,674,032			
14	Total	29	38,230,494,667				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	61,929.842	4,495.695	13.775	0.000	52,651.185	71,208.499
18	Bedrooms	17,001.038	3,219.658	5.280	0.000	10,355.992	23,646.084
19	Bathrooms	-6,119.732	3,973.455	-1.540	0.137	-14,320.538	2,081.074
20	Square Feet	24.197	1.748	13.844	0.000	20.590	27.804
21	Age	-2,877.174	197.901	-14.538	0.000	-3,285.621	-2,468.727
22	Time On Market	-633.994	73.960	-8.572	0.000	-786.639	-481.348
23							
24							

Figure 4.5 Dropping asking price from the model

In this final version of the model, all the variables are significant. The overall r^2 has only dropped from 0.993 in the original model to 0.992 in this final model. As expected, dropping insignificant variables had little impact on r^2.

In this example, all the variables that were insignificant in the original model ended up being dropped from the final model and no additional



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.996					
5	R Square	0.993					
6	Adjusted R Square	0.991					
7	Standard Error	3,416.728					
8	Observations	30					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	5	37,950,317,903	7,590,063,581	650.166	0.000	
13	Residual	24	280,176,764	11,674,032			
14	Total	29	38,230,494,667				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	61,929.842	4,495.695	13.775	0.000	52,651.185	71,208.499
18	Bedrooms	17,001.038	3,219.658	5.280	0.000	10,355.992	23,646.084
19	Bathrooms	-6,119.732	3,973.455	-1.540	0.137	-14,320.538	2,081.074
20	Square Feet	24.197	1.748	13.844	0.000	20.590	27.804
21	Age	-2,877.174	197.901	-14.538	0.000	-3,285.621	-2,468.727
22	Time On Market	-633.994	73.960	-8.572	0.000	-786.639	-481.348
23							
24							

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.996					
5	R Square	0.992					
6	Adjusted R Square	0.991					
7	Standard Error	3,509.236					
8	Observations	30					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	37,922,626,237	9,480,656,559	769.863	0.000	
13	Residual	25	307,868,430	12,314,737			
14	Total	29	38,230,494,667				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	56,174.704	2,567.304	21.881	0.000	50,887.247	61,462.162
18	Bedrooms	12,322.397	1,095.710	11.246	0.000	10,065.741	14,579.052
19	Square Feet	25.617	1.525	16.798	0.000	22.476	28.758
20	Age	-2,906.610	202.309	-14.367	0.000	-3,323.272	-2,489.948
21	Time On Market	-658.871	74.129	-8.888	0.000	-811.542	-506.200
22							
23							
24							

Figure 4.6 The final model with both Number of Bathrooms and Asking Price dropped from the model

variables were dropped, so you may be wondering why it was necessary to work through this process. As will be demonstrated with a later model, this is not always the case. Because we cannot usually tell in advance when significance will change as variables are dropped, it is always necessary to go through this process when more than one variable is insignificant.

Before we move on, we will take a moment to consider this model from a business, rather than a statistical, standpoint. How might a model like this be used? One possible use is appraisal. Because the model quantifies the selling price of a house based on the house's attributes, an appraiser (tax appraiser or loan appraiser) can plug in the attributes of a house under consideration and get an estimate of the value of that house. That estimate is not exact because things like condition and aesthetics also play a role, but it is a good starting point.

Likewise, a homeowner considering making an addition such as adding a bedroom, or a contractor trying to sell an addition, could use the model to estimate the improvement in the value of that addition. This, in turn, might affect what the homeowner is willing to pay or the contractor is able to charge.

Box 4.2

Using Regression to Schedule Meter Reading

Scheduling service personnel is very difficult when the services they perform are not routine. For example, scheduling calls for a plumber or cable repair person requires that you have at least a good estimate of how long each job will take and how long the travel between jobs will take. However, job time varies greatly depending on the specific situation of the job, and travel time can vary greatly depending on the time of day. We will explore these issues by way of an example from the electric utility industry. Although the resulting model is specific to the electric utility industry, the approach and techniques generalize to a great many service industries.

Electric utility companies must periodically read their customer's electric meter for billing purposes. Even in this high-tech world, many electric utilities get those readings by sending a meter reader out to walk through residential neighborhoods and commercial areas to physically look at each meter and record its readings. Research is ongoing on techniques for having the meters send their readings back to a central computer automatically, either over the power lines or via a cell phone network, but for many companies, physically reading the meters each month is cheaper.

One utility company reads meters on a 21-day cycle. That is, a meter reader reads one new route each day for 21 working days and then starts the set over again. With weekends and holidays, this 21-day cycle results in the customer getting a bill about once a month. The collection of meters that is read by a meter reader on a given day is called a *route*. Routes remain static for, on average, 2 years and each meter reader keeps the same set of 21 routes during this 2 years. This allows the meter readers to become familiar with their routes. When new construction takes place, those meters are added to the nearest route. Because new construction is rarely evenly dispersed, this results in some routes growing much more than other routes. Every 2 years, each meter reading office *reroutes*—that is, they reallocate meters among routes to try to level the workload.

In this box, we will work with actual meter-reading data from a utility company and see how multiple regression can take that data and develop a model that can be used to estimate (i.e., forecast) the time that would be required by any set of meters. This can make the rerouting process much easier and can result in routes that, at least at the start of the 2-year period, have the workload more equally distributed.

This utility company is divided into districts, and many of the districts are divided into local offices. Each local office has an assigned area for meter readings, as well as other tasks. Meter-reading routes never cross local office or district lines so the process of rerouting is constrained to optimizing the routes within each local office independently. Additionally, because each meter reader reads 21 routes, each local office can have either 21 routes, 42 routes, or 63 routes, and so on.

Another major consideration in designing routes is how hard to make the routes. Meter readers are expected to read meters for 6 hours per day. They have 1 hour in the morning to get their paperwork ready and drive to the start of their route. At the end of the day, they have 1 hour to drive back to the office, process their paperwork, and turn in any money they collected for past-due bills. Thus do you design the routes so that only an experienced meter reader who is familiar with the route and working in good weather can possibly finish it in 6 hours, or do you design the routes so an inexperienced meter reader who is unfamiliar with the route and working in poor weather has time to finish? If you choose the former, then many routes will not be finished. If you choose the latter, then experienced meter readers on familiar routes working in good weather will finish in well under 6 hours.

The term "route" is somewhat misleading. When reading a route, a meter reader will walk between consecutive meter locations if they are close together, as is usually the case for residential meters. If consecutive meters are located a considerable distance apart, as is sometimes the case with commercial meters, the meter reader will use a company-furnished automobile to drive between meter locations. However, one route does not have to be all walking or all driving. Most neighborhoods and commercial office parks are too small to take a meter reader all day to finish. Typically, a meter reader will read meters in one area for a time, then drive to another area and begin reading again. A route then may consist of two, three, or even more route segments.

The Data

The data used in this analysis were collected from experienced meter readers who were reading routes with which they were familiar. When a route consisted of more than one segment, each segment was measured and recorded individually. When anything unusual happened that significantly changed the reading time for that segment, that observation was dropped from the data set.

Additionally, certain routes have special circumstances that require significant amounts of time and are always a factor. For example, one of the meters at a local airport is on a radio tower on the opposite side of a controlled runway. To read this meter, the meter reader must go to the Federal Aviation Administration (FAA) office at the airport tower and request that an FAA person drive him across the runway in an FAA car. Once the FAA car is at the runway, both going and coming back, the operator must contact the control tower and wait for clearance to cross the active runway. As you can imagine, this greatly increases the time required to read this meter. Although few meters take this long to resolve, any route segment with a regular special circumstance was excluded from this analysis.

The variables collected for this analysis are explained next. Although other variables might have been more helpful, this set was selected because it either was available from information already stored by the utility company or was easy to measure for a given route segment.

Time. This is the time required by each meter reader to read a given route segment, recorded in minutes. Utility company meter readers use handheld computers to enter the meter readings, and these record the time the reading was entered so the time for a route segment could be calculated as the time for the last reading minus the time for the first reading.

Number of residential (or nondemand) meters. Electric meters fall into two major categories: nondemand and demand meters. Nondemand meters have a continuously moving display showing the number of kilowatt-hours of electricity that have been consumed since the meter was installed. If the reading last month was 40,000 and the reading this month is 41,200, then 1,200 kilowatt-hours were consumed between the two readings. With a nondemand meter, the meter reader simply records the reading and continues on. Nondemand meters are used almost exclusively for residents, although small business applications—roadside signs, apartment laundry rooms, and the like—might also use nondemand meters. Nondemand meters are quick to read.

Number of demand meters. Demand meters, on the other hand, take much longer to read. Like a nondemand meter, demand meters have a kilowatt-hour consumption meter that must be read and recorded. However, they also have a second meter, the demand, that records the peak consumption for the last month. Because this meter records a peak, it must be reset each month in case the following peak is lower. The demand is a significant part of a commercial electric bill so, to keep the customer from resetting the peak, the reset knob is locked with a color-coded plastic seal. After reading and recording the peak, the meter reader must break the seal, reset the meter, and install a new seal of a different color. Naturally, reading a demand meter takes significantly longer than reading a nondemand meter. Additionally, business meters (i.e., demand meters) tend to be further apart than residential meters (i.e., nondemand meters), a fact that the regression model will also consider.

Number of locations. Reading 100 meters in an apartment complex with 20 meters per building would take less time than reading 100 meters on the sides of 100 houses. This variable records the number of separate locations for each route segment.

Number of collects. When a residential electric bill is 2 months past due, this utility company expects its meter readers to try to collect for that bill as they read their routes. They are given bill collection cards in the morning that they must sequence into their routes. These cards are marked as either *collects* or *cuts*. With a collect, the meter reader knocks on the door and requests payment. If payment is given, then the meter reader collects that money and gives them a receipt. If no one is home, the meter reader leaves a preprinted note on the door. If the customer refuses to pay or is not home, no other action is taken.

Number of cuts. If the card is marked as a cut, then all the previously mentioned steps take place, but if the meter reader does not receive payment for any reason, he cuts off power to that house. This is done by cutting a seal on the meter box, removing the meter box cover, pulling out the electric meter, putting plastic sleeves on plugs on the back of the meter, reinstalling the meter, reinstalling the face plate, and locking the meter. As complicated as it sounds, it can be completed in under 60 seconds by an experienced meter reader. Although the number of collects and cuts will vary month to month, some routes are statistically much more likely to have a higher number of collects and cuts than are others. Interestingly, this is not always related to the average income of the neighborhood. Meter readers only attempt collections and cut off power for residential customers. Collections and cutoffs of commercial and other classes of customers are handled by a special bill collector.

Miles walked. It would be nice to know the real number of miles a meter reader needed to walk for a route. Residential meters can be on the front, on either side, on the back of a house, or even in the basement. This requires walking up the front yard, perhaps down one side, and perhaps around the back of the house. Commercial meters can be located anywhere around a building or in a power room or maintenance room inside or even on the roof. Measuring all these distances would take too long and would require the cooperation of all the meter readers. Miles walked then is a surrogate. It is simply the mileage as measured by driving a car down the street along the route. Naturally, the miles walked by the meter reader would be greater than this, but regression can account for this.

Miles driven. When a route segment must be driven, this is simply the mileage recorded using the automobile odometer. This does not include travel to and from the route because that is not part of the 6 hours allotted to meter reading.

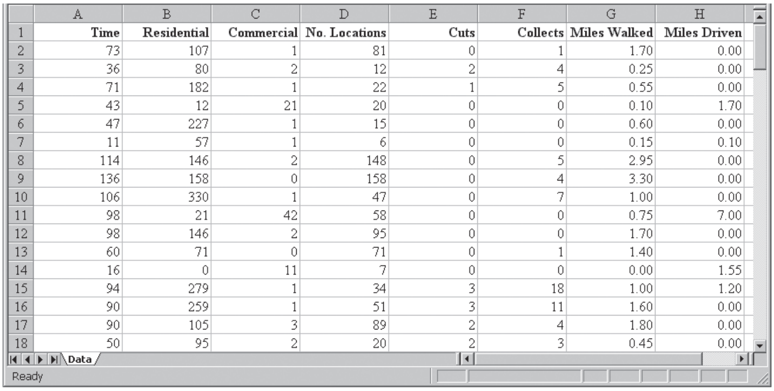
There are a number of additional factors that can affect meter-reading times on a route. These considerations are best classified as random variations or white noise in the meter-reading process. No effort was made to quantify or measure these. Examples include the following:

- *Unfriendly dogs*. An aggressive dog inside a fence is a problem when the electric meter is inside that same fence. The meter reader must either try to coerce the dog into allowing him entry or take the time to knock on the door and get the owner to control the dog while he reads the meter. An aggressive dog running loose can cause the meter reader difficulty for any number of houses.
- *Fences*. Fenced yards require more walking to gain access through the gates or the meter reader must climb over the fence. Locked gates only exacerbate this problem.
- *Bushes*. Bushes make it hard to get close to meters and to see them.
- *Traffic*. While on the driving portion of a route, traffic can delay the meter reader.

Experience indicates that the variations affect many routes in a fairly random fashion. For this reason, they were not measured or included in this analysis.

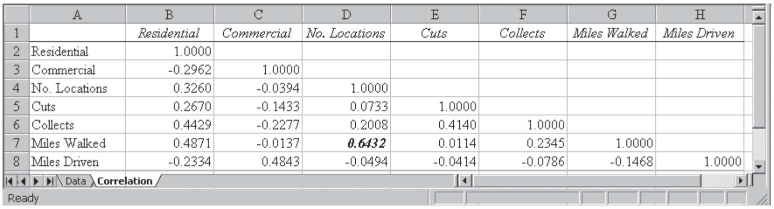
Data Analysis

Figure 4.7 shows the top of the data file. The data are stored in the *Data* tab of the Meter.xls worksheet. Figure 4.8 shows correlation analysis on the independent variables. Multicollinearity is not much of a problem, with only number of locations and miles walked clearing the 0.60 hurdle.



	A	B	C	D	E	F	G	H
1	Time	Residential	Commercial	No. Locations	Cuts	Collects	Miles Walked	Miles Driven
2	73	107	1	81	0	1	1.70	0.00
3	36	80	2	12	2	4	0.25	0.00
4	71	182	1	22	1	5	0.55	0.00
5	43	12	21	20	0	0	0.10	1.70
6	47	227	1	15	0	0	0.60	0.00
7	11	57	1	6	0	0	0.15	0.10
8	114	146	2	148	0	5	2.95	0.00
9	136	158	0	158	0	4	3.30	0.00
10	106	330	1	47	0	7	1.00	0.00
11	98	21	42	58	0	0	0.75	7.00
12	98	146	2	95	0	0	1.70	0.00
13	60	71	0	71	0	1	1.40	0.00
14	16	0	11	7	0	0	0.00	1.55
15	94	279	1	34	3	18	1.00	1.20
16	90	259	1	51	3	11	1.60	0.00
17	90	105	3	89	2	4	1.80	0.00
18	50	95	2	20	2	3	0.45	0.00

Figure 4.7 The top of the data file for the meter-reading data. The full data set has 121 observations



	A	B	C	D	E	F	G	H
1		*Residential*	*Commercial*	*No. Locations*	*Cuts*	*Collects*	*Miles Walked*	*Miles Driven*
2	Residential	1.0000						
3	Commercial	-0.2962	1.0000					
4	No. Locations	0.3260	-0.0394	1.0000				
5	Cuts	0.2670	-0.1433	0.0733	1.0000			
6	Collects	0.4429	-0.2277	0.2008	0.4140	1.0000		
7	Miles Walked	0.4871	-0.0137	***0.6432***	0.0114	0.2345	1.0000	
8	Miles Driven	-0.2334	0.4843	-0.0494	-0.0414	-0.0786	-0.1468	1.0000

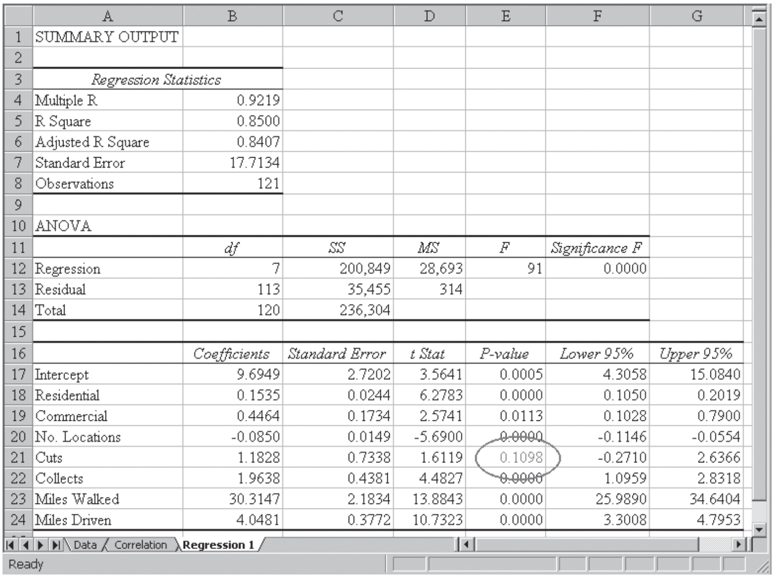
Figure 4.8 Correlation analysis results on the independent variables for the meter-reading data

Figure 4.9 shows the initial regression run on the data. The variations in the independent variables explain 85 percent of the variations in the dependent variable. Only the Number of cuts was insignificant, so that variable was dropped from the analysis. Figure 4.10 shows the resulting regression. This time, all the variables are significant, and still about 85 percent of the variation is explained.

This gives this resulting regression equation:

Resulting Regression Equation

Times to Read Route (Minutes) = 10.2563 + 0.1596
(Number of Residential Meters)
+0.4439(Number of Commercial Meters) – 0.0834(Number of Locations)
+2.2049(Number of Collects) + 29.7468(Miles Walked)
+4.0522(Miles Driven)

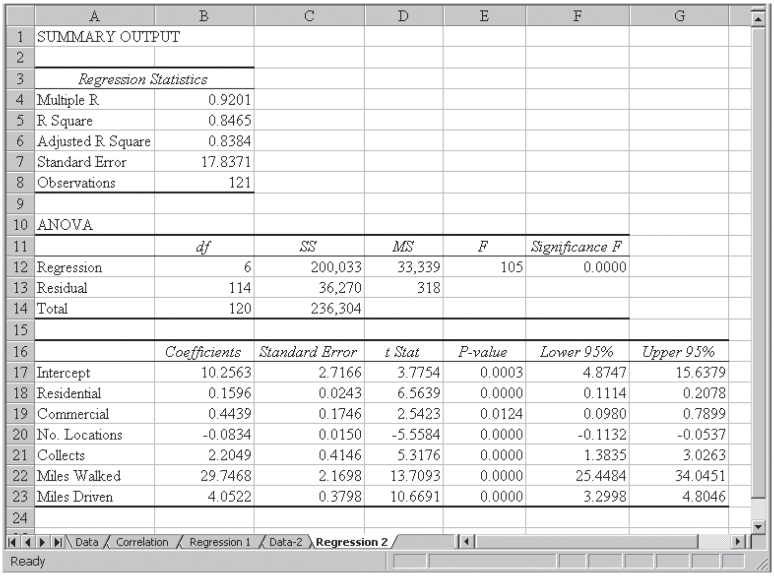


	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9219					
5	R Square	0.8500					
6	Adjusted R Square	0.8407					
7	Standard Error	17.7134					
8	Observations	121					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	7	200,849	28,693	91	0.0000	
13	Residual	113	35,455	314			
14	Total	120	236,304				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	9.6949	2.7202	3.5641	0.0005	4.3058	15.0840
18	Residential	0.1535	0.0244	6.2783	0.0000	0.1050	0.2019
19	Commercial	0.4464	0.1734	2.5741	0.0113	0.1028	0.7900
20	No. Locations	-0.0850	0.0149	-5.6900	0.0000	-0.1146	-0.0554
21	Cuts	1.1828	0.7338	1.6119	0.1098	-0.2710	2.6366
22	Collects	1.9638	0.4381	4.4827	0.0000	1.0959	2.8318
23	Miles Walked	30.3147	2.1834	13.8843	0.0000	25.9890	34.6404
24	Miles Driven	4.0481	0.3772	10.7323	0.0000	3.3008	4.7953

Figure 4.9 The initial regression results on the meter-reading data

Because Time was measured in minutes, this equation tells us that adding 1 residential meter to the route, while holding everything else constant, would add about 10 seconds (0.1596 × 60 seconds) to the time it takes to read a route, whereas adding a commercial meter would add about 27 seconds. Each collect adds a little over 2 minutes, each mile walked adds almost half an hour, and each mile driven adds a little over 4 minutes.

What is harder to understand, at first, is why the coefficient for Number of locations is negative. After all, adding more locations should increase the work. Recall, however, this is adding one location while *holding everything else constant*. That means no additional meters and no additional walking, so increasing the number of locations reduces the meter density and implies that the meters must be closer together because walking mileage does not change. This complexity is likely the reason for the negative coefficient, and its magnitude is so small that it has little impact on the results so it also could just be a statistical anomaly.



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9201					
5	R Square	0.8465					
6	Adjusted R Square	0.8384					
7	Standard Error	17.8371					
8	Observations	121					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	6	200,033	33,339	105	0.0000	
13	Residual	114	36,270	318			
14	Total	120	236,304				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	10.2563	2.7166	3.7754	0.0003	4.8747	15.6379
18	Residential	0.1596	0.0243	6.5639	0.0000	0.1114	0.2078
19	Commercial	0.4439	0.1746	2.5423	0.0124	0.0980	0.7899
20	No. Locations	-0.0834	0.0150	-5.5584	0.0000	-0.1132	-0.0537
21	Collects	2.2049	0.4146	5.3176	0.0000	1.3835	3.0263
22	Miles Walked	29.7468	2.1698	13.7093	0.0000	25.4484	34.0451
23	Miles Driven	4.0522	0.3798	10.6691	0.0000	3.2998	4.8046
24							

Figure 4.10 The final regression results on the meter-reading data

So What?

The resulting equation uses only easily obtainable data for each route. Using this equation would give management an easy way to manipulate route contents during a rerouting in such a manner that the resulting routes require very similar times to complete. This should lead to greater equality among the meter-reading employees and less dissatisfaction.

Including Qualitative Data in Multiple Regression

So far, all the variables that we have used in multiple regression have been ratio-scale data. Some of them, such as square feet in the last model, have been continuous, whereas others, such as number of bathrooms in the last model, have been discrete, but they have all had meaningful numbers attached to them. In this section, we will see how to include qualitative data in multiple regression. For example, in the sales model we have mentioned several times, you might want to include whether a competitor is having a sale in the model. This is an example of a qualitative variable because there is no meaningful number that can be attached to the yes or no answer to if the competitor is having a sale.

Qualitative data are very useful in business. We might want to indicate the make of the machines in a model to predict when maintenance is required, we might want to include the season of the year in a model to predict demand, or we might want to include a flag when demand was influenced by a special event. All of these situations can be handled in the same way.

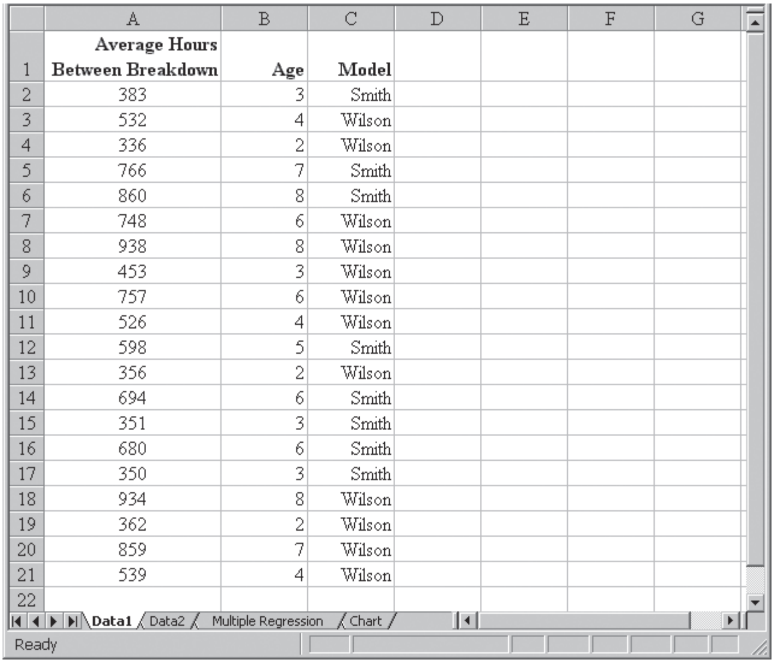
When only two possible conditions exist, such as with gender or the presence or absence of a special event, we use a special variable in the regression model. This variable goes by several names: *dichotomous variable*, *indicator variable*, or *dummy variable*. The dummy variable takes on a value of one when the condition exists and a value of zero when it does not exist.[3] For example, we would use a value of one when the special event happened and a value of zero for those periods where it did not happen.[4] For gender, we would arbitrarily choose either one or zero for male and use the other value for female.

Once the dummy variable is defined, no other special considerations are required. We run the multiple regression the same way, we test overall significance the same way, and we decide which variables to keep and which to discard exactly the same way. Dummy variables can be dropped for insignificance just like any other variable. An example follows.

Dummy Variable Example

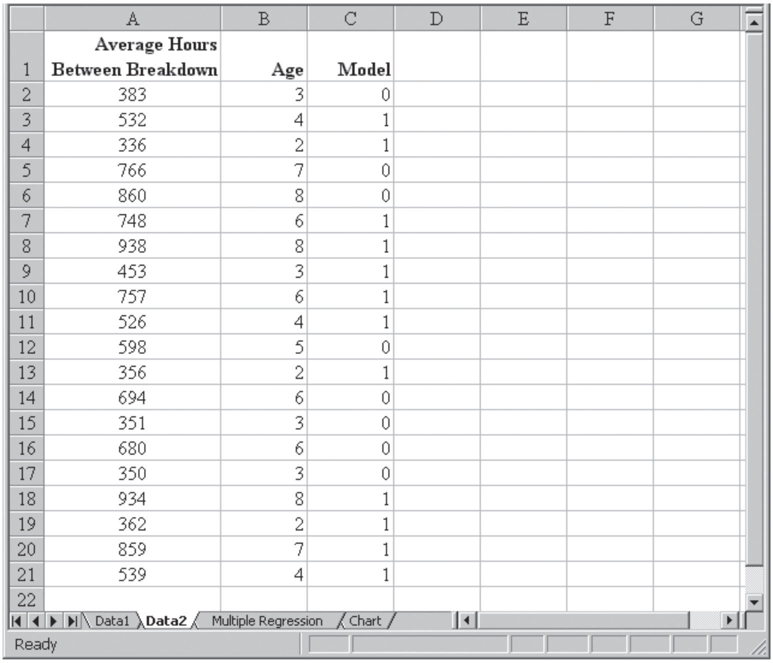
The worksheet Dummy.xls contains fictitious data on two models of machines, a Wilson and a Smith, along with their average hours between breakdown and their age. These data are shown in Figure 4.11. By coding the Smith as a zero and a Wilson as a one, these data can be used in multiple regression. The coded data are shown in Figure 4.12, and the resulting data are shown in Figure 4.13. Note that the dummy variable we created is significant in the model, as is the age of the machine.

Including a dummy variable in multiple regression causes the intercept to shift and nothing more. For this reason, dummy variables are also called *intercept shifters*. This can best be seen using the previous example.



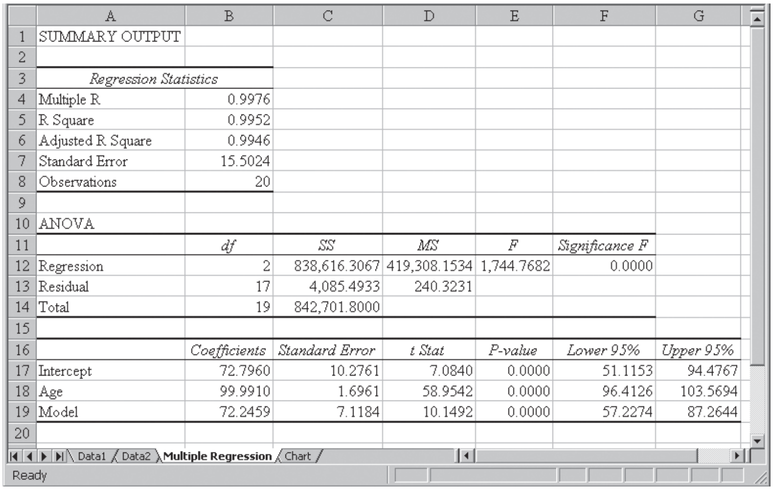
	A	B	C
1	Average Hours Between Breakdown	Age	Model
2	383	3	Smith
3	532	4	Wilson
4	336	2	Wilson
5	766	7	Smith
6	860	8	Smith
7	748	6	Wilson
8	938	8	Wilson
9	453	3	Wilson
10	757	6	Wilson
11	526	4	Wilson
12	598	5	Smith
13	356	2	Wilson
14	694	6	Smith
15	351	3	Smith
16	680	6	Smith
17	350	3	Smith
18	934	8	Wilson
19	362	2	Wilson
20	859	7	Wilson
21	539	4	Wilson
22			

Figure 4.11 The original data set with the machine name



	A	B	C
1	Average Hours Between Breakdown	Age	Model
2	383	3	0
3	532	4	1
4	336	2	1
5	766	7	0
6	860	8	0
7	748	6	1
8	938	8	1
9	453	3	1
10	757	6	1
11	526	4	1
12	598	5	0
13	356	2	1
14	694	6	0
15	351	3	0
16	680	6	0
17	350	3	0
18	934	8	1
19	362	2	1
20	859	7	1
21	539	4	1
22			

Figure 4.12 The modified data set with the machine name coded using a dummy variable



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9976					
5	R Square	0.9952					
6	Adjusted R Square	0.9946					
7	Standard Error	15.5024					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	838,616.3067	419,308.1534	1,744.7682	0.0000	
13	Residual	17	4,085.4933	240.3231			
14	Total	19	842,701.8000				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	72.7960	10.2761	7.0840	0.0000	51.1153	94.4767
18	Age	99.9910	1.6961	58.9542	0.0000	96.4126	103.5694
19	Model	72.2459	7.1184	10.1492	0.0000	57.2274	87.2644
20							

Figure 4.13 The resulting regression run

Dummy Variable Example Continued

From Figure 4.13, we get the following regression equation:

Regression Equation

Hours Between Breakdown = 72.7960 + 99.9910(Age) + 72.2459(Model)

However, Model can only take on the values of either zero or one. Substituting these values into the equation, we get the following:

Regression Equations Considering Dummy Variable

When Model = 0,
Hours Between Breakdown = 72.7960 + 99.9910(Age) + 72.2459(0),
which reduces to
Hours Between Breakdown = 72.7960 + 99.9910(Age).

When Model = 1,
Hours Between Breakdown = 72.7960 + 99.9910(Age) + 72.2459(1),
which reduces to
Hours Between Breakdown = 72.7960 + 99.9910(Age) + 72.2459,
which finally reduces to
Hours Between Breakdown = 145.0419 + 99.9910(Age).

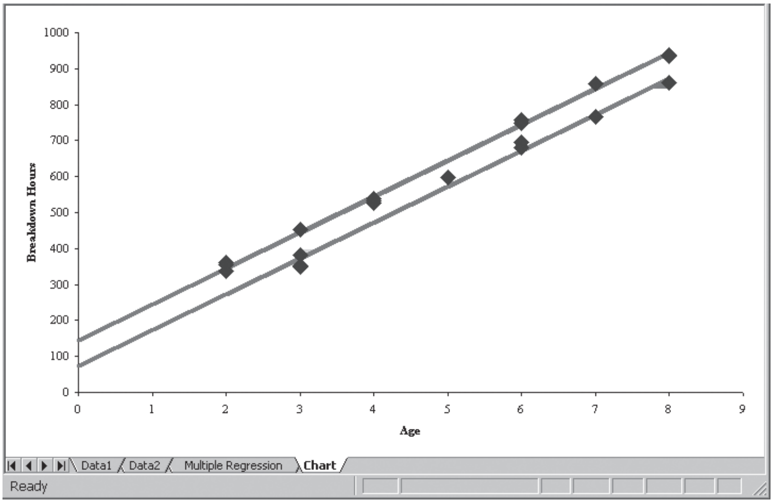


Figure 4.14 Charting the two regression lines, one when the dummy variable is zero and the other when the dummy variable is one

Thus the only difference between the two equations is the intercept. Figure 4.14 illustrates this using a chart.

More Than Two Possible Values

Of course, many times we wish to include a qualitative variable in the model that has more than two possible conditions. Examples might include race or eye color. We cannot simply define the variable using more than two values for our one dummy variable. For example, it would not be correct to have a dummy variable for competitor sale where 1 = no sale, 2 = minor sale, 3 = major sale, and 4 = clearance sale. The reason is based on the fact that dummy variables shift the intercept. Setting the variable up this way presupposes that the shift in the intercept between no sale and minor sale is the same as the shift between minor sale and major sale and the shift between no sale and minor sale is twice that of between no sale and clearance sale. Of course, we do not know this in advance, and it is likely not the case anyway.

Although we cannot code the single dummy variable in this fashion, we can include this data using multiple dummy variables. In this case, we would need three dummy variables. The first one would be a one for when there was no sale and zero otherwise. The second would be one for where there was a minor sale and zero otherwise. And the third would be one for when there was a major sale and zero otherwise. We do not need a fourth variable for other sales because a zero for all three of these dummy variables would automatically tell us that there must be a clearance sale. In fact, including this fourth, unneeded, dummy variable would force at least one of the dummy variables to be insignificant because any three would uniquely define the fourth. In general, you need $n – 1$ dummy variables to code n categories. In this case, you end up with $n – 1$ parallel lines[5] and $n – 1$ different intercepts.

The use of dummy variables when there are a large number of categories can greatly expand the number of variables in use. One of the pieces of information in the author's dissertation was *state*. Coding this into dummy variables required 49 (or 50 – 1) different variables. When the number of variables grows in this fashion, you must make sure you have an adequate sample size to support the expanded number of variables. As before, we recommend a minimum of 5 observations for each variable, including each dummy variable, with 10 per variable being even better. This was not a problem for the author because he had over 22,000 observations.

It is also possible to include more than one dummy variable in a regression model. For example, we might want to include both sale types and whether it is a holiday period in our model. When we have multiple dummy variables, each one is coded as described as done previously without consideration of any other qualitative variables that might need coding. That is, we would need one dummy variable for whether it is a holiday period and then three more for sale type, assuming the four categories previously discussed. The result would be a great deal of intercept shifting.

As you can imagine, the list of qualitative data you might need to include in a business model is quite long. Although no means comprehensive, that list includes model number, model characteristics, state or region, country, person or group, success or failure, and many more. Some of these are explored in Box 4.3.

Dummy Variable as Dependent Variable

It is also possible to use a dummy variable as the dependent variable. When you do this, it is not called regression. Rather, it is called *discriminate analysis*. Other than the name change, discriminate analysis is performed the same way as regression. This is discussed in more detail in Box 4.3.

Box 4.3

The Business of Getting Elected to Congress

As stated previously, businesses often need to include a wide range of qualitative data in statistical models. While the following example is not a business example in the truest sense of the word, it does illustrate the use of qualitative data both as a dependent variable and as independent variables.

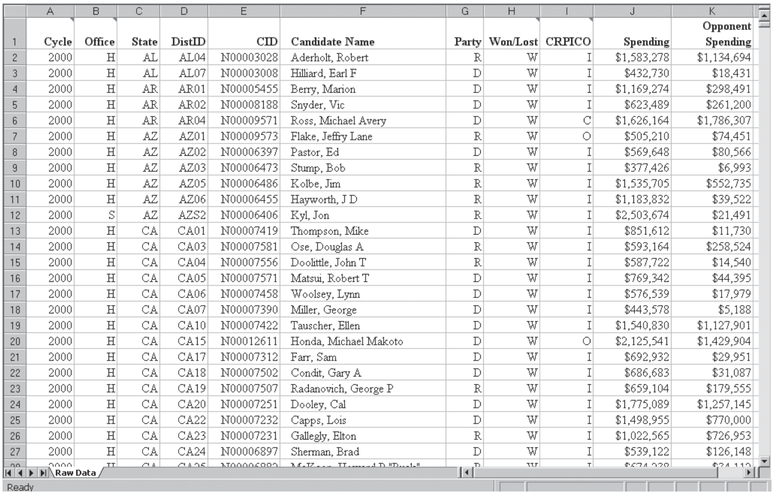
It takes a lot of money to get elected to Congress. Once elected, it takes a lot of money to stay elected. In this box, we will use a form of multiple regression analysis called *discriminate analysis* to see what affects who gets elected to Congress. As was described in the chapter, discriminate analysis is nothing more than multiple regression where the dependent variable is a dummy variable. In this case, the dependent variable will be whether they won their election.

The data for this sidebar came from Douglas Weber, a researcher at the Center for Responsive Politics. PresidentialElection.com describes the Center for Responsive Politics as follows:

> The Center for Responsive Politics is a non-partisan, non-profit research group based in Washington, D.C. that tracks money in politics, and its effect on elections and public policy. The Center conducts computer-based research on campaign finance issues for the news media, academics, activists, and the public at large. The Center's work is aimed at creating a more educated voter, an involved citizenry, and a more responsive government.[6]

The data for this analysis come from the 1996–2000 political campaigns for the U.S. Congress. These data are stored in the Excel file CandidateSpending1996-2000.xls in the *Raw Data* tab. The top of this data file is shown in Figure 4.15. Altogether, there are 2,086 entries. The variables stored are as follows:

- *Cycle*. This is the election year. Members of the House of Representatives are elected every 2 years. Members of the



	A	B	C	D	E	F	G	H	I	J	K
1	Cycle	Office	State	DistID	CID	Candidate Name	Party	Won/Lost	CRPICO	Spending	Opponent Spending
2	2000	H	AL	AL04	N00003028	Aderholt, Robert	R	W	I	$1,583,278	$1,134,694
3	2000	H	AL	AL07	N00003008	Hilliard, Earl F	D	W	I	$432,730	$18,431
4	2000	H	AR	AR01	N00005455	Berry, Marion	D	W	I	$1,169,274	$298,491
5	2000	H	AR	AR02	N00008188	Snyder, Vic	D	W	I	$623,489	$261,200
6	2000	H	AR	AR04	N00009571	Ross, Michael Avery	D	W	C	$1,626,164	$1,786,307
7	2000	H	AZ	AZ01	N00009573	Flake, Jeffry Lane	R	W	O	$505,210	$74,451
8	2000	H	AZ	AZ02	N00006397	Pastor, Ed	D	W	I	$569,648	$80,566
9	2000	H	AZ	AZ03	N00006473	Stump, Bob	R	W	I	$377,426	$6,993
10	2000	H	AZ	AZ05	N00006486	Kolbe, Jim	R	W	I	$1,535,705	$552,735
11	2000	H	AZ	AZ06	N00006455	Hayworth, J D	R	W	I	$1,183,832	$39,522
12	2000	S	AZ	AZS2	N00006406	Kyl, Jon	R	W	I	$2,503,674	$21,491
13	2000	H	CA	CA01	N00007419	Thompson, Mike	D	W	I	$851,612	$11,730
14	2000	H	CA	CA03	N00007581	Ose, Douglas A	R	W	I	$593,164	$258,524
15	2000	H	CA	CA04	N00007556	Doolittle, John T	R	W	I	$587,722	$14,540
16	2000	H	CA	CA05	N00007571	Matsui, Robert T	D	W	I	$769,342	$44,395
17	2000	H	CA	CA06	N00007458	Woolsey, Lynn	D	W	I	$576,539	$17,979
18	2000	H	CA	CA07	N00007390	Miller, George	D	W	I	$443,578	$5,188
19	2000	H	CA	CA10	N00007422	Tauscher, Ellen	D	W	I	$1,540,830	$1,127,901
20	2000	H	CA	CA15	N00012611	Honda, Michael Makoto	D	W	O	$2,125,541	$1,429,904
21	2000	H	CA	CA17	N00007312	Farr, Sam	D	W	I	$692,932	$29,951
22	2000	H	CA	CA18	N00007502	Condit, Gary A	D	W	I	$686,683	$31,087
23	2000	H	CA	CA19	N00007507	Radanovich, George P	R	W	I	$659,104	$179,555
24	2000	H	CA	CA20	N00007251	Dooley, Cal	D	W	I	$1,775,089	$1,257,145
25	2000	H	CA	CA22	N00007232	Capps, Lois	D	W	I	$1,498,955	$770,000
26	2000	H	CA	CA23	N00007231	Gallegly, Elton	R	W	I	$1,022,565	$726,953
27	2000	H	CA	CA24	N00006897	Sherman, Brad	D	W	I	$539,122	$126,148

Figure 4.15 The top of the candidate data file

Senate are elected every 6 years, but the elections are staggered so some senators are up for election every 2 years. In this data set, the values for cycle are 1996, 1998, and 2000.

- *Office*. This is the office for which the candidate is running. In this data set, the values are H (House) or S (Senate).
- *State*. This is the two-digit postal code for the state that the candidate is seeking to represent.
- *DistID*. This is an identification number for the district from which the candidate is running.
- *CID*. This is a universal identification number for a candidate that is assigned by the Center for Responsive Politics and that stays constant throughout the candidate's career.
- *Candidate Name*. This is the name of the candidate.
- *Party*. This is the party of the candidate. Most of the candidates are either Democrats (D) or Republicans (R). A few are third-party candidates (3), independents (I), or Libertarians (L).
- *Won/Lost*. This tells if the candidate won (W) the election, lost (L) the election, or if the election was undecided (U). This variable will end up being the dependent variable.
- *CRPICO.* This is a code indicating whether the candidate is an incumbent (I), challenger (C), or if the seat is open (O).

An incumbent already holds the office for which he or she is running, a challenger runs against an incumbent, and an open election is one where there is no incumbent running.

- *Spending.* This is how much money the candidate spent on the election.
- *Opponent Spending*. This is how much money the candidate's opponents spent on the election.

Many of these variables must be modified or converted before they can be used for discriminate analysis:

- *Cycle*. This is a numeric variable and can be used as is.
- *Office*. This variable was converted to a dummy variable, with zero for House and one for Senate.
- *State*. Different states would reasonably have different spending levels for offices as well as many other likely differences. In a complete analysis, the values for the 50 states would be converted to 49 dummy variables in order to capture those effects. Given Excel's difficulty in handling large numbers of variables, it was decided to drop the state value from the analysis.
- *DistID*. This has no statistical value and was dropped for analysis.
- *CID*. This has no statistical value and was dropped for analysis.
- *Candidate Name*. This has no statistical value and was dropped for analysis.
- *Party*. There are five possible values for this variable, so four dummy variables are required. Those dummy variables are Democrats, Republicans, Libertarians, and independents. Of course, any entry with a zero for all four would be a third-party candidate.
- *Won/Lost*. This is the dependent dummy variable, so it was moved to the left side of the data set to allow the independent variables to be contiguous. A value of one was used for a win, so a zero represented either a loss or an undecided election.

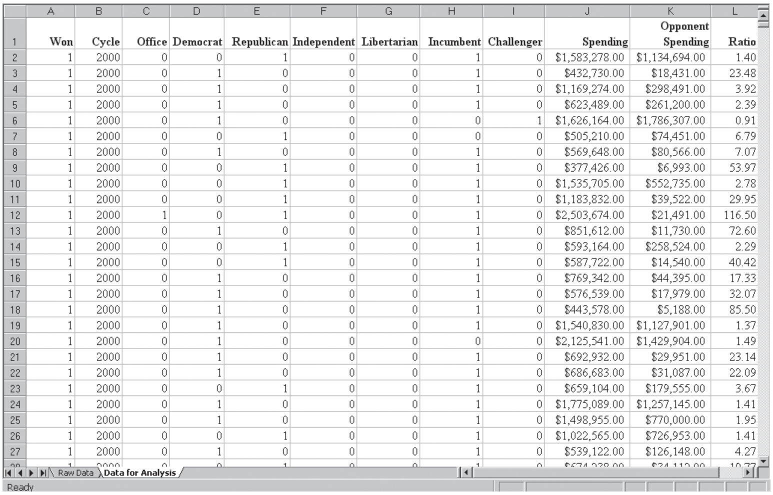
Remember that a dummy variable can only have two values (zero and one), and when used as the dependent variable, there can only be one dummy variable, so it was not possible to separate loss and undecided.

- *CRPICO.* There were three possible values, so two dummy variables are required. They are incumbent and challenger. A zero for both variables indicates an open election.
- *Spending.* This variable was used as is.
- *Opponent Spending*. This variable was used as is.
- *Ratio*. A new variable was created from the ratio of Spending divided by Opponent Spending.

The top of this modified data set is shown in Figure 4.16.

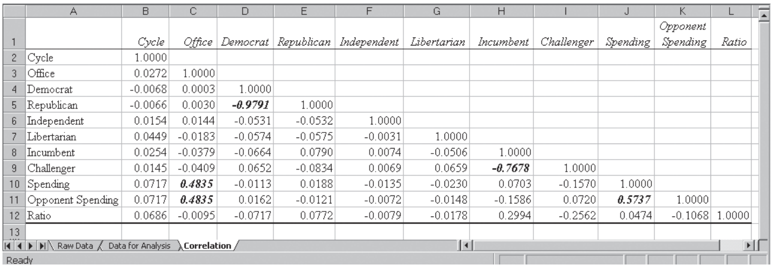
The Analysis

The first step is to perform correlation analysis on the independent variables. Those results are shown in Figure 4.17. Surprisingly, there are only four pairs of variables where multicollinearity is likely to be a problem:



	A	B	C	D	E	F	G	H	I	J	K	L
1	Won	Cycle	Office	Democrat	Republican	Independent	Libertarian	Incumbent	Challenger	Spending	Opponent Spending	Ratio
2	1	2000	0	0	1	0	0	1	0	$1,583,278.00	$1,134,694.00	1.40
3	1	2000	0	1	0	0	0	1	0	$432,730.00	$18,431.00	23.48
4	1	2000	0	1	0	0	0	1	0	$1,169,274.00	$298,491.00	3.92
5	1	2000	0	1	0	0	0	1	0	$623,489.00	$261,200.00	2.39
6	1	2000	0	1	0	0	0	0	1	$1,626,164.00	$1,786,307.00	0.91
7	1	2000	0	0	1	0	0	0	0	$505,210.00	$74,451.00	6.79
8	1	2000	0	1	0	0	0	1	0	$569,648.00	$80,566.00	7.07
9	1	2000	0	0	1	0	0	1	0	$377,426.00	$6,993.00	53.97
10	1	2000	0	0	1	0	0	1	0	$1,535,705.00	$552,735.00	2.78
11	1	2000	0	0	1	0	0	1	0	$1,183,832.00	$39,522.00	29.95
12	1	2000	1	0	1	0	0	1	0	$2,503,674.00	$21,491.00	116.50
13	1	2000	0	1	0	0	0	1	0	$851,612.00	$11,730.00	72.60
14	1	2000	0	0	1	0	0	1	0	$593,164.00	$258,524.00	2.29
15	1	2000	0	0	1	0	0	1	0	$587,722.00	$14,540.00	40.42
16	1	2000	0	1	0	0	0	1	0	$769,342.00	$44,395.00	17.33
17	1	2000	0	1	0	0	0	1	0	$576,539.00	$17,979.00	32.07
18	1	2000	0	1	0	0	0	1	0	$443,578.00	$5,188.00	85.50
19	1	2000	0	1	0	0	0	1	0	$1,540,830.00	$1,127,901.00	1.37
20	1	2000	0	1	0	0	0	0	0	$2,125,541.00	$1,429,904.00	1.49
21	1	2000	0	1	0	0	0	1	0	$692,932.00	$29,951.00	23.14
22	1	2000	0	1	0	0	0	1	0	$686,683.00	$31,087.00	22.09
23	1	2000	0	0	1	0	0	1	0	$659,104.00	$179,555.00	3.67
24	1	2000	0	1	0	0	0	1	0	$1,775,089.00	$1,257,145.00	1.41
25	1	2000	0	1	0	0	0	1	0	$1,498,955.00	$770,000.00	1.95
26	1	2000	0	0	1	0	0	1	0	$1,022,565.00	$726,953.00	1.41
27	1	2000	0	1	0	0	0	1	0	$539,122.00	$126,148.00	4.27

Figure 4.16 The top of the candidate data file after the modifications discussed have been made

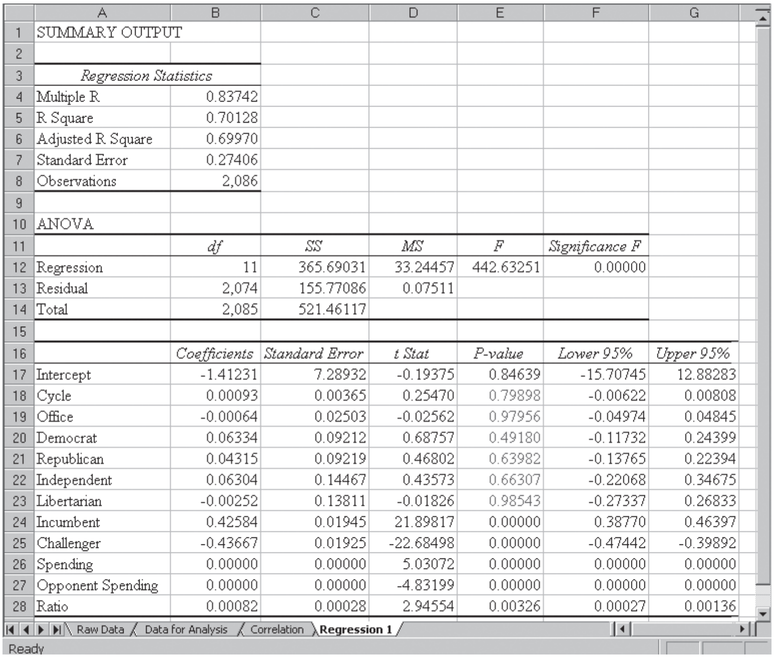


	A	B	C	D	E	F	G	H	I	J	K	L
1		*Cycle*	*Office*	*Democrat*	*Republican*	*Independent*	*Libertarian*	*Incumbent*	*Challenger*	*Spending*	*Opponent Spending*	*Ratio*
2	Cycle	1.0000										
3	Office	0.0272	1.0000									
4	Democrat	-0.0068	0.0003	1.0000								
5	Republican	-0.0066	0.0030	***-0.9791***	1.0000							
6	Independent	0.0154	0.0144	-0.0531	-0.0532	1.0000						
7	Libertarian	0.0449	-0.0183	-0.0574	-0.0575	-0.0031	1.0000					
8	Incumbent	0.0254	-0.0379	-0.0664	0.0790	0.0074	-0.0506	1.0000				
9	Challenger	0.0145	-0.0409	0.0652	-0.0834	0.0069	0.0659	***-0.7678***	1.0000			
10	Spending	0.0717	***0.4835***	-0.0113	0.0188	-0.0135	-0.0230	0.0703	-0.1570	1.0000		
11	Opponent Spending	0.0717	***0.4835***	0.0162	-0.0121	-0.0072	-0.0148	-0.1586	0.0720	***0.5737***	1.0000	
12	Ratio	0.0686	-0.0095	-0.0717	0.0772	-0.0079	-0.0178	0.2994	-0.2562	0.0474	-0.1068	1.0000
13												

Figure 4.17 Correlation analysis of the independent variable in the candidate data file

1. *Democrat/Republican (–0.9791)*. This is not surprising because almost all the candidates are either Democrats or Republications, you would expect a near perfect correlation, and we get it. The negative correlation only means that as the Democrat dummy variable goes up from zero to one, the Republican dummy variable moves in the opposite direction, from one to zero. The fact that the value is not exactly 1.00 simply indicates the presence of a few candidates who are neither Democrats nor Republicans.
2. *Opponent Spending/Office (0.4835)*. This does not pass our 0.60 threshold, but given the very small values for the other pairs, it is noticeable. Since Office is a dummy variable with a zero for House and one for Senate, this indicates a strong likelihood that nonincumbent spending is higher for the Senate than for the House. This is not surprising because the Senate is both more prestigious and a statewide seat requiring statewide campaigning.
3. *Spending/Office (0.4835)*. Like opponents, incumbents spend more for the Senate than for the House. What is surprising is that the correlation is the same for both spending variables.
4. *Challenger/Incumbent (–0.7678)*. This is also not surprising because someone who is not a incumbent must be a challenger. The reason for the less-than-perfect correlation is that some of the elections are open elections where there is no incumbent.

The initial regression results are shown in Figure 4.18. Overall, the model is able to explain a little over 70 percent of the variation in the results. No doubt, candidate positions on specific issues and character



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.83742					
5	R Square	0.70128					
6	Adjusted R Square	0.69970					
7	Standard Error	0.27406					
8	Observations	2,086					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	11	365.69031	33.24457	442.63251	0.00000	
13	Residual	2,074	155.77086	0.07511			
14	Total	2,085	521.46117				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-1.41231	7.28932	-0.19375	0.84639	-15.70745	12.88283
18	Cycle	0.00093	0.00365	0.25470	0.79898	-0.00622	0.00808
19	Office	-0.00064	0.02503	-0.02562	0.97956	-0.04974	0.04845
20	Democrat	0.06334	0.09212	0.68757	0.49180	-0.11732	0.24399
21	Republican	0.04315	0.09219	0.46802	0.63982	-0.13765	0.22394
22	Independent	0.06304	0.14467	0.43573	0.66307	-0.22068	0.34675
23	Libertarian	-0.00252	0.13811	-0.01826	0.98543	-0.27337	0.26833
24	Incumbent	0.42584	0.01945	21.89817	0.00000	0.38770	0.46397
25	Challenger	-0.43667	0.01925	-22.68498	0.00000	-0.47442	-0.39892
26	Spending	0.00000	0.00000	5.03072	0.00000	0.00000	0.00000
27	Opponent Spending	0.00000	0.00000	-4.83199	0.00000	0.00000	0.00000
28	Ratio	0.00082	0.00028	2.94554	0.00326	0.00027	0.00136

Figure 4.18 The initial regression run

issues accounted for much of the remaining percentage. Additionally, issues like these no doubt affected the candidate's ability to raise money, so some of those issues would be reflected in the spending variable.

Cycle is not significant, which seems to indicate that the impact of the various variables has not changed over the relatively short period of time represented in this data set. Office is not significant, indicating that election patterns are fairly consistent for the House and Senate. The Democrat and Republican dummy variables are not significant, which is surprising. This seems to indicate that spending and being an incumbent are much more important than party affiliations. Along the same lines, Libertarian is also not significant, but because there were only seven Libertarian candidates in this period, that is not surprising.

Incumbent was significant with a positive coefficient, as expected. That is, being an incumbent strongly helps your chance of being elected. Along the same lines, being a challenger was also significant and, as would be expected, had a negative coefficient. Spending and Opponent Spending are both significant and both have positive coefficients. They appear to be zero because the dependent variable is either zero or one and spending is measured in dollars and so has values in the millions. This large difference in the units yields very small coefficients. The spending ratio is also significant with a positive coefficient, indicating that the higher a candidate's spending relative to his opponent's spending, the better his or her chance of being elected. Again, the magnitude of the spending ratio coefficient is due to the magnitude of the ratios and not to its importance. Note that the largest ratio was over 650.

Normally, we would need to work through these variables, dropping them one at a time, to figure out which to drop. That work has been done but is not shown. All the variables that are insignificant in the first model end up dropping out of the final model, although when either the Democrat or Republican dummy variable is left in, whichever variable that is left in is almost significant.

Additionally, the two spending variables were both divided by 1,000,000, yielding spending expressed in millions of dollars. This is a linear transformation, so it has no effect on correlation or r^2 but does keep the coefficients for the spending variables from being so small. Also, transforming the spending variables has no effect on their ratio, so that variable stays the same. While not shown in a figure, this reduced data set is stored in the *Data2* tab of the worksheet.

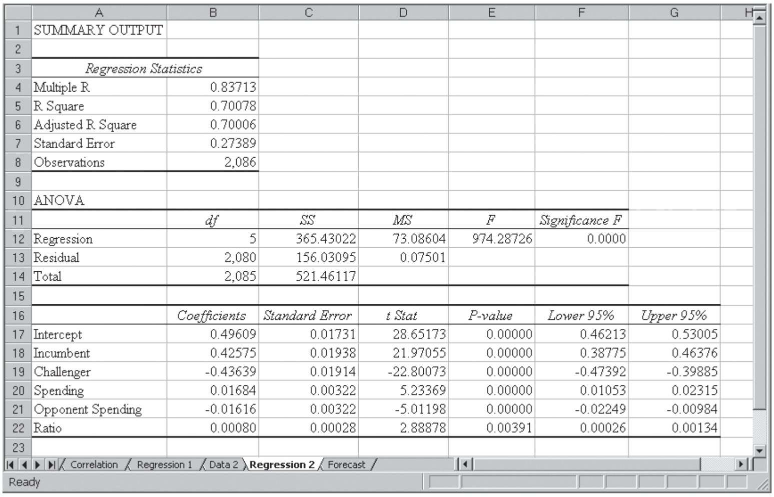
The regression on this final, reduced data set is shown in Figure 4.19. Notice that everything is significant. Variations in these six variables explain 70 percent of the variation in who won the election. This gives the following equation:

Final Regression Equation

$$\begin{gathered}Y = 0.49609 + 0.42575(\text{Incumbent}) - 0.43639(\text{Challenger}) + \\ 0.01684(\text{Spending, in millions}) - 0.01616(\text{Opponent Spending,} \\ \text{in millions}) + 0.00080(\text{Ratio})\end{gathered}$$

All these variables have the sign that we would expect.

If we are careful, in discriminate analysis, then it is possible to use the magnitude of the final coefficients to analyze the relative impacts of the independent variables. Great care is required because the



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.83713					
5	R Square	0.70078					
6	Adjusted R Square	0.70006					
7	Standard Error	0.27389					
8	Observations	2,086					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	5	365.43022	73.08604	974.28726	0.0000	
13	Residual	2,080	156.03095	0.07501			
14	Total	2,085	521.46117				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	0.49609	0.01731	28.65173	0.00000	0.46213	0.53005
18	Incumbent	0.42575	0.01938	21.97055	0.00000	0.38775	0.46376
19	Challenger	-0.43639	0.01914	-22.80073	0.00000	-0.47392	-0.39885
20	Spending	0.01684	0.00322	5.23369	0.00000	0.01053	0.02315
21	Opponent Spending	-0.01616	0.00322	-5.01198	0.00000	-0.02249	-0.00984
22	Ratio	0.00080	0.00028	2.88878	0.00391	0.00026	0.00134
23							

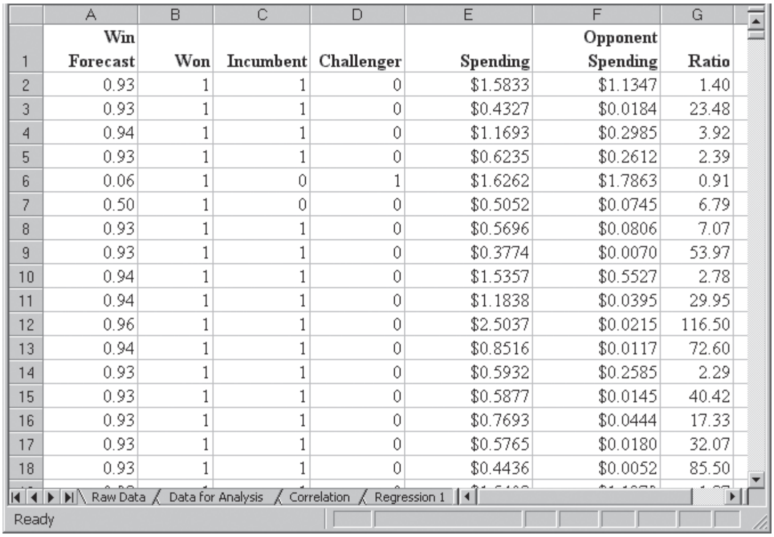
Figure 4.19 The final regression model for the candidate data file

magnitude of the coefficients, which is what we will be analyzing, is greatly influenced by the units in which the variable was measured. Of course, not all variables have a problem with units, as will be seen.

Understanding the Coefficients

Recall that our dependent variable was a dummy variable that had a value of one if the candidate won the election and a value of zero if the candidate lost the election. While all the observations in the data set have a value of zero or one for this variable, the resulting regression equation is not restricted to this 0–1 range. When this regression equation was applied to the 2,086 observations in this data set, on the *Forecast* tab, values for the forecasted dependent variable ranged from –0.43 to 1.47. Nevertheless 2,051 of 2.086 (98.3 percent) of the forecasts did fall within this range. Some of these results are shown in Figure 4.20.

Everything else being equal, the closer a candidate's score is to one, the higher the likelihood that they will win the election. Likewise, the closer a candidate's score is to zero, the lower the likelihood that they will win the election. So the forecast that results from the regression equation can roughly be treated as a probability of winning. This is only a rough approximation because 30 percent of the variation is unexplained and because values below zero or greater than one are possible.



	A	B	C	D	E	F	G
1	Win Forecast	Won	Incumbent	Challenger	Spending	Opponent Spending	Ratio
2	0.93	1	1	0	$1.5833	$1.1347	1.40
3	0.93	1	1	0	$0.4327	$0.0184	23.48
4	0.94	1	1	0	$1.1693	$0.2985	3.92
5	0.93	1	1	0	$0.6235	$0.2612	2.39
6	0.06	1	0	1	$1.6262	$1.7863	0.91
7	0.50	1	0	0	$0.5052	$0.0745	6.79
8	0.93	1	1	0	$0.5696	$0.0806	7.07
9	0.93	1	1	0	$0.3774	$0.0070	53.97
10	0.94	1	1	0	$1.5357	$0.5527	2.78
11	0.94	1	1	0	$1.1838	$0.0395	29.95
12	0.96	1	1	0	$2.5037	$0.0215	116.50
13	0.94	1	1	0	$0.8516	$0.0117	72.60
14	0.93	1	1	0	$0.5932	$0.2585	2.29
15	0.93	1	1	0	$0.5877	$0.0145	40.42
16	0.93	1	1	0	$0.7693	$0.0444	17.33
17	0.93	1	1	0	$0.5765	$0.0180	32.07
18	0.93	1	1	0	$0.4436	$0.0052	85.50

Figure 4.20 Forecasting winning using the candidate data set

The intercept is 0.49609 or very nearly 50 percent. This is exactly what we would expect. Without considering money or whether or not a candidate is an incumbent, with two strong parties, a candidate should have about a 50-50 chance.

Recall from above that the Democrat variable is almost significant. If it is left in for the final regression, none of the above coefficients changes more than a minor amount, and the Democrat variable has a coefficient of 0.02079. That is, during the time range associated with this data set, being a Democrat has a small (0.02079) positive impact on the probability of winning. Of course, one of the problems with handicapping political races using historical data is the 2002 elections, where being a Democrat had a negative impact. Because the data we are analyzing stops at 2000, the impact of the 2002 elections is not included. This illustrates the difficulty of using historical data to forecast elections.

Being the incumbent had a large (0.42575) positive impact on the probability of winning, while being a challenger had a large (–0.43639) negative impact on the probability of winning. Recall that spending is now measured in millions of dollars, and spending an additional one million dollars has only a small (0.01684) impact on the probability of winning. Of course, had spending been measured in tens of millions of dollars, the coefficient would be larger (0.1684), but the overall impact of spending would be the same, thus the caveat regarding the units used to measure the variables. Opponent spending had about the same impact (–0.01616 versus 0.01684) for a million dollars spent, thus candidate and opponent spending tend to offset one another. The ratio of opponent spending has only a minor impact (0.00080) on the probability of winning, so it takes a large imbalance to cause a large swing in the probabilities. Of course, these data are all for national offices with very large spending levels. This observation is not likely to be true for state or local elections with their relatively small budgets.

Conclusion

This box has shown how election data can be used to build a model for predicting the relative chances of a political hopeful being elected to Congress based on the candidate's and the candidate's opponent's incumbent and spending status. The resulting model was able to explain about 70 percent of the variation.

Now imagine that, rather than political data, this worksheet contained income, debt, and spending data on consumers. Also imagine that rather than election results, the dependent variable was a dummy variable where one represented a good credit risk and zero represented a bad credit risk. If that were the case, then column A in Figure 4.20 could contain credit scores rather than electability scores. Statistical analysis of credit history similar to what is presented here is, in fact, exactly how credit scores are developed.

Because regression, and therefore discriminate analysis, can have only one dependent variable, it is only possible to use a dummy variable as the dependent variable when you only have two possible categories, such as repaying a loan (or not) or making a sale. These types of models are often used by financial institutions where the dummy dependent variable represents whether someone is a good credit risk. The purpose of discriminate analysis is to assign each observation into one of the two categories described by the dependent variable. In other words, we wish to discriminate between the two possible outcomes. The development of this type of model is left to interested readers.

Because regression can only have one dependent variable, regression-based discriminate analysis can only support two categories. There are more advanced approaches for handling more than two categories. Interested students are referred to an advanced reference.

Testing the Validity of the Regression Model

There are three main problems, or diseases, that can affect multiple regression:

1. Multicollinearity
2. Autocorrelation
3. Heteroscedasticity

We will look at spotting and treating each of the problems individually.

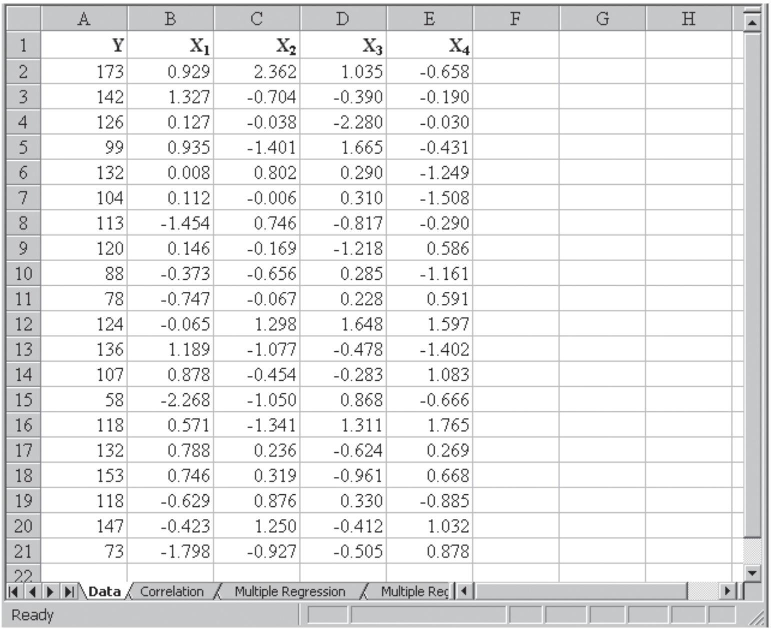
Multicollinearity

Multicollinearity is a major problem that affects almost every set of data to some degree. It is the sole reason we cannot just drop all the insignificant variables at once. As we will see, it can also cause coefficients to be hard to understand, as well as an array of other problems. Were it not for multicollinearity, developing multiple regression models would be an order of magnitude easier.

The best way to see the impact of multicollinearity is to see how well multiple regression performs without multicollinearity. An example of this follows.

Example With No Multicollinearity

Figure 4.21 shows a data set that has one dependent variable, Y, and four independent variables, X_1 through X_4. The independent variables were constructed such that they have absolutely no multicollinearity.[7] Because multicollinearity is correlation between the independent variables, the quickest way to test for multicollinearity is via a correlation



	A	B	C	D	E
1	Y	X_1	X_2	X_3	X_4
2	173	0.929	2.362	1.035	-0.658
3	142	1.327	-0.704	-0.390	-0.190
4	126	0.127	-0.038	-2.280	-0.030
5	99	0.935	-1.401	1.665	-0.431
6	132	0.008	0.802	0.290	-1.249
7	104	0.112	-0.006	0.310	-1.508
8	113	-1.454	0.746	-0.817	-0.290
9	120	0.146	-0.169	-1.218	0.586
10	88	-0.373	-0.656	0.285	-1.161
11	78	-0.747	-0.067	0.228	0.591
12	124	-0.065	1.298	1.648	1.597
13	136	1.189	-1.077	-0.478	-1.402
14	107	0.878	-0.454	-0.283	1.083
15	58	-2.268	-1.050	0.868	-0.666
16	118	0.571	-1.341	1.311	1.765
17	132	0.788	0.236	-0.624	0.269
18	153	0.746	0.319	-0.961	0.668
19	118	-0.629	0.876	0.330	-0.885
20	147	-0.423	1.250	-0.412	1.032
21	73	-1.798	-0.927	-0.505	0.878

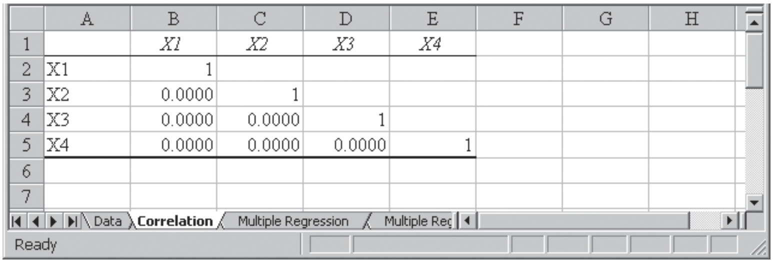
Figure 4.21 Made-up data set containing no multicollinearity

matrix containing just the independent variables. This is shown in Figure 4.22. As you can see, no correlation, and therefore no multicollinearity, is present.

Figure 4.23 shows the initial regression run. Notice that X_3 and X_4 are not significant. Additionally, notice the equation:

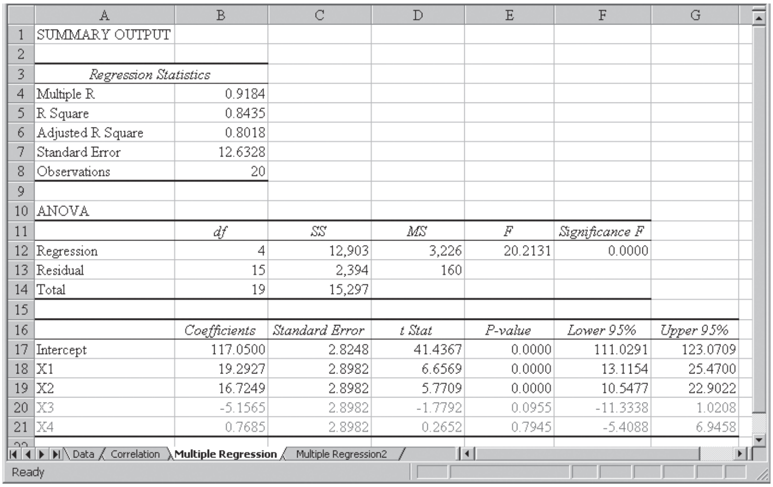
Regression Equation

$$\hat{Y} = 117.0500 + 19.2927X_1 \ + \ 16.7249X_2 - 5.1565X_3 + 0.7685X_4$$



	A	B	C	D	E
1		*X1*	*X2*	*X3*	*X4*
2	X1	1			
3	X2	0.0000	1		
4	X3	0.0000	0.0000	1	
5	X4	0.0000	0.0000	0.0000	1

Figure 4.22 Results of running correlation analysis on this fictitious data set that contains no multicollinearity



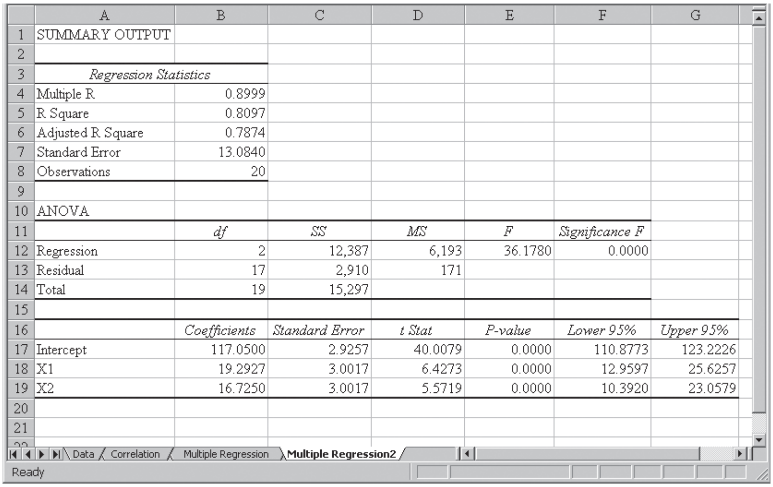
	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9184					
5	R Square	0.8435					
6	Adjusted R Square	0.8018					
7	Standard Error	12.6328					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	12,903	3,226	20.2131	0.0000	
13	Residual	15	2,394	160			
14	Total	19	15,297				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	117.0500	2.8248	41.4367	0.0000	111.0291	123.0709
18	X1	19.2927	2.8982	6.6569	0.0000	13.1154	25.4700
19	X2	16.7249	2.8982	5.7709	0.0000	10.5477	22.9022
20	X3	-5.1565	2.8982	-1.7792	0.0955	-11.3338	1.0208
21	X4	0.7685	2.8982	0.2652	0.7945	-5.4088	6.9458

Figure 4.23 Results of the initial regression run on this fictitious data set that contains no multicollinearity

Now we will simply drop the two insignificant variables. The results are shown in Figure 4.24. Notice the resulting equation:

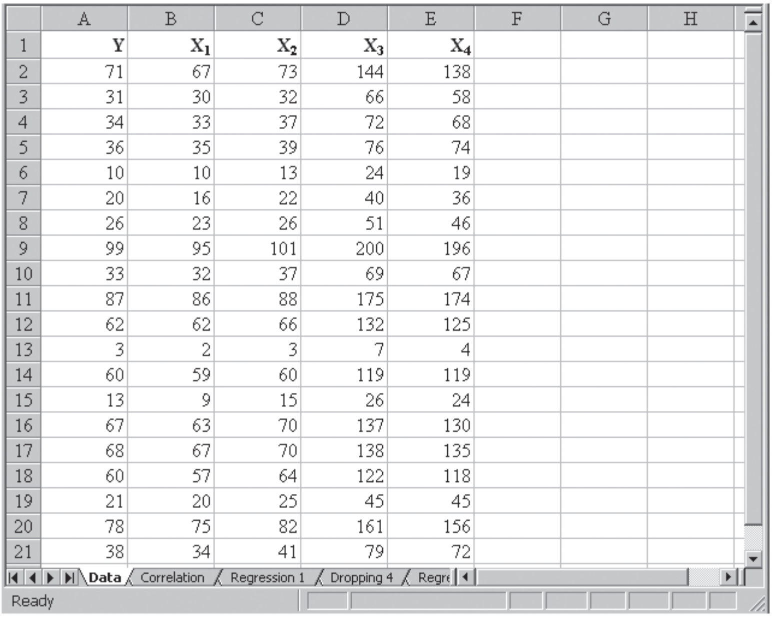
Regression Equation After Dropping Two Variables

$$\hat{Y} = 117.0500 + 19.2927X_1 + 16.7249X_2$$



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.8999					
5	R Square	0.8097					
6	Adjusted R Square	0.7874					
7	Standard Error	13.0840					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	12,387	6,193	36.1780	0.0000	
13	Residual	17	2,910	171			
14	Total	19	15,297				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	117.0500	2.9257	40.0079	0.0000	110.8773	123.2226
18	X1	19.2927	3.0017	6.4273	0.0000	12.9597	25.6257
19	X2	16.7250	3.0017	5.5719	0.0000	10.3920	23.0579
20							
21							

Figure 4.24 Results of the final regression run with two variables dropped on this fictitious data set that contains no multicollinearity



	A	B	C	D	E	F	G	H
1	Y	X_1	X_2	X_3	X_4			
2	71	67	73	144	138			
3	31	30	32	66	58			
4	34	33	37	72	68			
5	36	35	39	76	74			
6	10	10	13	24	19			
7	20	16	22	40	36			
8	26	23	26	51	46			
9	99	95	101	200	196			
10	33	32	37	69	67			
11	87	86	88	175	174			
12	62	62	66	132	125			
13	3	2	3	7	4			
14	60	59	60	119	119			
15	13	9	15	26	24			
16	67	63	70	137	130			
17	68	67	70	138	135			
18	60	57	64	122	118			
19	21	20	25	45	45			
20	78	75	82	161	156			
21	38	34	41	79	72			

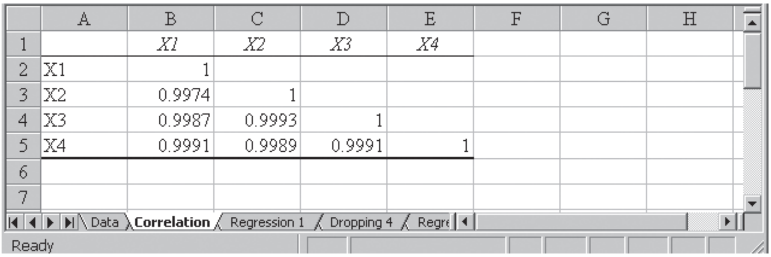
Figure 4.25 The high-multicollinearity fictitious data

Thus the intercept and coefficients for X_1 and X_2 did not change at all. Additionally, there were only minor changes for the *t* statistic for X_1 and X_2 and neither changed significance.

As the example shows, multiple regression behaves very smoothly when there is no multicollinearity. It is the presence of multicollinearity that causes much of our difficulties. Now we will look at an example with more extreme multicollinearity.

High-Multicollinearity Example

The data in the HighMulticollinearity.xls worksheet were especially constructed to have a high degree of multicollinearity. These data are shown in Figure 4.25. Figure 4.26 shows the resulting correlation matrix of just the independent variables. Notice that each pair-wise correlation exceeds 0.99. This is high multicollinearity indeed.

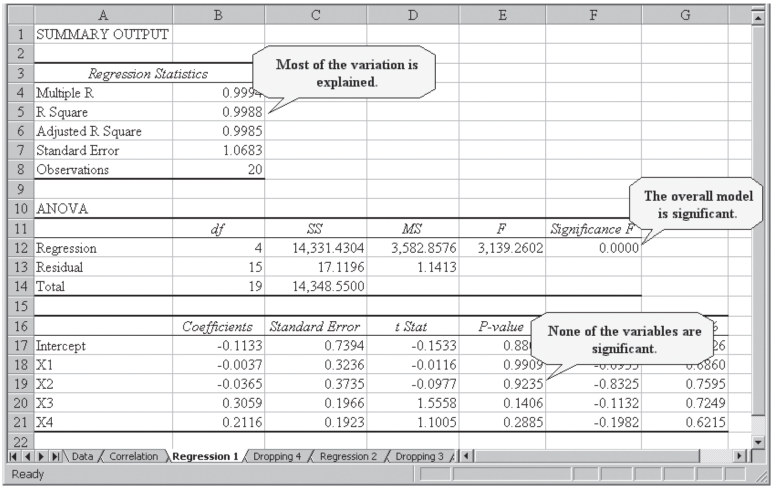


	A	B	C	D	E	F	G	H
1		*X1*	*X2*	*X3*	*X4*			
2	X1	1						
3	X2	0.9974	1					
4	X3	0.9987	0.9993	1				
5	X4	0.9991	0.9989	0.9991	1			
6								
7								

Figure 4.26 Correlation analysis on the high-multicollinearity fictitious data

Figure 4.27 shows the resulting multiple regression run. Notice the following:

- The r^2 value is 0.9988 so almost 100 percent of the variation in Y is being explained by the variation in the four independent variables. From this perspective, you could not ask for a better model.
- The overall model is significant. That is, it passes the F-test. This is to be expected given the high r^2 value.
- None of the independent variables is significant. Here, we have an overall model that is significant yet none of the variables used to construct the model is significant. This is a very clear indicator of multicollinearity.



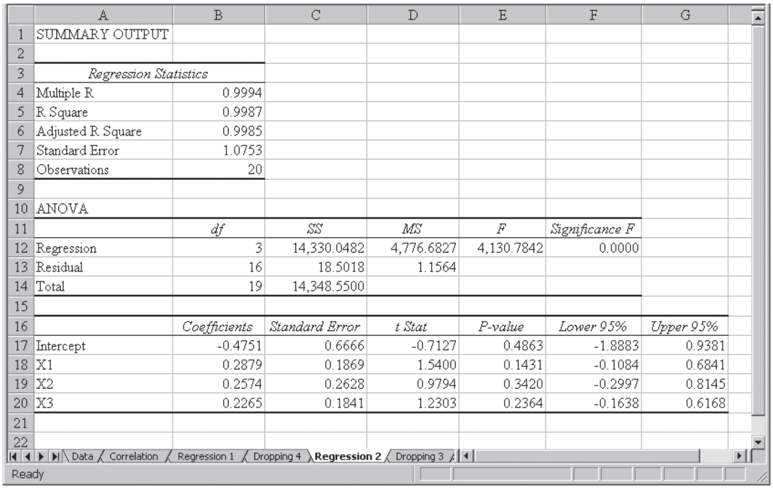
	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9994					
5	R Square	0.9988					
6	Adjusted R Square	0.9985					
7	Standard Error	1.0683					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	4	14,331.4304	3,582.8576	3,139.2602	0.0000	
13	Residual	15	17.1196	1.1413			
14	Total	19	14,348.5500				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	[illegible]	[illegible]
17	Intercept	-0.1133	0.7394	-0.1533	0.88[illegible]	[illegible]	[illegible]26
18	X1	-0.0037	0.3236	-0.0116	0.9909	[illegible]	0.6860
19	X2	-0.0365	0.3735	-0.0977	0.9235	-0.8325	0.7595
20	X3	0.3059	0.1966	1.5558	0.1406	-0.1132	0.7249
21	X4	0.2116	0.1923	1.1005	0.2885	-0.1982	0.6215
22							

Most of the variation is explained.

The overall model is significant.

None of the variables are significant.

Figure 4.27 Initial regression run on the high-multicollinearity fictitious data



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9994					
5	R Square	0.9987					
6	Adjusted R Square	0.9985					
7	Standard Error	1.0753					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	14,330.0482	4,776.6827	4,130.7842	0.0000	
13	Residual	16	18.5018	1.1564			
14	Total	19	14,348.5500				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-0.4751	0.6666	-0.7127	0.4863	-1.8883	0.9381
18	X1	0.2879	0.1869	1.5400	0.1431	-0.1084	0.6841
19	X2	0.2574	0.2628	0.9794	0.3420	-0.2997	0.8145
20	X3	0.2265	0.1841	1.2303	0.2364	-0.1638	0.6168
21							
22							

Figure 4.28 Regression run with the X$_4$ ***variable dropped on the high-multicollinearity fictitious data***

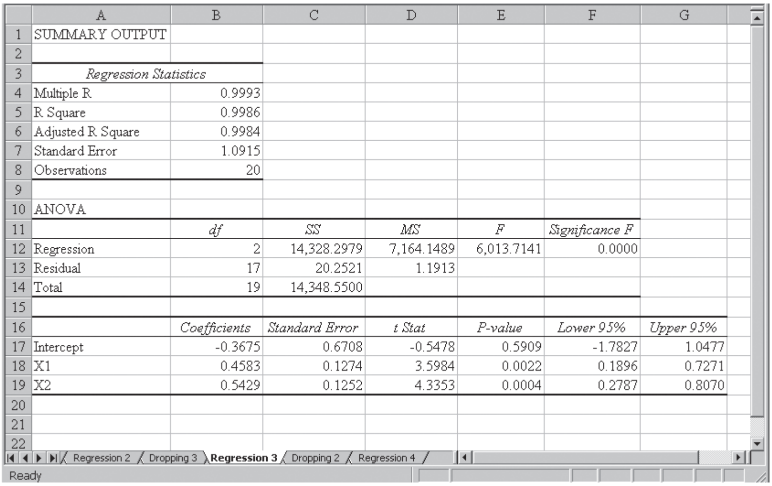
Normally, we would need to drop all four variables one at a time and record the resulting r^2 values in order to decide which to drop. However, the way the data was constructed for this example guarantees about the same impact regardless of the variable dropped, so we will simply drop X_4. The resulting multiple regression run is shown in Figure 4.28.

This time, notice the following:

- The r^2 value does not change much, going from 0.9988 to 0.9987.
- The overall model is still significant.
- Again, none of the remaining independent variables is significant.
- The slope coefficients for X_1 and X_2 change dramatically, going from negative to positive. This will end up being one of the important signs of multicollinearity.

This time, we will drop X_3. The resulting multiple regression run is shown in Figure 4.29. This time, notice the following:

- The r^2 value does not change much, going from 0.9987 to 0.9986.
- The overall model is still significant.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.9993
R Square	0.9986
Adjusted R Square	0.9984
Standard Error	1.0915
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	14,328.2979	7,164.1489	6,013.7141	0.0000
Residual	17	20.2521	1.1913		
Total	19	14,348.5500			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0.3675	0.6708	-0.5478	0.5909	-1.7827	1.0477
X1	0.4583	0.1274	3.5984	0.0022	0.1896	0.7271
X2	0.5429	0.1252	4.3353	0.0004	0.2787	0.8070

Figure 4.29 Regression run with the X$_3$ ***and*** X$_4$ ***variables dropped on the high-multicollinearity fictitious data***

- This time, the two remaining slope coefficients are significant.
- Both of the remaining slope coefficients nearly double in magnitude.

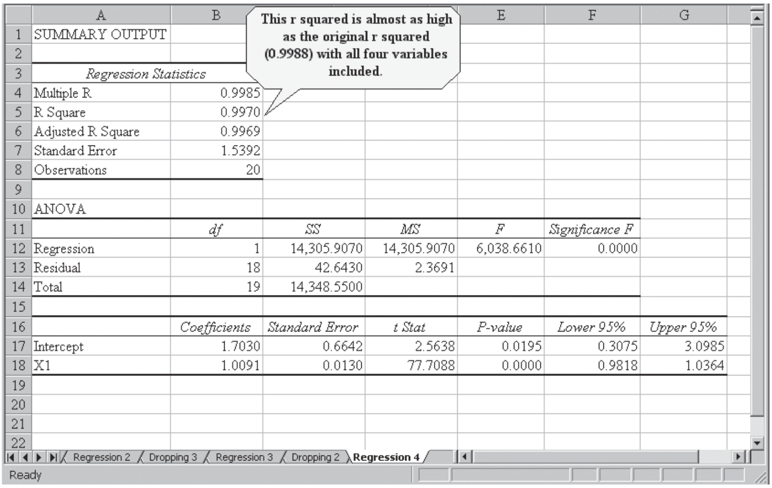
Although this model meets all our criteria, we will go ahead and drop X_2. The resulting simple regression run is shown in Figure 4.30. This time, notice the following:

- The r^2 value does not change much, going from 0.9986 to 0.9970.
- The overall model is still significant.
- The single remaining independent variable is significant.

Clearly, the model suffered serious problems relating to its high degree of multicollinearity, although the final model in Figure 4.30 no longer has a multicollinearity. Do you see why?

What Causes Multicollinearity?

Independent variables are selected based on their theoretical relationship with the dependent variable, not their statistical suitability for using in



This r squared is almost as high as the original r squared (0.9988) with all four variables included.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.9985					
5	R Square	0.9970					
6	Adjusted R Square	0.9969					
7	Standard Error	1.5392					
8	Observations	20					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	1	14,305.9070	14,305.9070	6,038.6610	0.0000	
13	Residual	18	42.6430	2.3691			
14	Total	19	14,348.5500				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	1.7030	0.6642	2.5638	0.0195	0.3075	3.0985
18	X1	1.0091	0.0130	77.7088	0.0000	0.9818	1.0364
19							
20							
21							
22							

Figure 4.30 Regression run with the X_2, X_3, and X_4 variables dropped on the high-multicollinearity fictitious data

multiple regression. Oftentimes, a natural relationship exists between these variables.

For example, in the sales forecasting model we have been discussing, two of the variables we would naturally collect are levels of advertising and competitor actions, most likely in the form of competitor advertising. It is reasonable to assume that the higher our advertising spending, the more our competitors are going to spend on advertising. That is, when our advertising spending goes up, competitor spending is likely to go up, and when our spending goes down, competitor spending is likely to go down. In other words, these two independent variables are highly correlated, and therefore we have multicollinearity. That natural multicollinearity does not mean that we should not collect data on both variables. Remember, the decision on the independent variables to start with is a theoretical decision, not a statistical decision. Rather, it simply means that both variables may not make it to the final model or, if they do, the final model will have multicollinearity between these two variables. Researchers must understand this relationship if they are to interpret the results properly. After all, if competitor advertising does not make it into the final model, they need to understand why.

Flawed data-collection methods can also introduce multicollinearity into the model. For example, if sales data were only collected from stores with a high level of competition, the data set would show a stronger relationship between advertising spending and competitor spending on advertising than would be the case if all locations were included in the sample. That would introduce an unnaturally high level of multicollinearity between the two variables.

Spotting Multicollinearity

In the previous examples, we have already seen some of the ways in which multicollinearity can be spotted. In general, all of the following are indicators of multicollinearity:

The first indicator is *a high correlation between independent variables in the correlation matrix*. Unless they have been artificially created, as shown previously, all independent variables will have some correlation between them and this correlation will show up in the correlation matrix. A rule of thumb is that a value of 0.60 or higher in the correlation matrix is an indicator of multicollinearity strong enough to be concerned about. This is the easiest rule to use, so it is recommended that you produce a correlation matrix on each data set prior to running multiple regression. Because you want a high degree of correlation between each independent variable and the single dependent variable, you should only include the independent variables in this correlation matrix so the high values between the variables and the dependent variable does not confuse its interpretation.

The second indicator is *a low tolerance*. One drawback to testing for multicollinearity using correlations is that it only spots pairwise multicollinearity because it is based on pairwise correlation. It is possible to test to see there is multicollinearity between more than two variables; that is, if two or more independent variables combined together can explain another independent variable. To do this, you run multiple regression with one of the independent variables as the dependent variable and the remaining independent variables (*not* the dependent variable) as the independent variables. The tolerance is then computed as follows:

Tolerance

$$1 - r^2$$

for this reduced multiple regression run. It is the case that the smaller the tolerance, the greater the multicollinearity regarding the independent variable being used as the dependent variable. The smallest possible value for tolerance is zero, and a good rule of thumb is that anything below 0.20 indicates a problem with multicollinearity. Two notes are in order. First, with *k* independent variables, there will be *k* measures of tolerance, as each independent variable is used, in turn, as the dependent variable. Second, tolerance requires at least three independent variables because with just two independent variables, the correlation coefficient is adequate for measuring multicollinearity. Finally, due to the difficulty of performing these repeated multiple regression runs, tolerance as a diagnostic for multicollinearity is not emphasized in this textbook.

The third indicator is *important theoretical variables that are not significant*. There are two main reasons why an important theoretical variable might not be significant: Either the theory is wrong or there is multicollinearity.[8] If you are confident that the theory is correct, then the cause is most likely that another variable is robbing the theoretically important variable of its explanatory ability—in other words, multicollinearity.

The fourth indicator is *coefficients that do not make sense theoretically*. In the sales forecast example, the theory says that increasing advertising spending should increase sales, so we would expect a positive slope coefficient for advertising spending. If that does not happen, then either the theory is wrong or, once again, multicollinearity is causing another variable to rob the theoretically important variable of its explanatory ability and therefore, in the process, altering its coefficient.

Note that it is rarely possible to evaluate coefficients theoretically beyond their signs. This is because the magnitudes of the coefficients are determined by the units of the independent variable, the units of the dependent variable, and if multicollinearity is present, the units of the collinear variables. Change any of these units and the coefficient changes. That is, measure sales in thousands of dollars rather than dollars or advertising spending in minutes of television time instead of dollars and the advertising spending coefficient changes. However, regardless of the units, the sign of the slope coefficient should behave according to theory.

The fifth indicator is when you notice that *dropping a variable causes dramatic shifts in the remaining coefficients*. If there is no multicollinearity, then the explanatory power of a variable does not change as other variables come and go. We saw this in the previous example artificially created without multicollinearity. Therefore, when coefficients shift as other variables come and go, it is an indicator of multicollinearity. Thus we see that the greater the shift, the larger the multicollinearity. As discussed previously, the biggest concern is when the coefficients change signs or when values change by an order of magnitude.

The sixth indicator is when you notice that *dropping a nonoutlier observation causes dramatic shifts in the coefficients*. Although rarely used in practice, dropping a single observation that is not an outlier should not cause much of a shift in the coefficients. When it does, that is a sign of multicollinearity. Of course, this is also a sign that the observation is possibly an outlier so you must be careful in its use. One of the reasons this is not used much in practice is that it is the least likely of all the approaches to generate an observable effect, plus researchers rarely wish to drop useful data.

Treating Multicollinearity

When multicollinearity is present, any or all of the following can be used to treat it.

Fix the sampling plan. It goes without saying that when the multicollinearity was introduced by a poor approach to gathering the data, new data should be collected using a better sampling plan. It is much better to work with good data than it is to try to fix bad data.

Transform the collinear variables. Multicollinearity is a *linear* correlation between two (or more) variables. Transforming one or more of these variables in a nonlinear fashion can reduce or eliminate the multicollinearity. Nonlinear transformations include taking the log, squaring, and taking the square root. Multiplying by a number or adding a number are both linear transformations and will not change multicollinearity.

The trouble with transforming the data is that it changes the data. Changing the data makes it tougher to theoretically interpret the data. For example, we know that a positive slope coefficient for advertising indicates that spending more money on advertising increases sales. If sales and advertising are both measured in thousands of dollars, then a coefficient of 0.50 would indicate that for every additional thousand dollars spent on advertising, sales go up by $500. But what would the coefficient mean if we were using the log or square root of advertising dollars? For this reason, variable transformations are usually only used in models intended for prediction where there is little or no interest in understanding the underlying processes.

Transform the data set. An advanced statistical process called factor analysis can be used to transform a collinear data set, or any subset of that data set, into new, uncorrelated variables that explain the same variation as the original data set. However, these new and uncorrelated variables are even more manipulated than the simple variable transformations discussed previously, making them that much harder to theoretically interpret. Students interested in this topic should consult an advanced statistical textbook such as Philip Bobko's *Correlation and Regression: Principles and Applications for Industrial/Organizational Psychology and Management* (1995, McGraw-Hill). Factor analysis was the technique used to create the completely uncorrelated variables used in an earlier example in this chapter.

Use an advanced multiple regression approach. A type of multiple regression called *ridge regression* is more adept at working with collinear data. Excel is not able to perform ridge regression.

Drop one of the collinear variables. After all, if the two variables are explaining the same, or mostly the same, variation, it makes little sense to include both in the model. When the multicollinearity between two variables is too high, it is rarely the case that both will end up being significant. Thus the model development procedures discussed previously will automatically cause one in the pair of collinear variables to be dropped. Even if they both end up being significant, they may end up biasing the coefficients to such an extent that one of them must be dropped so the remaining coefficients make theoretical sense.

Do nothing. If the collinear variables are all significant, then they help improve the fit of the model. If the model is to be used mainly for prediction, then any theoretical problems with the coefficients will not be a problem. Even when the model is to be used for understanding, multicollinearity does not always have such a strong impact as to cause the coefficients not to make theoretical sense.

Autocorrelation

One of the assumptions of regression, both simple and multiple, is that the error terms (ε) are independent of each other. Stated another way, ε_i is uncorrelated with ε_{i-1} or ε_{i-2} or ε_{i-3} and so on. When correlation between one or more of these error terms exists, it is called *autocorrelation*.

Autocorrelation is only an issue when we have *time-series* data—that is, data that were measured at different points in time. For example, if we have quarterly measures of demand for several years, it is likely that the demand for any quarter was related to quarterly demand a year ago, so it is likely that ε_i is correlated with ε_{i-4}. This is called a lag four correlation.

This issue of lagged correlation can easily be illustrated with a figure. Figure 4.31 shows a one-period lag in the correlation of the error terms.

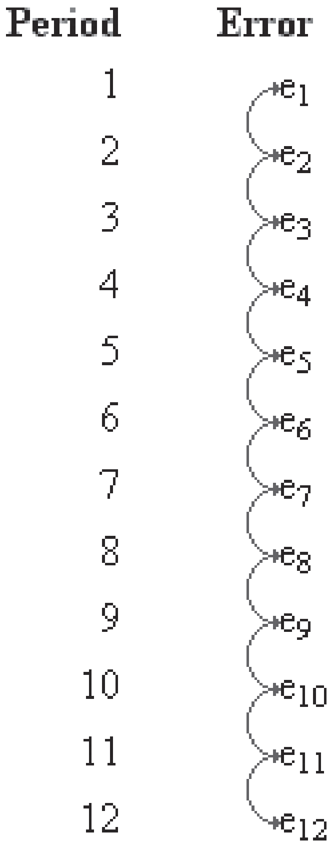


Figure 4.31 An illustration of lag one correlation

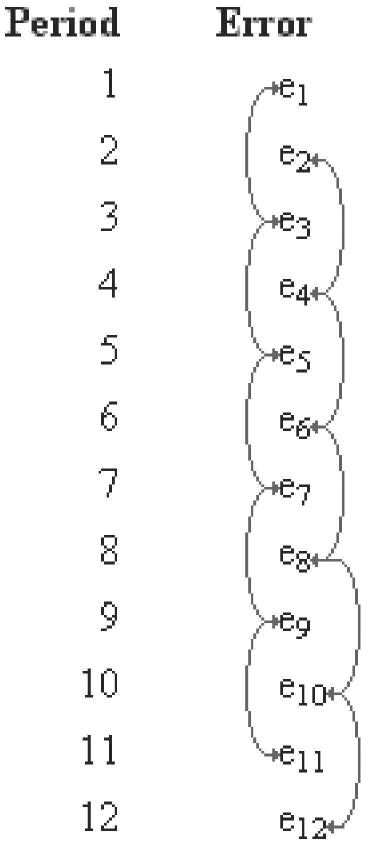


Figure 4.32 An illustration of lag two correlation

This is called *first-order autocorrelation*. Figure 4.32 shows a two-period lag in the correlation. This is called *second-order autocorrelation*.

When the data are not time series, there is no reason to be concerned about autocorrelation. After all, there is not likely to be any relationship between the different dependent variable observations, so there is unlikely to be any correlation of the error terms. Although it is technically possible for the error terms to be correlated for non-time- series data (called *cross sectional* data), we need not be concerned with this rare occurrence. After all, if the data are not time-series data, there is no specific order for the data. Therefore, we could simply rearrange the sequence of the data and alter any lag correlations of the error terms.

Durbin-Watson Test

Statistical software, like SPSS or SAS, can compute a Durbin-Watson test to easily spot first-order autocorrelation. The hypotheses for the Durbin-Watson test are the following:

$$H_0 : \rho_1 = 0$$
$$H_0 : \rho_1 \neq 0$$

Of course, we can also perform a one-tailed version of the test. The hypotheses use a ρ_1 because the Durbin-Watson test can only spot first-order autocorrelation.

The calculated Durbin-Watson statistic value can take on values between zero and four. A value of two, which is in the middle of this range, indicates no autocorrelation. A value of zero indicates positive autocorrelation and a value of four indicates negative autocorrelation. When a table is not available, a rule of thumb is that values of *d* between 1.5 and 2.5 indicates no autocorrelation. The Durbin-Watson test statistic *d* is defined with the following equation:

Durbin-Watson Test Statistics

$$d = \frac{\sum_{i=2}^{n} \left(e_i - e_{i-1}\right)^2}{\sum_{i=1}^{n} e_i^2}$$

A couple of notes are in order regarding this formula. First, notice that although we are testing for the significance of a correlation, neither a sample (*r*) or assumed population (ρ) correlation coefficient is used in the calculations. Second, note that the top of the formula is summed from "$i = 2$ to n" whereas the bottom is summed from "$i = 1$ to n." This is because the top measures the lagged squares and $e_1 - e_0$ is not defined.

The Durbin-Watson test first appears to work differently than the other hypothesis tests we have looked at. For a one-tailed test,[9] the critical values (d_L and d_U)[10] do not divide the distribution into acceptance and rejection regions. Rather, they divide the distribution into five different regions, as noted in Table 4.1.

Table 4.1. Durbin-Watson Outcome Regions

Area	Outcome
$0 – d_L$	Positive autocorrelation
$d_L – d_U$	Test is inconclusive
$d_U – 4\text{-}d_U$	No autocorrelation
$4\text{-}d_U – 4\text{-}d_L$	Test is inconclusive
$4\text{-}d_L – 4$	Negative autocorrelation

The lack of clear boundaries between the acceptance and rejection regions is, in fact, not due to the Durbin-Watson test working differently than other hypothesis tests. Rather, the actual boundaries of the test depend on the regression coefficients. Because printed tables cannot easily reflect this, the ranges where the test is inconclusive represent the range of possible values from the table for different values of the regression coefficients.

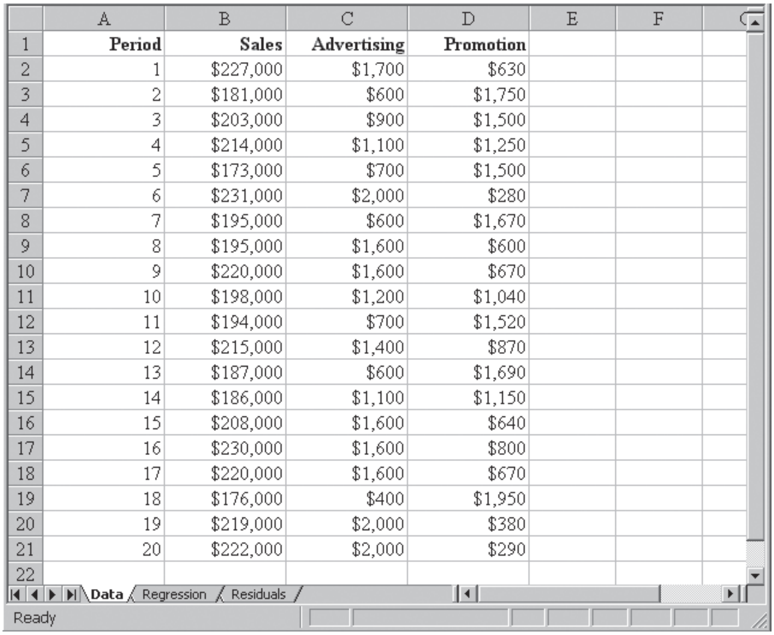
Example

It is very common for businesses to collect data, such as sales data, over time. Because these data are collected over time, it often has autocorrelation. We will explore this issue with an example.

Figure 4.33 shows sales data for 20 periods, along with the advertising and promotion data for the same period that will be used to explain the sales. That is, Sales is the dependent variable and Advertising and Promotion are the independent variables. This figure shows the data in Excel, and it is saved in Durbin-WatsonExample.xls if you wish to experiment with the data; however, the regression analysis will be performed using SPSS.

Figure 4.34 shows the resulting multiple regression run. Note that the overall model is significant, and both of the independent variables are significant. This model explains 79.1 percent of the variation in sales.

The value of *d* from the model summary area is 2.087. Because we do assume that the autocorrelation is either positive or negative and *d* is between 1.5 and 2.5, we use a two-tailed test and conclude that there is no autocorrelation.



Period	Sales	Advertising	Promotion
1	$227,000	$1,700	$630
2	$181,000	$600	$1,750
3	$203,000	$900	$1,500
4	$214,000	$1,100	$1,250
5	$173,000	$700	$1,500
6	$231,000	$2,000	$280
7	$195,000	$600	$1,670
8	$195,000	$1,600	$600
9	$220,000	$1,600	$670
10	$198,000	$1,200	$1,040
11	$194,000	$700	$1,520
12	$215,000	$1,400	$870
13	$187,000	$600	$1,690
14	$186,000	$1,100	$1,150
15	$208,000	$1,600	$640
16	$230,000	$1,600	$800
17	$220,000	$1,600	$670
18	$176,000	$400	$1,950
19	$219,000	$2,000	$380
20	$222,000	$2,000	$290

Figure 4.33 Data for Durbin-Watson example in Excel

The value of *d* can be approximated using the following formula:

Approximating Durbin-Watson *d*

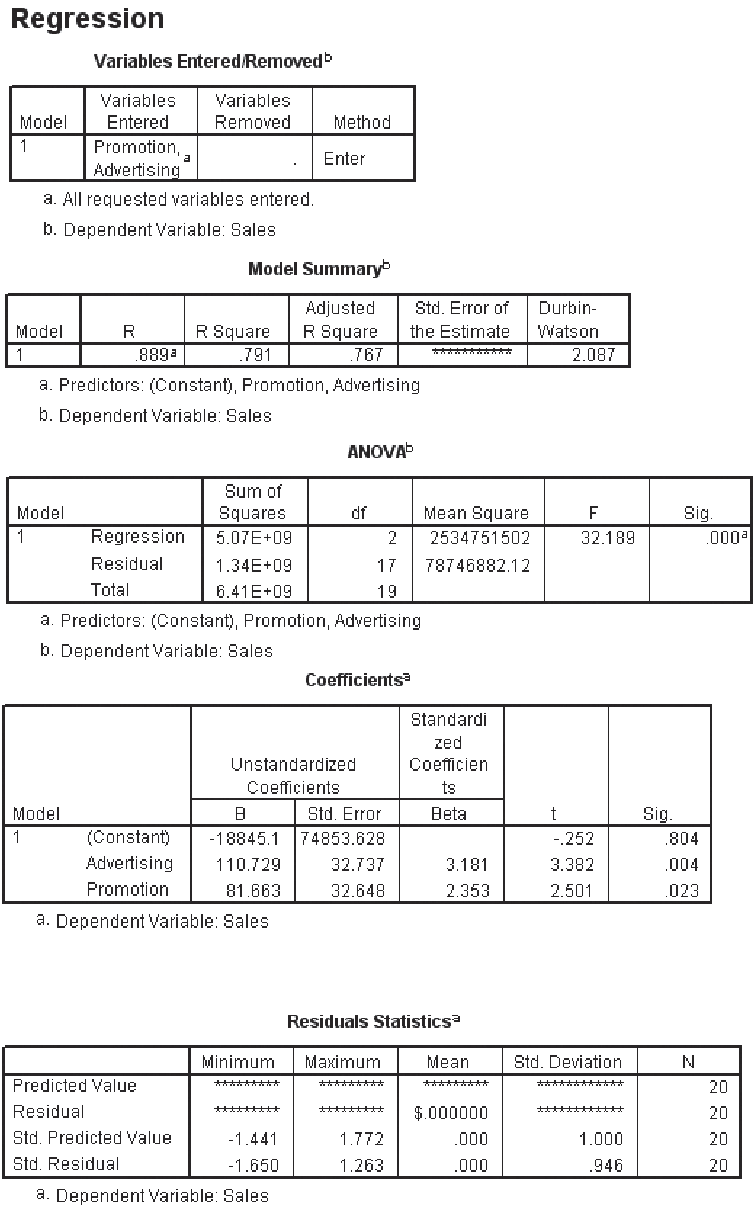
$$2(1-r)$$

where r is the correlation coefficient that measures the association between successive residuals.

Although not shown, the regression model for the Durbin-Watson example was computed in Excel and the residuals were saved. The *Residuals* tab of the Durbin-WatsonExample.xls worksheet computes the correlation coefficient between successive residuals. That value is –0.06491. Applying the approximation formula above, we have 2[1 – (–0.06491)], which equals 2(1 + 0.06491), which is about 2.12983. That is close to the 2.0869 calculated by SPSS.

Treating Autocorrelation

There are two things you can do to minimize the possibility of having autocorrelation when dealing with time-series data. The first is to



Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Promotion, Advertising[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Sales

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.889[a]	.791	.767	***********	2.087

a. Predictors: (Constant), Promotion, Advertising

b. Dependent Variable: Sales

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5.07E+09	2	2534751502	32.189	.000[a]
	Residual	1.34E+09	17	78746882.12		
	Total	6.41E+09	19			

a. Predictors: (Constant), Promotion, Advertising

b. Dependent Variable: Sales

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-18845.1	74853.628		-.252	.804
	Advertising	110.729	32.737	3.181	3.382	.004
	Promotion	81.663	32.648	2.353	2.501	.023

a. Dependent Variable: Sales

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	*********	*********	*********	************	20
Residual	*********	*********	$.000000	************	20
Std. Predicted Value	-1.441	1.772	.000	1.000	20
Std. Residual	-1.650	1.263	.000	.946	20

a. Dependent Variable: Sales

Figure 4.34 The results of running multiple regression using SPSS

transform the data. When variables are measured in dollars over time, state those dollars in a constant unit, such as discounted dollars. When the dollars are left in their raw form,[11] changes in the buying power of the dollar over time are built into the dependent variable but will not be accounted for by any independent variable. Furthermore, those changes over time are fairly regular and therefore correlated with one another, leading to autocorrelation. Converting to discounted dollars removes this time-based source of variation. In business, this is by far the most common transformation. Another transformation that works well, especially with economic and financial data, is to restate the variables as a percentage change.

When one or more independent variables are measured in their raw form, they have a built-in variation that is not being used to explain the dependent variable. As a result, this variation is transferred to the error term, where it becomes autocorrelation. Converting the variables to discounted dollars or percentage change removes this source of variation.

The second approach is to add a new independent variable, here called Period, where Period is simply a measure of the changes over time. This variable will explain this regular variation and keep it from reaching the error term. If Period is significant, then autocorrelation exists and Period treated it. If Period is not significant, then autocorrelation is not present.

The period variable should be a linear variable when every one-unit change in periods causes the same change in the period variable. The most common approach is to label the first period 1, the second 2, and so on. With annual data, using the actual year number would also work. For quarterly data, you could use sequential period numbers or year numbers with a.00, .25, .50, and .75 added for the different periods within each year.

You must be careful in labeling your periods. It is often tempting to use a labeling scheme that violates the assumption of equal units between the periods. For example, none of the following methods is appropriate:

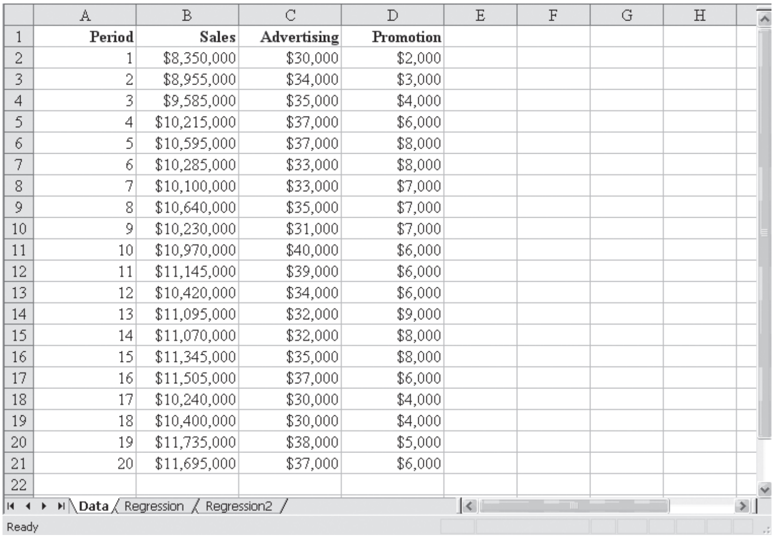
- With monthly data, using 2002.01, 2002.02, . . . ,2002.11, 2002.12, 1991.01, and so on
- With quarterly data, using 2002.1, 2002.2, 2002.3, 2002.4, 2003.1, and so on
- With quarterly data, using 1, 2, 3, 4, 1, 2, 3, 4, and so on

With each of these, the gap between some sequential pairs of periods is different from other sequential pairs, and this nonlinear ordering of the periods violates the linearity assumption of regression.

Second Durbin-Watson Example

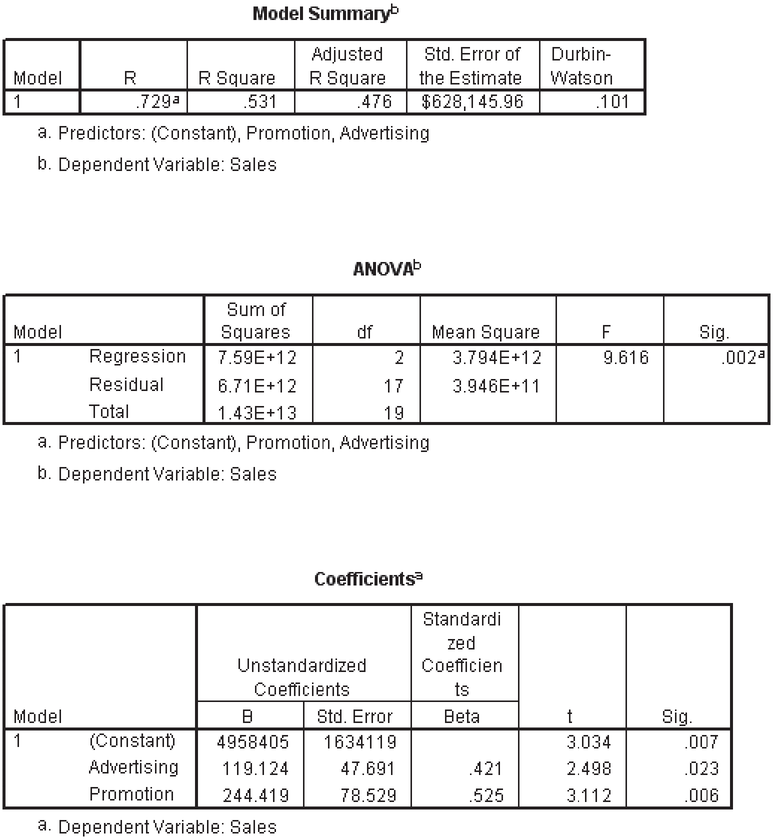
Figure 4.35 shows a set of data especially constructed to contain positive autocorrelation. The data are stored in the Excel file DW-2.xls for easy manipulation. Figure 4.36 shows the initial multiple regression run on this data. Note that the overall model is significant and both independent variables are also significant. However, the model only explains 53.1 percent of the variation in sales. This model has positive autocorrelation, as shown by the Durbin-Watson statistic of 0.101 shown in Figure 4.36.

To correct the autocorrelation, the period variable is added to the model as a third independent variable. That regression run is shown in Figure 4.37. Again, the overall model is significant and all three independent variables are significant. The explanatory power of the model goes up from 53.1 percent to 98.7 percent. This model has much less



Period	Sales	Advertising	Promotion
1	$8,350,000	$30,000	$2,000
2	$8,955,000	$34,000	$3,000
3	$9,585,000	$35,000	$4,000
4	$10,215,000	$37,000	$6,000
5	$10,595,000	$37,000	$8,000
6	$10,285,000	$33,000	$8,000
7	$10,100,000	$33,000	$7,000
8	$10,640,000	$35,000	$7,000
9	$10,230,000	$31,000	$7,000
10	$10,970,000	$40,000	$6,000
11	$11,145,000	$39,000	$6,000
12	$10,420,000	$34,000	$6,000
13	$11,095,000	$32,000	$9,000
14	$11,070,000	$32,000	$8,000
15	$11,345,000	$35,000	$8,000
16	$11,505,000	$37,000	$6,000
17	$10,240,000	$30,000	$4,000
18	$10,400,000	$30,000	$4,000
19	$11,735,000	$38,000	$5,000
20	$11,695,000	$37,000	$6,000

Figure 4.35 Data especially constructed to contain positive autocorrelation



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.729[a]	.531	.476	$628,145.96	.101

a. Predictors: (Constant), Promotion, Advertising

b. Dependent Variable: Sales

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.59E+12	2	3.794E+12	9.616	.002[a]
	Residual	6.71E+12	17	3.946E+11		
	Total	1.43E+13	19			

a. Predictors: (Constant), Promotion, Advertising

b. Dependent Variable: Sales

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	4958405	1634119		3.034	.007
	Advertising	119.124	47.691	.421	2.498	.023
	Promotion	244.419	78.529	.525	3.112	.006

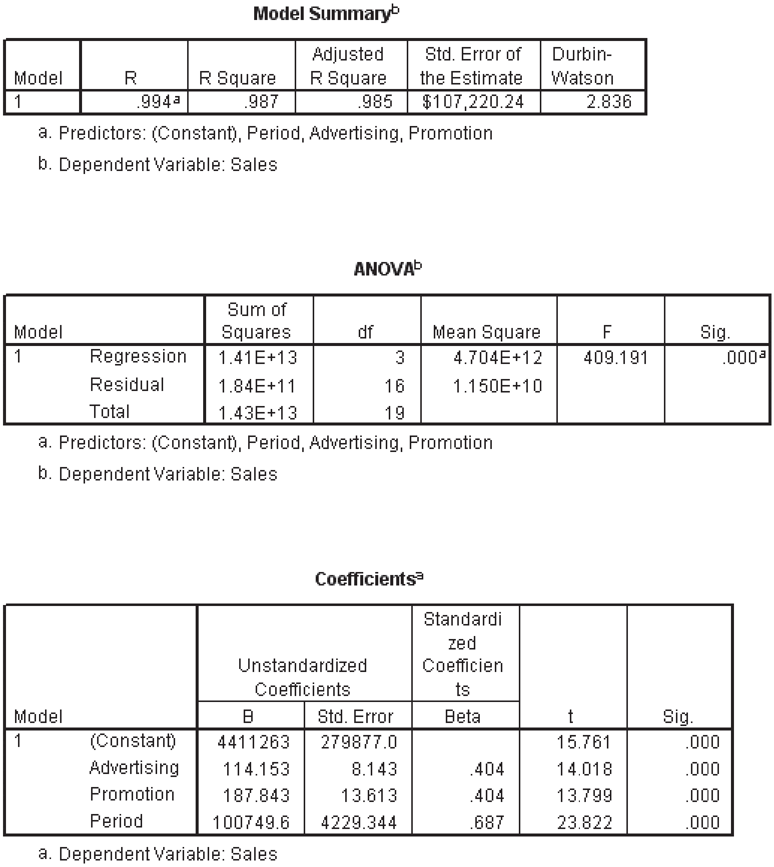
a. Dependent Variable: Sales

Figure 4.36 Initial regression on this fictitious data

autocorrelation, as shown by the Durbin-Watson statistic of 2.836 shown in Figure 4.37.

Heteroscedasticity

One of the assumptions of regression is that the error terms have equal variance. We called this homoscedasticity. Violating the assumption of homoscedasticity is called heteroscedasticity.[12] Although heteroscedasticity is a violation of the assumptions of regression, it is a fairly minor violation relative to multicollinearity and autocorrelation. However, when heteroscedasticity is high, the researcher may need to build a separate



Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.994[a]	.987	.985	$107,220.24	2.836

a. Predictors: (Constant), Period, Advertising, Promotion

b. Dependent Variable: Sales

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.41E+13	3	4.704E+12	409.191	.000[a]
	Residual	1.84E+11	16	1.150E+10		
	Total	1.43E+13	19			

a. Predictors: (Constant), Period, Advertising, Promotion

b. Dependent Variable: Sales

Coefficients[a]

		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	4411263	279877.0		15.761	.000
	Advertising	114.153	8.143	.404	14.018	.000
	Promotion	187.843	13.613	.404	13.799	.000
	Period	100749.6	4229.344	.687	23.822	.000

a. Dependent Variable: Sales

Figure 4.37 Redoing the regression model by adding period as a variable

regression model for each range of the error term. That is, the researchers need to build one model for the range where the variance is high and another for where it is low.

We will not be using a hypothesis test approach to testing for heteroscedasticity; rather, we will be using a visual approach. Specifically, we will be looking at a plot of the residuals (error terms) once the model has been built. Residuals are shown on the Y-axis. We must produce a plot for each independent variable that makes it into the final model. Fortunately, Excel makes this easy.

Example

As we saw earlier, our Celebrities.xls worksheet has six independent variables and all of them are significant except for Earnings. We will now drop Earnings, produce the final model, and get a plot of the residuals. To add the plots to the output, we simply check *Residual Plots* in the *Regression* dialog box. Excel produces the plots and stacks them on top of each other, as shown in Figure 4.38. You click on the plot you wish to see and that brings it to the front. Figures 4.39 through 4.42 show the individual plots.

Figure 4.39 shows Residuals plotted against Income Rank. As Income Rank increases, the residuals appear to be randomly distributed when

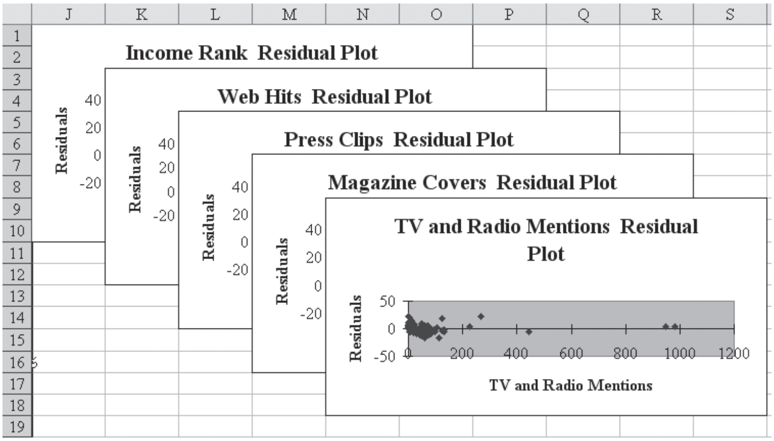


Figure 4.38 The stacked plot of residuals

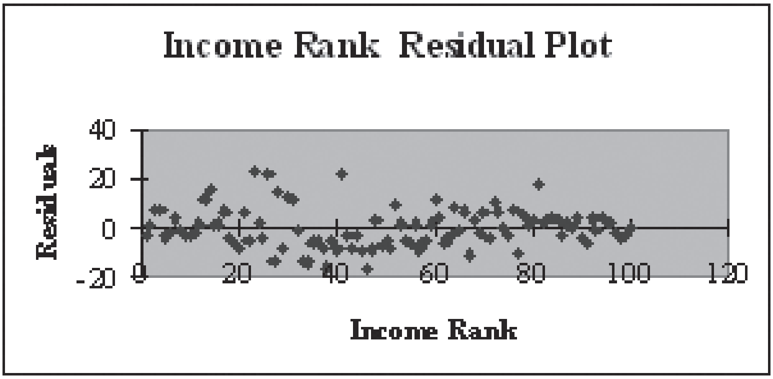


Figure 4.39 Residuals plotted against Income Rank

fairly consistent variability, so heteroscedasticity is not an issue. Figure 4.40 shows Residuals plotted against Web Hits. Clearly, the residuals are neither randomly distributed nor have equal variability, a clear sign of heteroscedasticity. Figure 4.41 shows Residuals plotted against Press Clippings. As with Figure 4.40, heteroscedasticity is clearly evident.

Figure 4.42 shows Residuals plotted against Magazine Covers. With this variable, it is much harder to visually gauge heteroscedasticity. Magazine Covers can take only a limited number of discrete values and so the residual values are bunched above these few values. Dummy variables have this problem as well as discrete variables, only with dummy variables, there are only two columns of points. Although the variability clearly declines as the number of Magazine Covers decreases, this may be

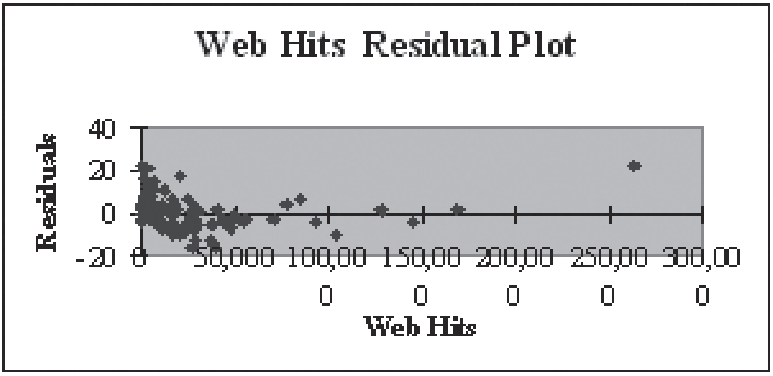


Figure 4.40 Residuals plotted against Web Hits

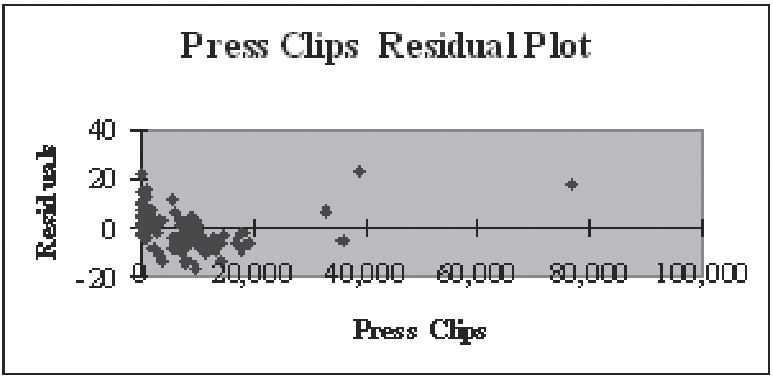


Figure 4.41 Residuals plotted against Press Clippings

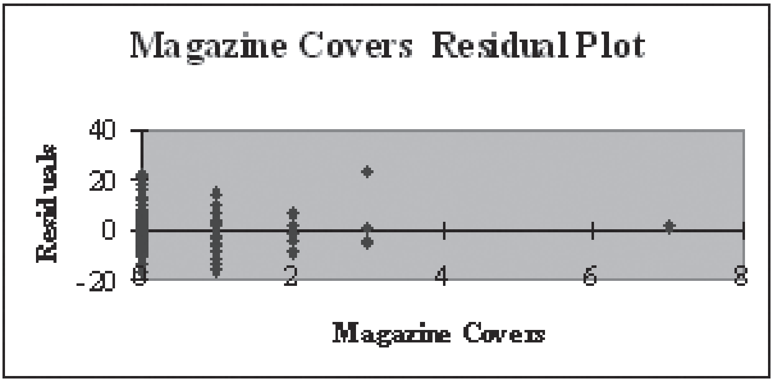


Figure 4.42 Residuals plotted against Magazine Covers

due more to the lower number of observations than to a decline in the variance. Therefore, the best we can say regarding heteroscedasticity is that we are uncertain.

Figure 4.43 shows Residuals plotted against TV and Radio Mentions. The values on the left side of the chart appear to be randomly distributed with fairly constant variability. The very few observations on the right clearly have less variability but there are too few values to make any real judgment. These could simply be outliers. The most reasonable conclusion is that the bulk of the data shows no heteroscedasticity.

The problem with gauging heteroscedasticity from these individual plots is that they do not show the error terms relative to all the data at once. It may be that large differences relative to one variable may not

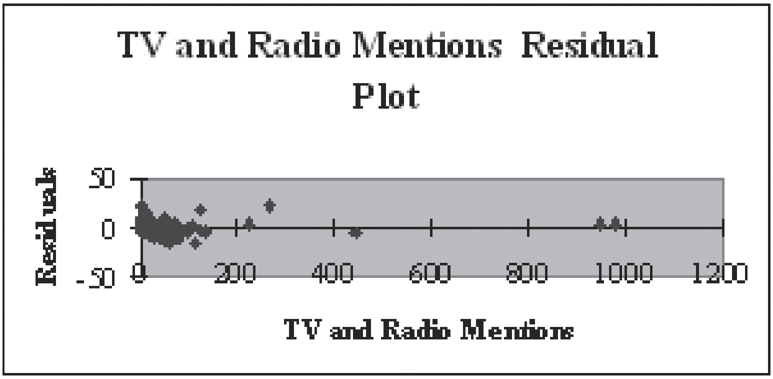


Figure 4.43 Residuals plotted against TV and Radio Mentions

be too large when other variables are included in the comparison. The way to gauge heteroscedasticity across all the variables at once is to use a chart that Excel does not automatically compute—a plot of the standardized residuals against estimated values of the dependent variable. A homoscedastic model will display a cloud of dots with no visible pattern, whereas heteroscedasticity will be characterized by a pattern, such as a funnel shape, indicating greater error as the dependent increases.

Using the Celebrities.xls worksheet discussed previously, multiple regression was rerun, using SPSS and requesting a plot of the residuals. This is shown in Figure 4.44. The Y-axis is the standardized residuals and the X-axis is the standardized predicted values. With homoscedasticity, you would expect the spread of the data to be fairly consistent as you move from left to right. Because this data has a fairly narrow spread on both ends with a wider spread in the middle, it is an indication of heteroscedasticity. This measure looks at all the variables at once. This plot can be produced in Excel by having regression save the residuals, manually build a formula to calculate the predicted values, and then manually produce the plot. This exercise is left for interested readers.

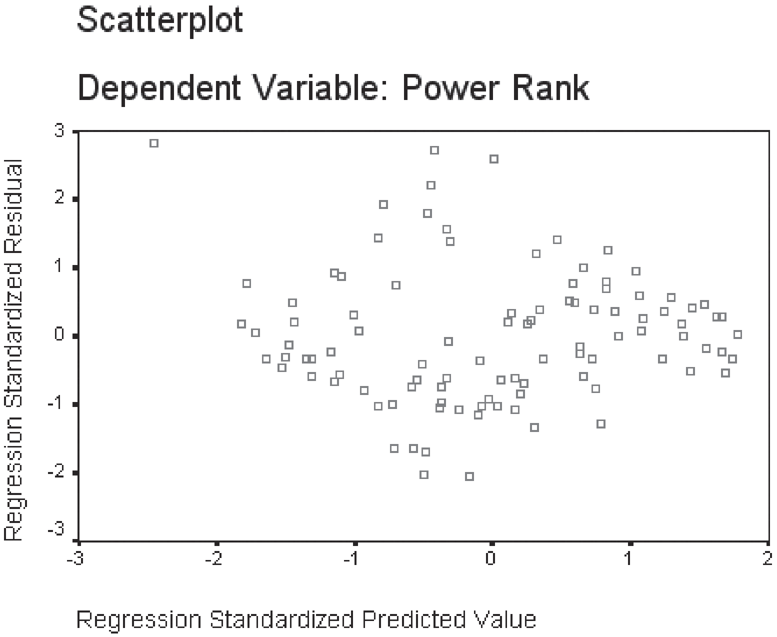


Figure 4.44 Plot of the residuals against predicted Y-values

Treating Heteroscedasticity

Heteroscedasticity is a violation of the assumptions of regression; however, its effect is much smaller than either multicollinearity or autocorrelation. Heteroscedasticity causes the least squares estimation method to be less efficient. One approach to dealing with heteroscedasticity is to use *weighted least squares*. Excel is not able to perform this, so we will not explore it further.

A second approach is to transform the variable exhibiting heteroscedasticity using a nonlinear transformation such as squares, square roots, or logs. These transformations are particularly effective when the data show high variability on one side of the residual plot and low variability on the other. However, a nonlinear transformation introduces all the theoretical interpretation concerns raised earlier.

A third approach is to simply realize that heteroscedasticity exists in the multiple regression model but to otherwise ignore it in the calculation of the model. Given the minor impact of heteroscedasticity, this is the approach often taken by researchers.

Summary

In this chapter we looked at model building, incorporating qualitative data, and testing the validity of the model. You should now be ready to build your own multiple regression models.

Notes

Introduction

1. Jiawi Han and Micheline Kamber. (2006). *Data mining: Concepts and techniques* San Francisco: Elsevier.

Chapter 1

1. Damodaran Online. The data page. Retrieved August 15, 2010, from http://pages.stern.nyu.edu/~adamodar/New_Home_Page/data.html as part of a larger data set.
2. U.S. Office of Personnel Management. *Federal Civilian Workforce Statistics: The Fact Book, 2007 Edition*. Retrieved from http://www.opm.gov/feddata/factbook/2007/2007FACTBOOK.pdf
3. American Public Transportation Association. *2010 Public Transportation Fact Book*, 61st ed. April 2010. Retrieved from http://www.apta.com/resources/statistics/Documents/FactBook/APTA_2010_Fact_Book.pdf
4. *Unlinked passenger trips* refers to the total number of passengers who board public transit vehicles. Each passenger is counted each time that person boards a vehicle even though the boarding may be the result of a transfer from another route while on the same journey. Thus unlinked passenger trips will be larger than actual ridership.
5. *Outlier* does not necessarily mean mistake or error. Here, the data for New York City *is* absolutely correct; it is just well outside the range of any other observation.
6. The actual calculation for this specific data set is –0.042, which is extremely close to zero. The exact value would depend on the number of observations included, but it would always be close to zero.
7. The assignment of Age as X and Tag Number as Y is completely arbitrary and for correlation does not matter. The resulting value for the correlation coefficient would be exactly the same if Age were assigned as Y and Tag Number were assigned as X. The proof of this is left to interested readers.
8. SPSS stands for Statistical Program for the Social Sciences, but everyone just calls it SPSS.
9. It actually does not hurt anything to include the cells with the column labels (e.g., "=CORREL(A1:A8,B1:B8)"), as they are just ignored by Excel.

10. The “bi-” in “bivariate” comes from the Latin *bis*, meaning “twice” and *bini*, meaning “in twos.”
11. It is helpful if the reader has a general knowledge of hypothesis testing on means and proportions as well as the Student *t*-distribution. However, students without familiarity with these topics should be able to read and comprehend most of this section.

Chapter 2

1. With simple regression, it is not uncommon to leave out the first subscript of the *X*s; that is, write $X_{1,1}$ as X_1, $X_{1,2}$ as X_2, and so on. If you do this, you must be careful as it can lead to confusion in multiple regression.
2. There is a second set of equations that can be used to calculate simple regression coefficients. Normally, we would avoid giving two equations that accomplish exactly the same thing. However, you may run into this set of equations in other courses—such as operation management or forecasting—so they are presented here for completeness:

$$b_0 = \overline{Y} - b_1\overline{X}$$

$$b_1 = \frac{\sum X \cdot Y - n \cdot \overline{X} \cdot \overline{Y}}{\sum X^2 - n\overline{X}^2}$$

This set of equations gives the same results and they have no computational advantage.

3. ANOVA is short for *an*alysis *o*f *va*riance.
4. This statement assumes α = 0.05, as it usually does in business statistics.

Chapter 3

1. The term is the combination of *homo*, from the Latin *homos* meaning one and the same or similar, and the Greek *skedastokos*, meaning able to disperse. Thus the term refers to having equal or the same variances.
2. The assumption of normally distributed error terms uncorrelated with one another automatically implies the independence of the error terms.
3. Interested students may wish to repeat our analysis using any or all of the counts for bronze, silver, or gold as the dependent variable to see how they compare with the results presented here.
4. The value of this coefficient is actually $4.69047952958057 \times 10^{-09}$ or 0.00000000469047952958057. It is this low not because the variable is

unimportant but rather because the units used to measure income (dollars) are so large relative to the units used to measure rank. As a general rule, you cannot gauge the importance of a variable by the magnitude of its coefficient.

5. Again, the value is small due to the units used to measure Web Hits.
6. Recall from the last chapter that correlation is only defined between pairs of variables.
7. In all these calculations, results are rounded for display but have been carried out in all the calculations.
8. Once any two of *SSR*, *SSE*, and *SST* are computed, you can always find the third using this formula.
9. Highlight the cells containing the formulas, right-click with the mouse, and select *Copy*. With the formulas still highlighted, right-click again and under the Paste options will be an icon of a clipboard with a 123 in the bottom right corner. Select this Paste option. This replaces the formulas with their current value.
10. As we will see later, in special cases, the overall model can be significant even when none of the independent variables is significant or at least when they are all included in the model.
11. "Auto" comes from the Greek *autos*, meaning same or self. Thus autocorrelation refers to correlation with oneself. In this case, the error terms are being correlated with themselves; more specifically, the correlation between pairs of error terms is taken at a constant interval.
12. The calculations are not exactly this straightforward. Rejecting 10 true null hypotheses (200 × 0.05) due to sampling error alone requires that all 200 null hypotheses be true—that is, that all 200 variables be insignificant. If most of the variables were truly significant as they were in this case, then the number of rejections of true null hypotheses would be correspondingly low because rather than having (200 × 0.05) we would have a number much smaller than 200 in this calculation.
13. Because of the way Excel handles *p*-values, you must divide the *p*-value by 2 for a one-tailed test.

Chapter 4

1. Most statistical packages support all or most of these. It is up to the operator to select the actual procedure to be used.
2. Because r^2 always goes up when you add variables, it will always go down when you drop a variable. However, when dropping an insignificant variable, this drop is often slight and may be hard to pick up if you normally format r^2 to four or five decimal points.

3. The values of one and zero make the math easier to understand and the model easier and are traditionally used, but any two numbers could be used and would have the same overall effect.
4. Again, this simply makes understanding the model easier. We could as easily code the data the other way—that is, zero for when the event happened and one when it did not happen.
5. Or planes when there are two nondummy independent variable, or hyperplanes when there are three or more nondummy independent variables.
6. http://www.presidentialelection.com/follow_the_money/ accessed on April 9, 2011.
7. For the interested reader, this was accomplished using an advanced statistical procedure called *factor analysis*. The operation of factor analysis is beyond the scope of this textbook. Additionally, factor analysis cannot be performed using Excel. The author created this data set using SPSS. Interested readers are referred to an advanced statistical reference book, such as Bobko (1995), *Correlation and Regression: Principles and Applications for Industrial/Organizational Psychology and Management*. New York: McGraw-Hill.
8. Of course, this can also be caused by a bad sample.
9. For a two-tailed test, we double the alpha value shown in the table, and the ranges $0 - d_L$ and $4\text{-}d_L - 4$ simply become rejection regions.
10. Refer to an advanced statistics textbook for Durbin-Watson tables.
11. In finance, this raw form is called *nominal* dollars.
12. Recall from earlier that part of this term comes from the Greek *skedastokos*, meaning able to disperse. The *hetero* comes from the Latin *heteros*, meaning other than usual, other, or different.

Index

OTHER TITLES IN OUR QUANTITATIVE APPROACHES TO DECISION MAKING COLLECTION

Donald Stengel, California State University, Fresno, Editor

- *Business Applications of Multiple Regression* by Ronny Richardson
- *Operations Methods: Waiting Line Applications* by Kenneth Shaw
- *Regression Analysis: Understanding and Building Business and Economic Models Using Excel* by J. Holton Wilson, Barry P. Keating, and Mary Beal-Hodges
- *Working With Excel: Refreshing Math Skills for Management* by Pricilla Chaffe-Stengel and Donald N. Stengel
- *Decision Analysis for Managers* by David Charlesworth
- *Multi-Objective Decision Analysis: Managing Trade-Offs and Uncertainty* by Clinton W. Brownley
- *Integrated Management of Processes and Information* by Kenneth Shaw
- *Business Applications of Operations Research* by Nag Bodhibrata
- *Regression Analysis: Unified Concepts, Practical Applications, Computer Implementation* by Bruce Bowerman, Emily Murphree, and Richard T. O'Connell
- *Experimental Design: Unified Concepts, Practical Applications, and Computer Implementation* by Bruce Bowerman, Emily Murphree, and Richard T. O'Connell

Business Applications of Multiple Regression

Second Edition

Ronny Richardson

This second edition of *Business Applications of Multiple Regression* describes the use of the statistical procedure called multiple regression in business situations, including forecasting and understanding the relationships between variables. The book assumes a basic understanding of statistics but reviews correlation analysis and simple regression to prepare the reader to understand and use multiple regression.

The techniques described in the book are illustrated using both Microsoft Excel and a professional statistical program. Along the way, several real-world data sets are analyzed in detail to better prepare the reader for working with actual data in a business environment.

This book will be a useful guide to managers at all levels who need to understand and make decisions based on data analysis performed using multiple regression. It also provides the beginning analyst with the detailed understanding required to use multiple regression to analyze data sets.

Dr. Ronny Richardson is a professor of operations management at Kennesaw State University. His research and teaching interests are in the areas of operations management, statistics, project management, and computers. He is the author of 20 books and over 500 published articles. He has consulted with several major companies in the areas of production and inventory control. Prior to teaching, he was an executive for Georgia Power Company.

QUANTITATIVE APPROACHES TO DECISION MAKING COLLECTION
Donald N. Stengel, *Editor*

e-ISBN 978-1-63157-060-5

90000

9 781631 570605

BEP BUSINESS EXPERT PRESS