Manuscript for

MANAGERIAL STATISTICS:

A CASE-BASED APPROACH

STATA EDITION

Peter Klibanoff,

Kellogg School of Management, Northwestern University

Alvaro Sandroni,

Kellogg School of Management, Northwestern University

Boaz Moselle,

Director, LECG Ltd., London

Brett Saraniti,

Hawaii Pacific University

Revised 2012

Contents

ACKNOWLEDGEMENTS

In addition to the people acknowledged in the first edition of the text, we would like to thank Shuyan Wu for her excellent assistance in helping us revise the book to use Stata statistical software. Thanks to Reza Kheirandish for identifying some typos in the previous edition and to Josh Cherry for some additional editing. We also thank Andrew Sfekas and Florian Zettelmeyer for their help and advice on building the custom menu in Stata. Stata is a product of StataCorp LP, and Releases 11 and 12 (StataCorp. 2009. *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP; StataCorp. 2011. *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP) were used to generate results, screenshots and output in this version of the text.

ACKNOWLEDGEMENTS FROM THE PREVIOUS EDITION

Many people have helped shape this text and influence our thinking about or presentation of the topics therein. The book grew out of a redesign of the core statistics course in the MBA curriculum at the Kellogg School of Management, Northwestern University. Initially, Peter Klibanoff and Boaz Moselle undertook this redesign. Alvaro Sandroni joined soon after in helping implement it. Brett Saraniti joined us more recently to help turn the notes into a text, while Boaz had left for the government and private sectors by this time and was less involved in the final phases. We owe great thanks to the many colleagues and former colleagues who developed some of the early material from which the redesign grew and who provided us with pedagogical advice and encouragement early in our careers. Larry Jones especially stands out in this regard and we also greatly valued the help from Matt Jackson, Alejandro Manelli, Tzachi Gilboa and Bob Weber. Don Jacobs, former Dean of Kellogg, deserves credit for motivating the redesign and for insisting on Microsoft Excel-based tools. We want to thank those colleagues who have taught Kellogg courses along with us based on our notes, and who also helped improve both the content and form of the material. Peter Esö, Christoph Kuzmics, Karl Schmedders, Eilon Solon and Rakesh Vohra helped make our teaching more successful and more fun by being an active part of our “teaching team”. Karl and Eilon jointly developed with us some of the material Chapter Two is based on. Peter Esö was extremely helpful in developing some of the Chapter Nine material and in creating the current form of Chapters Eight and Nine. Bob Weber was generous enough to let us adopt and expand his KStat collection of macros for Microsoft Excel. Two of Peter Klibanoff's former doctoral students helped in producing the notes on which the book is based: Emre Ozdenoren helped integrate the use of KStat and Excel throughout, and Sayan Chakraborty helped with the material in Chapter One. Piotrek Kuszewski helped with the Test Bank materials. We thank the numerous doctoral students who have been our teaching assistants for their help and contributions. Finally, this text literally would not exist without the joy and challenge of teaching our many MBA students at Kellogg over the years. It was they who motivated the whole orientation of the text and whose feedback convinced us that not only were we on the right track, but that students elsewhere could benefit from, learn from and enjoy a text based on our approach to managerial statistics.

We would like to acknowledge the work of our reviewers who provided comments and suggestions of ways to continue to improve our text. Thanks to: Suryapratim Banerjee, Vanderbilt University; Bruce Barrett, University of Alabama; Bernard Dickman, Hofstra University; Walter J. Mayer, University of Mississippi; David R. McKenna, Boston College; Muhammad A. Obeidat, Southern Polytechnic State University; Alan Olinsky, Bryant College; Victor R. Prybutok, University of North Texas; Helen M. Roberts, Montclair State University; Subarna K. Samanta, The College of New Jersey; and Edward D. White III, Air Force Institute of Technology.

We would like to thank the team at Thomson/South-Western responsible for our text. We thank our senior acquisitions editor Charles McCormick, Jr., our senior developmental editor Alice Denny, our senior marketing manager Larry Qualls and George Werthman for believing in the project enough to take us on and for their patience as it developed. We thank our production manager at Thomson, Kelly Hoard, and our production manager at GEX, Inc., Sarah McKay, for their great work, communication skills and patience in seeing this through to the final product.

Finally, all of us thank our families for bearing with the extra time and energy that this project took from us away from them.

Peter Klibanoff
Boaz Moselle
Alvaro Sandroni
Brett Saraniti

CHAPTER 1

DOUBLE E (EE): AN INTRODUCTION TO PROBABILITY DISTRIBUTIONS AND ESTIMATION

This chapter introduces us to the Double E (EE) chain of consumer electronics stores and their struggle to improve operations by using some basic statistical analysis. EE's main problem is dealing with pseudo customers who utilize its sales staff's time and expertise and then buy the products online or elsewhere. The case motivates the use of data to diagnose and help construct solutions to the company's issues. The topics introduced include means, standard deviations, variances, proportions, normal and t-distributions, sampling, the sampling distribution of the sample mean, confidence intervals for means and proportions, and some associated Excel and Stata functions.

The techniques developed in this case will establish a foundation for more sophisticated analysis discussed later.

1.1 EE: Uncertainty and Probability

EE is a chain of stores selling consumer electronics in the United States. Over the last decade, it has expanded to more than 4,000 stores spread across the country, thereby becoming one of the largest retailers of consumer electronics in the country. However, of late, EE's profits have been declining. The primary reasons for this are suspected to be falling quality of service and growing competition. EE has decided to deal with the problems aggressively and wants to come up with fast and effective solutions. In this chapter, we will see how probability and basic statistics will be useful to EE in a number of areas. Furthermore, many topics introduced in this chapter will be used and referred to repeatedly throughout the remainder of the book.

PROBABILITY DISTRIBUTION

Much of what EE deals with, or encounters in the course of its operations, involves fluctuating quantities. For example, it experiences variations in its weekly sales, the number of items turned in for repair each week, the number of items a customer buys during one visit, the length of time a salesperson spends with a single customer, the end-of-quarter profits, etc. One convenient way of summarizing the fluctuations is to use a **probability distribution**. A probability distribution makes possible the calculation of the chance that a variable lies in a given range. For example, a probability distribution for weekly sales allows us to calculate the chance that the weekly sales will be in a given range (e.g., weekly sales between $10,000 and $50,000).

A **continuous probability distribution** is one in which the variable can assume any value within a range. This means that if a variable can take the values, *a* and *b*, it can assume any value between *a* and *b*. Graphically, a continuous probability distribution can be represented by a curve (see Figure 1.1).

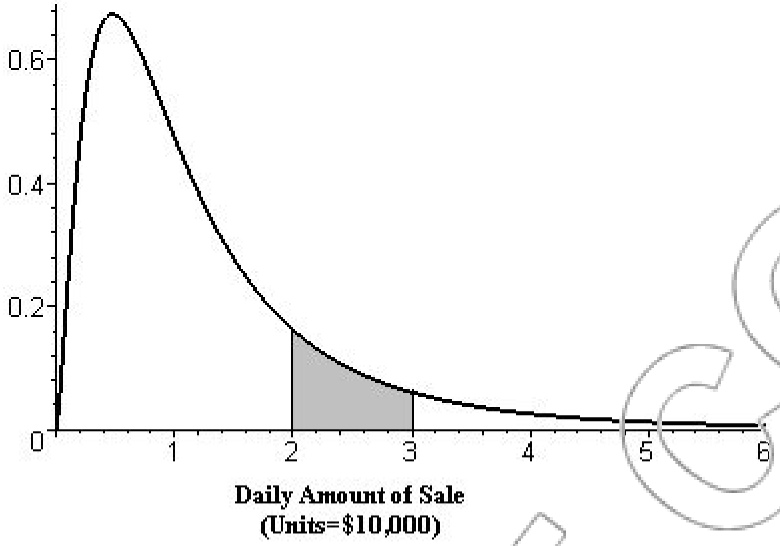


Figure 1.1: Graph of probability distribution describing the daily sales (in dollars) at an EE store.

One variable that would typically be described by a continuous distribution is the dollar amount of sales in a day at an EE outlet. The area under the curve within a given range gives the probability of sales falling in that range. For example, in Figure 1.1, the probability that the dollar amount of sales on a given day is between $20,000 and $30,000 is equal to the area of the shaded region. Since something always has to happen, the total area under the curve for any probability distribution is equal to one.

A **discrete probability distribution** is one in which the variable only takes on a certain countable number of values. For instance, the number of customers who buy flat panel televisions tomorrow in a given store follows a discrete probability distribution with possible values of {0, 1, 2, 3, 4, 5 or more}. The tools developed in this text will rely on continuous distributions. In fact, though the dollar amount of sales is discrete (we cannot divide pennies any further), we have assumed for simplicity that it is described by a continuous distribution. We will frequently use this standard trick to our advantage. For purposes of convenience, it often pays to approximate discrete distributions by continuous distributions.

1.2 The Mean

We will now introduce three of the most widely used attributes of a probability distribution, namely, the mean, the variance, and the standard deviation. We start with the mean. The mean of a distribution measures the average (or expected) value of that distribution. The mean is often our best single prediction for a variable’s value. Consider the sales manager of an EE store. He knows that the weekly sales of desktop personal computers (PCs) can be described by a probability distribution. The mean sales provide him with a single number around which the actual weekly sales will vary. It is usually denoted by the Greek letter μ (“mu”).

What the mean does for a probability distribution is similar to what the average does for a group of numbers. The mean is also calculated much like the average of a group of numbers. Before learning how this is done, let us review how one computes the average of a group of numbers. Suppose the sales manager at an EE store observes the sales of desktop PCs for 5 weeks in succession. Let us take them to be 19, 25, 20, 25 and 27. To get the average sales of desktops per week during this period, she needs to sum up these numbers and divide by five. The average weekly number sold is equal to the following:

$$\text{Average sales} = (19+25+20+25+27)/5 = 116/5 = 23.2$$

This means that, on average, 23.2 desktop PCs were sold each week at the store during this time period.

1.3 The Variance and Standard Deviation

Knowing the mean is not always enough to compare two probability distributions. If a particular distribution has a higher mean than a second one, all the values of the first one are not necessarily higher than the second one. To illustrate this, consider the dollar amounts of sales in two of EE’s stores. Suppose they can be represented by the probability distributions shown in Figure 1.2. The means of the distributions are labeled μ_1 and μ_2. Though the mean of distribution 2 (μ_2) is higher than that of distribution 1 (μ_1), a value drawn from distribution 2 may be lower than one drawn from distribution 1. In fact, because distribution 2 is so spread out there is a greater probability of obtaining very low values than there is with distribution 1. This shows that having a measure of the spread around the mean is useful in addition to knowing the mean itself.

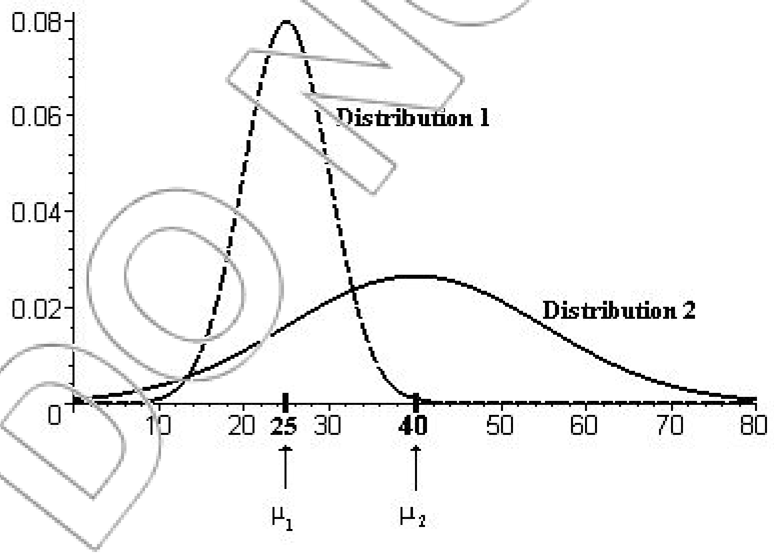


Figure 1.2: μ_1 = mean of distribution 1; μ_2 = mean of distribution 2.

The variance is the most frequently used measure of variation or spread of a distribution around the mean. The higher the variance of a distribution, the more likely it is for the variable to assume values far from the mean. Mathematically, the variance is the average squared deviation from the mean (i.e., for each possible value, subtract the mean, square the resulting number, and calculate the mean of these numbers using the probability distribution) and is usually denoted by σ^2 ("sigma squared"). Basically, it measures on average how "far" the actual sales are from their average.

Why is a number like the variance useful? Consider, for example, the sales manager at an EE store who is in charge of ordering inventories. To order inventories in the right quantities, she needs to account for the variability in weekly demand for different items sold at the store. She knows that probability distributions can be used to understand the demand fluctuations. To set the right inventory levels, knowing the mean is generally insufficient. She also needs to know how spread out the distribution for demand is about its mean. In other words, she needs to measure the variability in demand for that particular item. The variance and standard deviation of the probability distribution can do this for her.

THE MEAN AND VARIANCE OF FINANCIAL SECURITIES

One important application of mean and variance lies in finance. The return on any financial security fluctuates and can be described by a probability distribution. A security with a higher mean return than a second one provides higher returns on average. Obviously, any investor would prefer a higher mean return all else equal. However, this is not the only factor that influences the investment decisions of most investors. Investors' behavior suggests that they like high returns but dislike huge fluctuations or variations in the returns. Huge fluctuations suggest significant possibilities of very high or very low returns. This makes the security risky or volatile. The variance of the probability distribution used to describe the returns on a security is one measure of the risk associated with the security. The higher the variance becomes, the more risky the security is. A risk-sensitive individual takes into account both the means and the variances of securities while making investment decisions.[1]

STANDARD DEVIATION

One drawback of the variance is that, as a number, it can be hard to interpret. This is because it is measured in the square of the original variable's units. For example, the distribution of weekly sales measured in dollars will have a variance measured in dollars squared. Interpreting dollars squared is difficult. For this reason, it is common to use the square root of the variance, called the standard deviation, instead of the variance itself. The standard deviation is a measure of spread that is always in the same units as the original variable. Since the standard deviation is the square root of the variance, it is usually denoted by σ ("sigma").

1.4 Proportions

Working with variables with only two possible outcomes can sometimes be helpful. Consider the customers who come to an EE store. Some of them buy at least one product and some leave without buying any. The variable "customer buys at least one item" has two possible outcomes:

[1] In Chapter 4, we will revisit the connection between variance and risk in the context of capital budgeting and the CAPM model.

YES or NO. To use this variable numerically, we can say the variable takes the value 1 if the customer buys at least one product and 0 if he or she does not buy any. If we use 1 and 0 in this way, then the average, or mean, of the variable is the **proportion** of customers who buy at least one item. A specific illustration is the following. We look at any five EE customers. We observe if each customer buys an item or not on his or her visit to the store and assign the value 1 and 0 accordingly. For example, (see Figure 1.3), customers 1, 4, and 5 do not buy any items, and customers 2 and 3 do.

Customer Identity	Value of variable showing if an item is bought
Customer 1	0
Customer 2	1
Customer 3	1
Customer 4	0
Customer 5	0

Figure 1.3: This table shows if a customer bought an item.

Let us take the average of the values in the right-hand column. The average is 0.4. Notice that 0.4 (or 40%) is the proportion of these five customers who bought at least one item. Hence, the average of this variable gives the proportion of the five customers who bought at least one item.

When dealing with a variable with two outcomes coded as 0 and 1, instead of talking about the mean, we will sometimes use the proportion, which we denote by p. The proportion is always between 0 and 1. When p is the mean of the distribution of such a variable, p(1-p) and

$\sqrt{p(1-p)}$ will be its variance and standard deviation, respectively. So, for a variable with only two outcomes, 0 and 1, knowing the proportion tells you the mean, the variance, and the standard deviation.

1.5 The Normal Distribution

The normal distribution is one of the most common distributions in statistics. There is a whole family of normal distributions, one for each pair of means and standard deviations. Each normal distribution can be uniquely characterized by those two parameters.

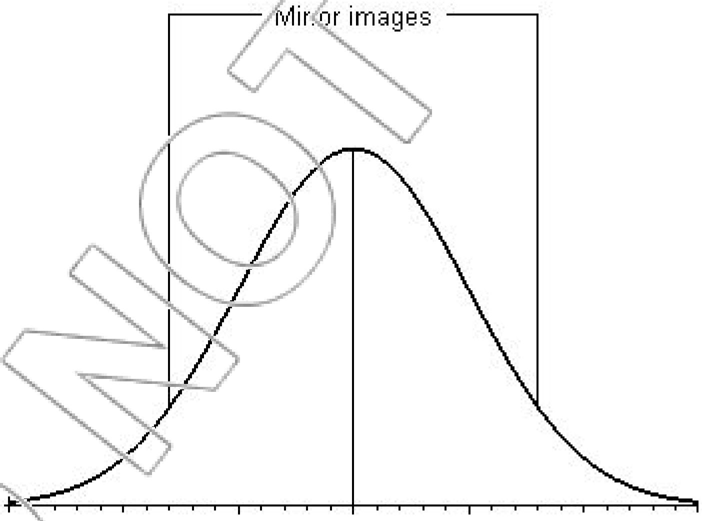


Figure 1.4: Normal distribution is symmetric and bell-shaped.

Characteristic features of a normal distribution are its bell shape and symmetry (see Figure 1.4). Symmetry of the distribution implies that if a vertical line is drawn along the middle of the distribution, the left and right halves will be mirror images of one another. The tails of a normal distribution approach, but never touch, the X-axis. Though they are possible, values far above or below the mean occur with small probability. Normal distributions with large standard deviations have shorter peaks and fatter tails than most. Distributions with smaller standard deviations have taller peaks with thin tails.

EXCEL FUNCTIONS

NORMDIST: The NORMDIST function in Excel calculates the area within a given range under a particular normal distribution. Directly, this function gives us the area to the left of a given value, but because the total area under the curve is equal to one, we can use the function to determine any area or probability for a normal distribution.

For example, suppose we want to find the area to the right of 36.5 under the normal distribution with mean of 28 and standard deviation of 7 (the area A as shown in the Figure 1.5).

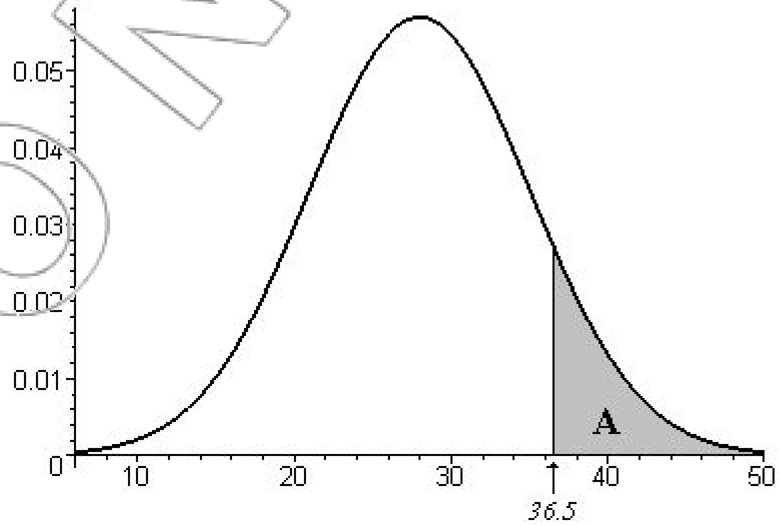
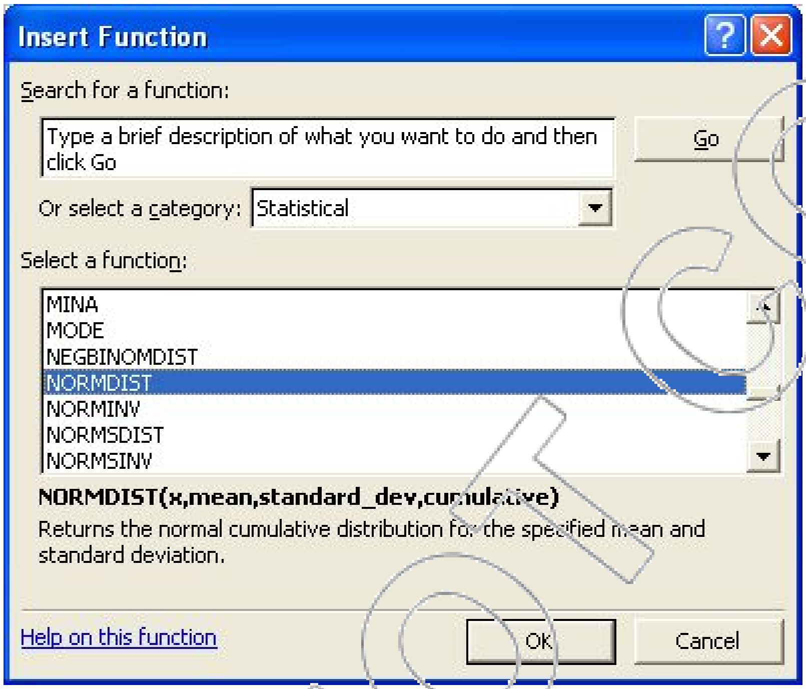


Figure 1.5: Normal distribution with mean of 28 and standard deviation of 7.

To calculate this area, open a worksheet in Excel. Select **INSERT>FUNCTION** from the menu and choose **Statistical** from the **Function Category** window. Then choose **NORMDIST** from the **Function Name** window as shown below.



Search for a function:

Type a brief description of what you want to do and then click Go

Or select a category: Statistical

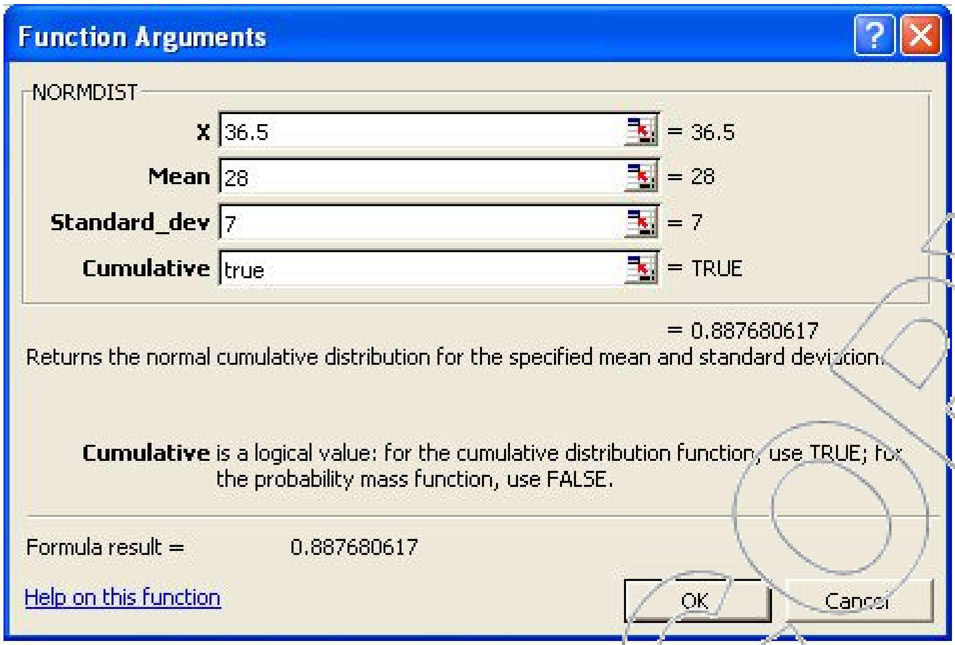
Select a function:

MINA
MODE
NEGBINOMDIST
NORMDIST
NORMINV
NORMSDIST
NORMSINV

NORMDIST(x,mean,standard_dev,cumulative)

Returns the normal cumulative distribution for the specified mean and standard deviation.

When you click **OK**, you will see a dialog box like this, and you can fill in the boxes with the appropriate values.



NORMDIST

X 36.5 = 36.5

Mean 28 = 28

Standard_dev 7 = 7

Cumulative true = TRUE

= 0.887680617

Returns the normal cumulative distribution for the specified mean and standard deviation.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability mass function, use FALSE.

Formula result = 0.887680617

Click **OK** to get the area to the left of 36.5. This area turns out to be 0.888 (rounding off to three decimal places). Since we wish to find the area to the right of 36.5, we have to calculate 1 minus 0.888. This means that area A, which equals the probability of being at least 36.5, is 1-0.888 = 0.112.

How can we find the area between two values under a normal distribution using the NORMDIST function? Suppose we want to find the area lying between 36.5 and 38 under the normal distribution with mean of 28 and standard deviation of 7. This is the region marked B in Figure 1.6. Observe that the area of B is equal to the area to the left of 38 minus the area to the left of 36.5. Therefore, you should find these two areas using Excel and subtract the smaller one from the larger. Earlier, we found that the area to the left of 36.5 is 0.888. (Typing **=NORMDIST(36.5, 28, 7, TRUE)** into a blank cell will also give you the same result.) Proceeding similarly, the area to the left of 38 is 0.923. Therefore, the area between 36.5 and 38 is 0.923-0.888 = 0.035.

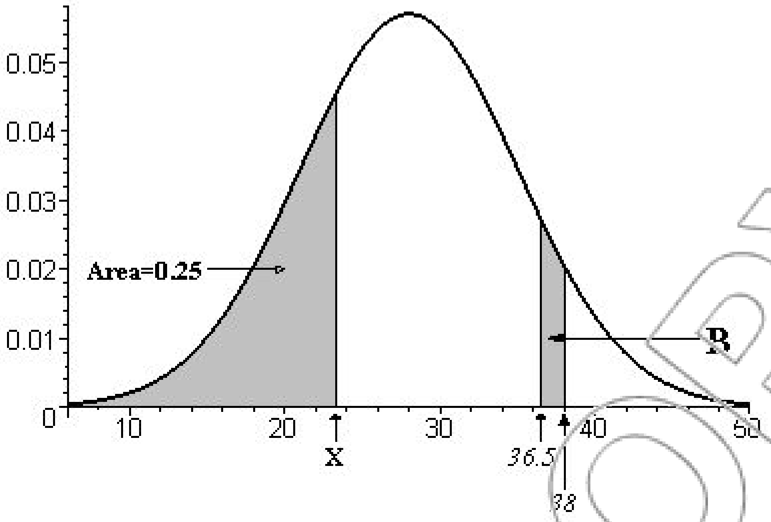
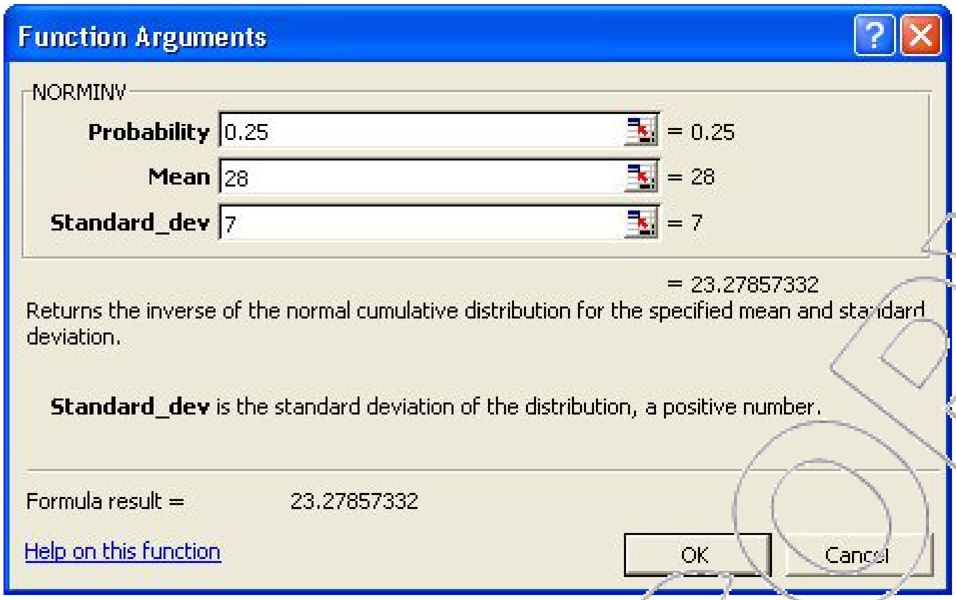


Figure 1.6: Normal distribution with mean of 28 and standard deviation of 7.

NORMINV: Consider once again the normal distribution with mean of 28 and standard deviation of 7. Suppose we want to find the value for which the probability of falling below that value is 0.25. In Figure 1.6, this is the point denoted by X. To find this value, select **INSERT>FUNCTION** from the menu and choose **Statistical** from the **Function Category** window. Then choose **NORMINV** from the **Function Name** window. When you click **OK**, you will see a dialog box like this (once we have filled in some of the boxes):



Function Arguments

NORMINV

Probability 0.25 = 0.25

Mean 28 = 28

Standard_dev 7 = 7

= 23.27857332

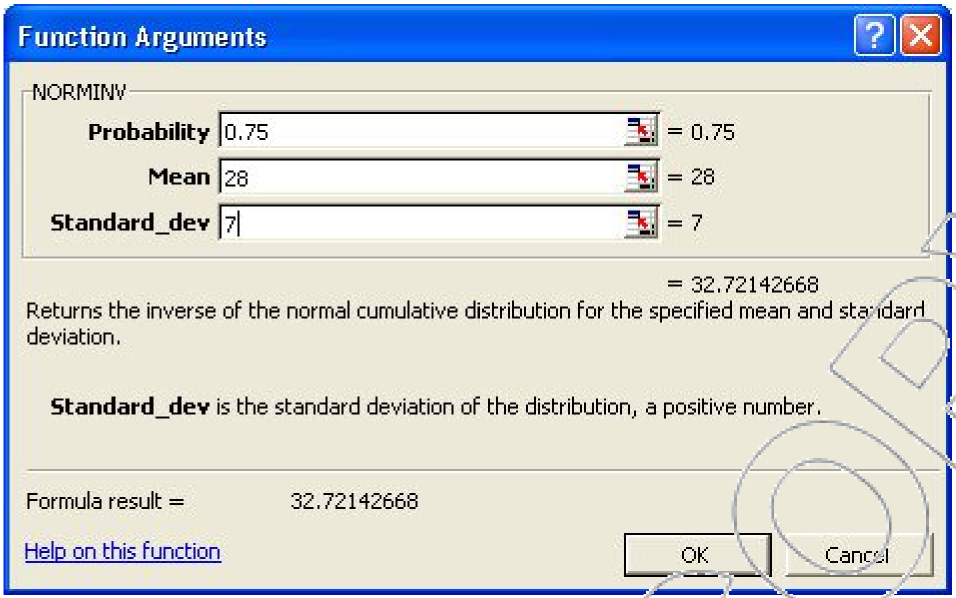
Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

Standard_dev is the standard deviation of the distribution, a positive number.

Formula result = 23.27857332

In the dialog box, type in the probability that you want to the left of the value (0.25 in this example). Type the mean and standard deviation of the normal distribution corresponding to **Mean** and **Standard_dev**, respectively. When you click **OK**, Excel returns the value of X as 23.279. In other words, the probability of obtaining a value below 23.279 from a normal distribution with mean of 28 and standard deviation of 7 is 0.25.

To calculate the value having a given probability to the right, you will need to input 1 minus that probability into NORMINV. For example, if you enter 0.75 as the probability, you find that the probability of obtaining a value above 32.721 from a normal distribution with mean of 28 and standard deviation of 7 is 0.25. The NORMINV function tells you what value will give you a certain probability to its left. At 32.721, we find 75% of the area to the left leaving 25% of the area under the curve to the right.



Function Arguments

NORMINV

Probability 0.75 = 0.75

Mean 28 = 28

Standard_dev 7 = 7

= 32.72142668

Returns the inverse of the normal cumulative distribution for the specified mean and standard deviation.

Standard_dev is the standard deviation of the distribution, a positive number.

Formula result = 32.72142668

Notice how both of these values we calculated with NORMINV are the same distance from the mean of 28. That is, |32.721-28| = 4.721 and |23.279-28| = 4.721. The symmetry of the normal distribution makes the distance from the mean (needed to get 25% of the area under the tail) the same in either direction.

STATA FUNCTIONS

You can find the area to the left of a particular value under a normal distribution and the value for which the area to the left is a given probability under a normal distribution by using the **normal(z)** and **invnormal(p)** commands in Stata, respectively. However, these two commands assume the normal distribution with mean of 0 and standard deviation of 1 (called the standard normal distribution). For this reason, we delay explaining these commands in detail until after discussing the standard normal distribution in the next section.

THE STANDARD NORMAL

The normal distribution with mean of 0 and standard deviation of 1 is called the standard normal or the z-distribution. Any normal distribution can be converted into the standard normal. The method of transforming a normal distribution into the standard normal is referred to as *standardization.* If a variable, X, has a normal distribution with mean of μ, and standard deviation of σ, then the variable $z = (X - \mu)/\sigma$ has a standard normal distribution. The new variable, z, measures the number of standard deviations X is away from the mean. For example, consider the weekly sales of microwaves at an EE store. Suppose that it is described by a normal distribution with mean of 25 and standard deviation of 5. If X denotes the variable *weekly sales of microwaves*, then the variable, $z = (X-25)/5$, will have the standard normal distribution.

Standardizing a normal variable is useful since it converts distances from the mean into units of standard deviations. This is important and helpful in drawing conclusions insensitive to the original units the variable was measured in. For example, stores A and B have weekly inventories of 30 and 20 microwaves, respectively. The weekly demand for microwaves in store A is normally distributed with mean of 25 and standard deviation of 5 (see Figure 1.7). For store B, the weekly demand is normally distributed but with mean of 16 and standard deviation of 3.5 (see Figure 1.8). Given this information, management wants to know which store has a higher probability of a stock out, i.e., running out of microwaves.

One way of answering this question is to do the following: To find the probability of a stock out in Store A, we look at the normal distribution with mean of 25 and standard deviation of 5 and find the area to the right of 30. Similarly, in Store B, we find the area to the right of 20 under the normal distribution with mean of 16 and standard deviation of 3.5. We can compare these two probabilities and see which store has a bigger chance of a stock out.

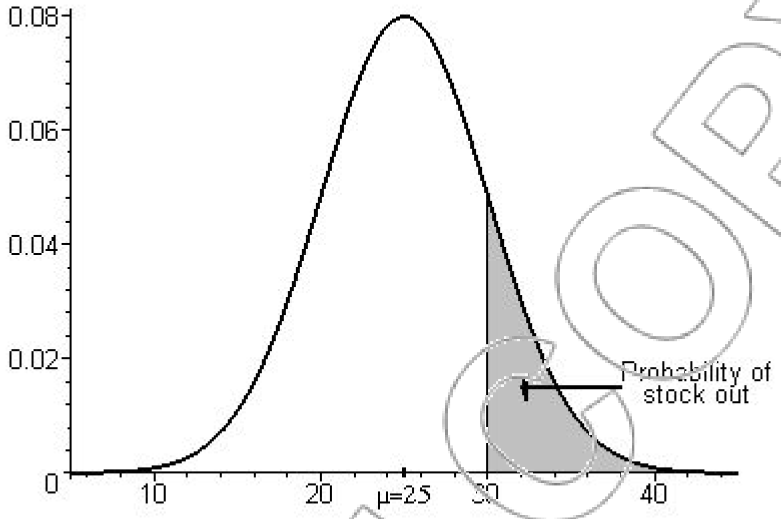


Figure 1.7: Shaded area represents the probability of a stock out in store A.

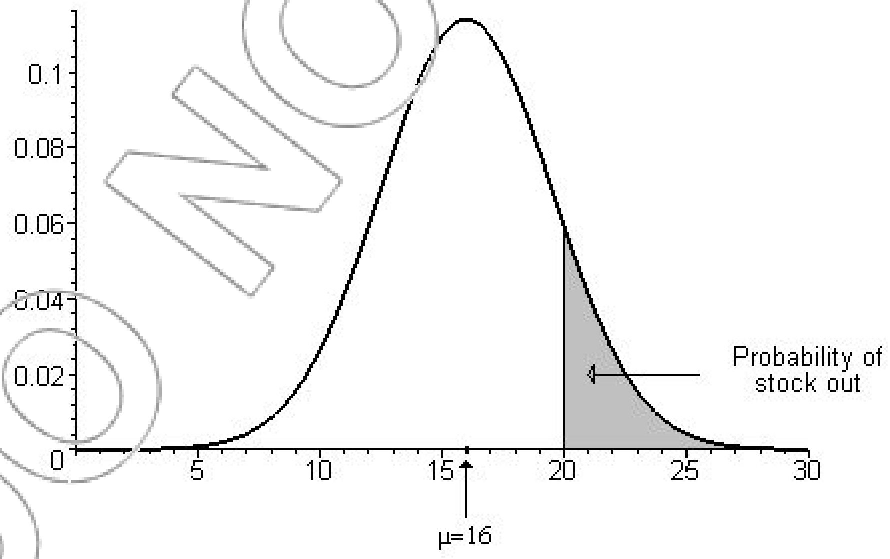


Figure 1.8: Shaded area represents the probability of a stock out in store B.

A simpler and more intuitive way of answering the above question would be to standardize the two distributions and compare them directly. This will give us the number of standard deviations 30 and 20 are away from their respective means. In store A, an inventory level of 30 is z_1 = (30-25)/5 = 1.00 standard deviation above the mean. For store B, the inventory level of 20 is z_2 = (20-16)/3.5 = 1.14 standard deviations above the mean (see Figure 1.9). The probability that a store suffers a stock out increases the fewer standard deviations its inventory level is above the mean. Since 1.00 is less than 1.14, the probability of a stock out in store A will be higher than that in store B. Standardization allows us to answer our question without finding the actual probabilities of stock outs in each store.

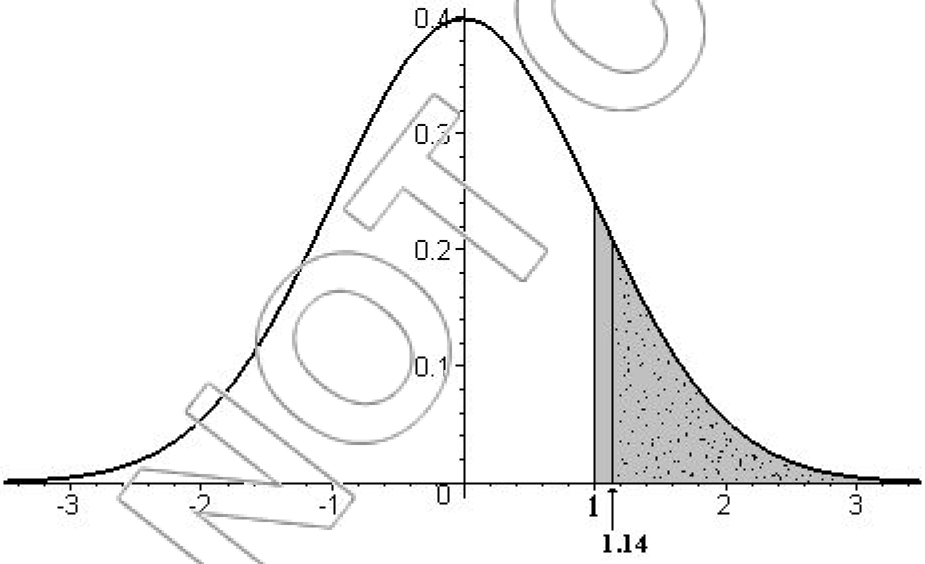


Figure 1.9: The standard normal distribution. The shaded area represents the probability of a stock out in store A. The dotted area represents the probability of a stock out in store B.

EXCEL FUNCTIONS

Excel has two functions that are useful when working with the standard normal. These are **NORMSDIST** and **NORMSINV**. As the names suggest, these functions are similar to the

NORMDIST and NORMINV functions we encountered earlier. However, unlike NORMDIST and NORMINV, the NORMSDIST and NORMSINV functions assume the distribution to be the standard normal.

STATA FUNCTIONS

normal(z): The normal(z) function in Stata calculates the area to the left of a given value z under a standard normal distribution. Therefore, to calculate the area to the left of a given value X that has a normal distribution with mean μ and standard deviation σ, you will need to first standardize the normal variable by using the equation z = (X-μ)/σ.

Consider again an example where we want to find the area to the right of 36.5 under the normal distribution with mean of 28 and standard deviation of 7. To calculate this area, open Stata. Type **display normal((36.5-28)/7)** in the Command box. Press **Enter**, and Stata will return the following:[2]

```
. display normal((36.5-28)/7)
0.88768068
```

Since this number is the area to the left of 36.5, to find the area to the right of 36.5, we have to calculate 1 minus this number. Using Stata to do this gives:

```
.display 1-normal((36.5-28)/7)
0.11231932.
```

[2] Note that in the actual Stata output, zero is omitted before the decimal. We have added a zero here to distinguish the decimal in the output from the period in front of the actual command.

To find the area between two values, say, 36.5 and 38, under the normal distribution with mean of 28 and standard deviation of 7, type **display normal((38-28)/7)-normal((36.5-28)/7)**. Press **Enter**, and Stata will calculate the area to be 0.03575559.

<u>invnormal(p):</u> The invnormal(p) command in Stata calculates the value for which the probability of falling below that value is p in the standard normal distribution. Consider once again the normal distribution with mean of 28 and standard deviation of 7. Suppose we want to find the value for which the probability of falling below that value is 0.25. In Stata, type **display invnormal(0.25)** in the Command box and press **Enter** to get:

```
. display invnormal(0.25)
-0.67448975
```

This tells us the area below -0.67449 in the standard normal distribution is 0.25. To convert this into a value in the normal distribution with mean 28 and standard deviation 7 we need to multiply by the standard deviation and then add the mean. Since -0.67449 = (X-28)/7, solving for X yields X = -0.67449*7+28 = 23.279. We could have done this directly in Stata by using the command **display 7*invnormal(0.25) + 28**.

1.6 The t-Distribution

The t-distributions are a common family of distributions in statistics. In fact, we will use them far more often than the normal distributions. The curve of a t-distribution is similar to a standard normal distribution. Like the standard normal, it is symmetric, bell-shaped, and has a mean of 0; however, all t-distributions have more area in the tails (i.e., fatter tails) than the standard normal.

t-distributions are characterized by a positive number called *degrees of freedom*. A t-distribution with a few degrees of freedom has very fat tails, and one with many degrees of freedom looks much like a standard normal. This is evident in Figure 1.10, where, as the degrees of freedom of a t-distribution increases (from 10 to 25 to 100), its shape resembles the standard normal.

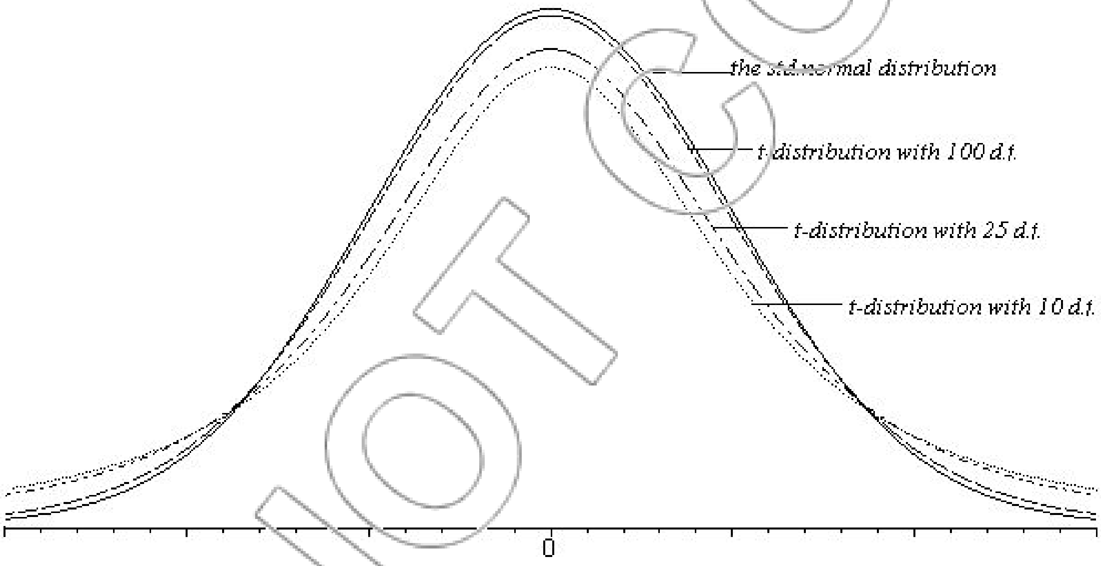


Figure 1.10: t-distributions converging to the standard normal as the degrees of freedom increases.

(The determination of the appropriate of degrees of freedom will be discussed further later on when we use t-distributions in connection with estimation.)

EXCEL FUNCTIONS

TDIST: The TDIST function gives the area under a t-distribution within a given range. Suppose we want to calculate the area to the right of 1 under a t-distribution with 20 degrees of freedom. This is the area marked A in Figure 1.11.

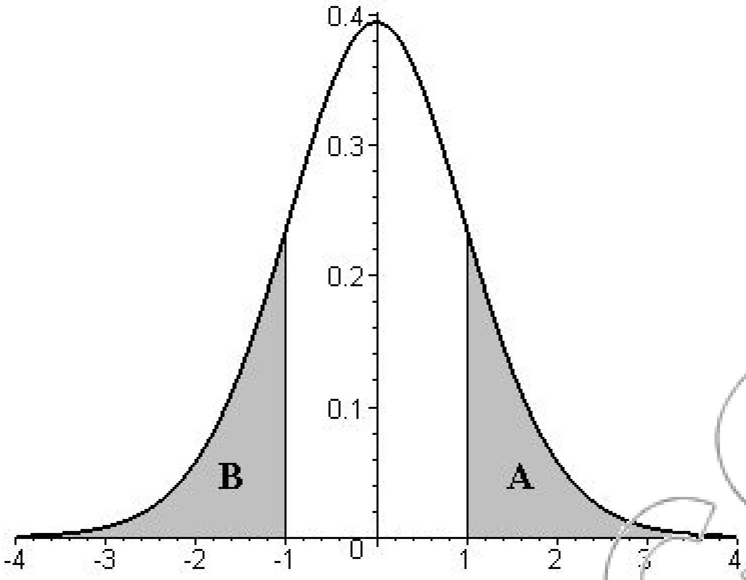
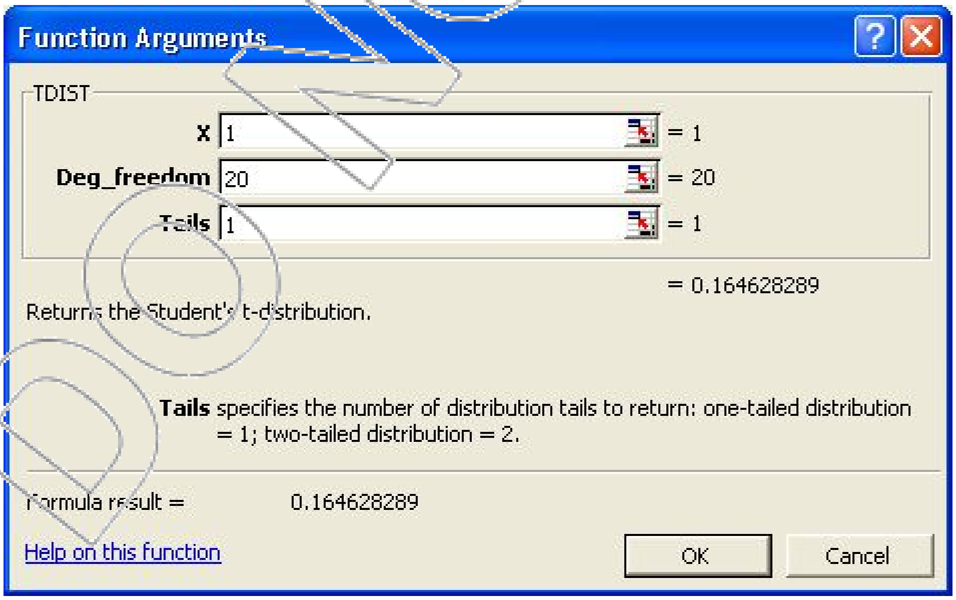


Figure 1.11: The t-distribution with 20 degrees of freedom. What are the areas of regions A and B?

In Excel click **INSERT>FUNCTION** and choose **Statistical** from the **Function Category** window. Then choose **TDIST** from the **Function Name** window. When you click **OK**, you will see a dialog box like this (once we have filled in some of the boxes):



X 1 = 1

Deg_freedom 20 = 20

Tails 1 = 1

= 0.164628289

Returns the Student's t-distribution.

Tails specifies the number of distribution tails to return: one-tailed distribution = 1; two-tailed distribution = 2.

Formula result = 0.164628289

In the dialog box, we choose the number, which is 1 in this case, to the right of which we want to find the area. Next, we must plug in the degrees of freedom of the t-distribution (in this case, 20). Since we want to find the area in one of the tails of the t-distribution, we type in 1 corresponding to **Tails**. Clicking **OK** now gives the area of region A to be about 0.165.

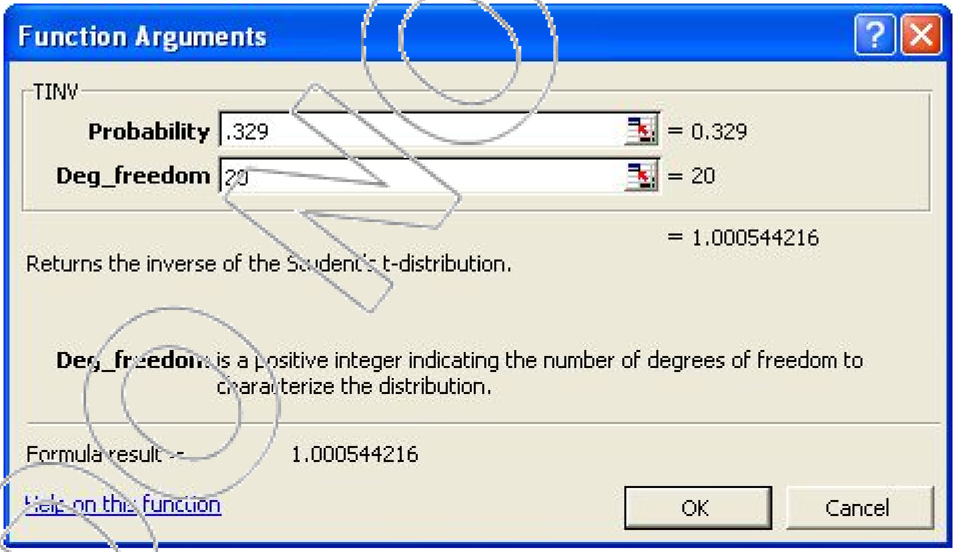
Suppose we want to find the area to the left of -1 (B in Figure 1.11). To do this, we have to make use of the symmetry of t-distributions since Excel does not accept a negative number as the first entry in the dialog box for TDIST. Symmetry ensures that for a variable Y with a t-distribution, Prob (Y<-1) = Prob (Y>1). In other words, the area to the right of 1 is the same as the area to the left of -1, i.e., the area of A is equal to area of B. Once we have realized this, we can determine the area of B by finding the area of A, Hence, the area of B = area of A = 0.165.

We might also be interested in knowing the area to the right of -1 under a t-distribution with 20 degrees of freedom. Since we cannot enter a negative number as the first entry of a TDIST dialog box, we cannot calculate this area directly. However, we can see from the symmetry in Figure 1.11 that Prob(Y>-1) = Prob(Y<1) = 1 - Prob(Y>1).

In English, that means the area to the right of -1 is equal to 1 minus the area to the right of 1. We know how to calculate the area to the right of 1 under a t-distribution with 20 degrees of freedom. In fact, we did this earlier. It is equal to the area of A in Figure 1.11, which we calculated to be 0.165. Therefore, the area to the right of -1 under a t-distribution with 20 degrees of freedom, equals 1 - 0.165 = 0.835.

Suppose we need to find the total area to the right of 1 and to the left of -1 for the t-distribution with 20 degrees of freedom. This is equal to the sum of areas A and B. You can do this by finding the area to the right of 1 and multiplying by 2. The required area becomes (2)(0.165) = 0.33. A more automatic way of doing this is to utilize the option of 2-Tails in the TDIST function. In the TDIST dialog box, type in **X** equal to 1, **Deg_freedom** equal to 20, and **Tails** equal to 2. Clicking **OK** gives the sum of the areas A and B, which is 0.329. The difference between 0.329 and 0.33 is solely due to round-off error.

TINV: Like the NORMINV and NORMSINV functions, the TINV function returns a number for a given probability/area. However, the TINV function operates in a different manner. Given an area under a t-distribution with a specified number of degrees of freedom, the TINV command returns a number to the right of which lies half the area entered. For example, referring to Figure 1.11, an area of about (0.5)(0.329) = 0.165 lies to the right of 1 under a t-distribution with 20 degrees of freedom. To see how TINV returns the desired number, click **INSERT>FUNCTION, choose Statistical** and choose **TINV** from the **Function Category** and click **OK**. The following Dialog box appears (after filling in the values):



Function Arguments

TINV

Probability .329 = 0.329

Deg_freedom 20 = 20

= 1.000544216

Returns the inverse of the Student's t-distribution.

Deg_freedom is a positive integer indicating the number of degrees of freedom to characterize the distribution.

Formula result = 1.000544216

In the dialog box, you will type 0.329 (the sum of areas A and B) for **Probability** and 20 as the **Deg_freedom**. When you click **OK**, Excel returns the value 1.0005. (Since we rounded 0.329 a little bit, the results here are off a little bit as well.) The function, therefore, returns a number to the right of which lies half the given area. The remaining half of the area lies to the left of the negative of the same number (in this case, -1).

Suppose we want to find the number to the right of which is an area of 0.0225 under a t-distribution with 14 degrees of freedom. To find the number using Excel, open the TINV dialog box and type in 0.045 [= (2)(0.0225)] as **Probability** and 14 as **Deg_freedom**. Excel returns the value 2.201.

STATA FUNCTIONS

ttail(n,t): The ttail(n,t) function in Stata gives the area to the right of t under a t-distribution with n degrees of freedom. Suppose that we want to calculate the area to the right of 1 under a t-distribution with 20 degrees of freedom. Typing **display ttail(20,1)** in the Command box and pressing **Enter** will generate the following:

```
. display ttail(20,1)
0.16462829
```

Note that the number t entered in the ttail(n,t) command may be positive or negative. For example, to calculate the area to the right of -1 under a t-distribution with 20 degrees of freedom, we simply type **display ttail(20,-1)** and get 0.83537171.

Stata does not automatically calculate the two-tailed area corresponding to a given value under a t-distribution. If, for example, we want to find the total area to the right of 1 and to the left of -1 for the t-distribution with 20 degrees of freedom, typing the command **display 2*ttail(20,1)** generates the answer (approximately 0.329).

invttail(n,p): The invttail(n,p) command in Stata calculates the value in a t-distribution with n degrees of freedom for which the probability of falling to the right of that value is p. Consider the example related to Figure 1.11, where we calculated the area to the right of 1 under a t-distribution with 20 degrees of freedom to be approximately 0.165. To see if 1 is indeed the number having area of 0.165 to its right in that t-distribution, using Stata, type **display invttail(20, 0.165)**, press **Enter**, and get:

```
. display invttail(20,0.165)
0.99842649
```

The result is roughly equal to 1. The discrepancy is due to our rounding of the 0.165. The usefulness of the invttail command will become clearer below when we study confidence intervals.

1.7 Estimating with Data

One of the reasons for EE's declining profits is the stiff challenge posed by its rivals. EE is facing increasingly tough competition from online retailers. Managers at EE suspect that a number of customers who come to an EE store get help from the salespeople in understanding and comparing different products but often stop short of buying the product. They would rather buy the chosen product from an online retailer. Online retailers, with lower operating expenses, overhead costs, and often a tax-advantage can afford to sell the product at a cheaper price than a brick-and-mortar retailer like EE. Such a phenomenon adds nothing to EE's revenues and reduces the quality of service provided to customers who buy from EE.

To cut down on the service provided to pseudo customers (customers who use EE to learn about a product but do not buy from EE) and increase the quality of service for its true customers, managers at EE have suggested several possible strategies. One of the suggested solutions is to set a refundable service charge for all customers seeking advice from a salesperson at EE. This service charge will be refunded in full if the customer goes on to buy the product from EE; otherwise, it will not be refunded. Before spending time debating the merits of various strategies such as these, EE must ascertain whether and to what extent such a problem exists. The manager might want to know the average time spent by a salesperson with pseudo customers per day, the average waiting time for a true customer (waiting time is defined by the length of time a true customer waits before being attended by a salesperson), and the proportion of pseudo customers. For instance, if pseudo customers do not take up much of the salespeople's time, then the problem of the sales force spending unproductive time with pseudo customers would not be so serious. Specifically, EE management, based on costs and industry benchmarks, has concluded that if less than 20% of a salesperson's day (approximately 1 hour and 36 minutes of an 8-hour day) is spent with pseudo customers, then the drain on service personnel by pseudo customers will not be considered a serious problem.

To estimate the average time spent with pseudo customers, the manager could chart the daily time spent by each salesperson with pseudo customers by going to (or contacting) each of the 4,000+ EE stores and subsequently find the average of those times. In practice, observing the service time spent by each salesperson with pseudo customers across all EE stores is costly. Even in situations where all the historical data could be collected, it is never possible to collect data on future service times. Thus, in all such situations, we will need to draw conclusions from a **sample** of the elements of interest rather than looking at the entire **population** of interest (here, time spent by salespersons with each past, present, and future pseudo customer).

Sample Size:

The sample size is the number of observations in the sample. This is denoted by n, i.e., n = 100 means there are 100 observations in the sample. In general, the larger the sample size, the more precise are the estimates based on that sample. When deciding on the size of the sample, one trades off the cost and time involved in collecting each observation against the value of more precise estimates.

ESTIMATING THE MEAN

The management team at EE would like to know the average time a salesperson spends attending to pseudo customers. However, all it has is the information in the sample. What is the best way to use the sample to estimate the population (or "true") mean? The best estimate of the true mean is the **sample mean**. The sample mean is calculated by adding all the values in the sample and dividing by the sample size.

It is important to distinguish between the population mean and the sample mean. Notationally, the population mean is denoted by μ, and the sample mean is denoted by $\bar{x}$ ("x-bar"). $\bar{x}$ is the estimator that we'll use to estimate μ.

COMPUTATION OF THE SAMPLE MEAN

Consider a sample of service times that the service manager has collected. It is stored in the file **service**. This file provides the observed service times spent with pseudo customers in a day by

100 salespersons. The size of the sample is 100. Service times have been measured in seconds and stored in the column named **servicetime**. To calculate the sample mean, we can use the **ktabstat** command in Stata. An easy way to invoke this command to calculate the sample mean and a number of other statistics for all of the variables in a dataset is through the **Univariate Statistics>Standard (ktabstat)** command on the **Core Statistics** custom menu. To do this, first load the **service.dta** file into Stata.[3] Now select **User>Core Statistics>Univariate Statistics>Standard (ktabstat)** from the drop-down menu (see Figure 1.12). You can also invoke the **ktabstat** command by typing **db ktabstat** in the Command box.

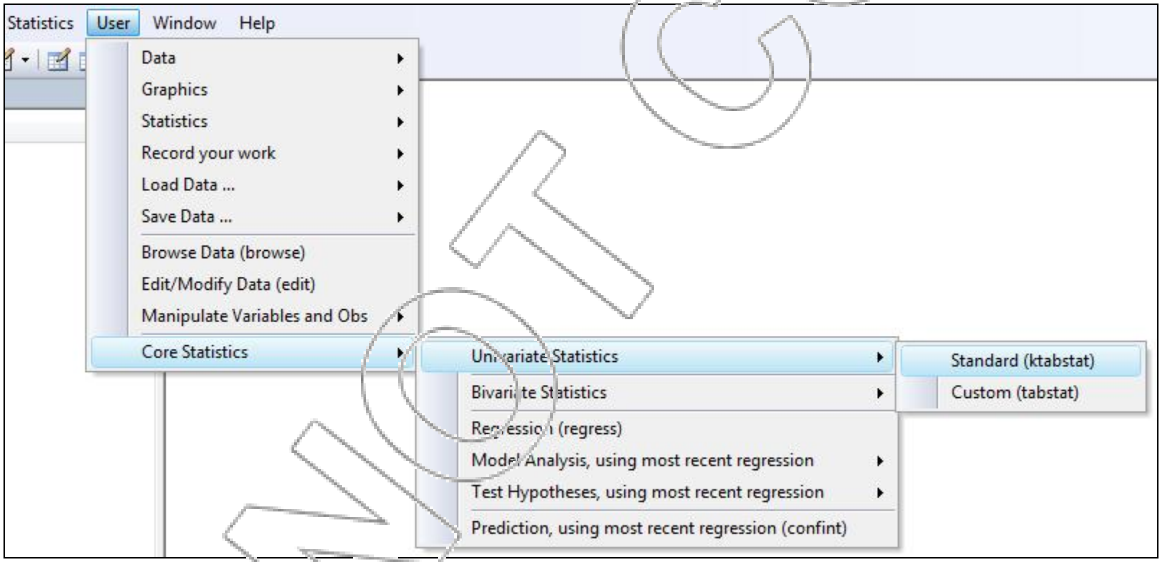
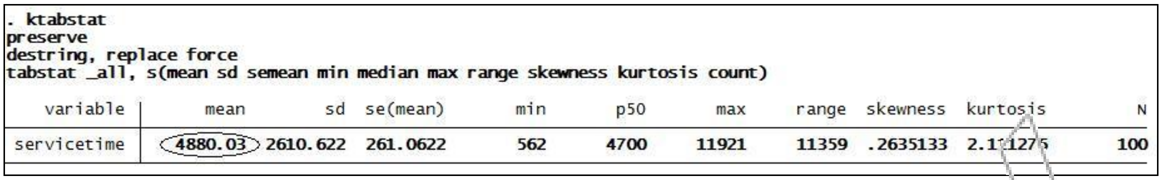


Figure 1.12 The Univariate Statistics command in the Core Statistics custom menu.

Click **OK** in the ensuing dialog box, and Stata will produce the output in Figure 1.13 that includes the sample mean, $\bar{x}$, as well as a number of other values to be explained later.[4] The sample mean is the number in the **mean** column, which is given as 4880.03 seconds.

[3] See the Appendix for instructions on loading, converting, and saving data files in Stata.

[4] As you can see from Figure 1.13, the analogous typed Stata command is **ktabstat**.



```
. ktabstat
preserve
destring, replace force
tabstat _all, s(mean sd semean min median max range skewness kurtosis count)

    variable |      mean        sd  se(mean)       min       p50       max     range  skewness  kurtosis         N
-------------+--------------------------------------------------------------------------------------------------
 servicetime |   4880.03  2610.622  261.0622       562      4700     11921     11359  .2635133  2.111276       100
----------------------------------------------------------------------------------------------------------------
```

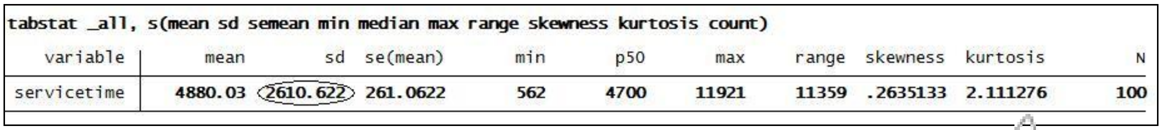
Figure 1.13: Univariate statistics for servicetime (mean).

How does this compare with the 1 hour 36 minutes threshold set by management? Since the threshold is 5760 seconds (equal to 1 hour 36 minutes), we see the sample mean is below it. We hope that this is because the sample mean reflects the actual mean, but we are unsure. Maybe we were lucky (or unlucky if it means we make a bad decision) with the sample we used. We must continue the analysis to quantify more precisely our confidence that the population mean is below management's threshold.

ESTIMATING THE STANDARD DEVIATION

The sample mean provides an estimate of the population mean. Is the time spent by most salespersons with pseudo customers similar to the mean? Are a few spending a long time while the others are spending a short time? To answer these questions, we must estimate the distribution's variance or the standard deviation. Since we'll mostly be working with the standard deviation later on, we'll focus on that now. The best estimate of the true standard deviation is the **sample standard deviation**. The sample standard deviation, s, is the estimator we use to estimate the population standard deviation, σ.

The **User>Core Statistics>Univariate Statistics>Standard (ktabstat)** command also calculates the sample standard deviation. The sample standard deviation is the number in the **sd** column (see Figure 1.14). For this data, s = 2610.622 seconds.



```
tabstat _all, s(mean sd semean min median max range skewness kurtosis count)
    variable |      mean        sd  se(mean)       min       p50       max     range  skewness  kurtosis         N
-------------+----------------------------------------------------------------------------------------------------
 servicetime |   4880.03  2610.622  261.0622       562      4700     11921     11359  .2635133  2.111276       100
------------------------------------------------------------------------------------------------------------------
```

Figure 1.14: Univariate Statistics for servicetime (standard deviation).

1.8 The Sampling Distribution

We have estimated the average time spent by an EE salesperson each day serving pseudo customers. To do this, we have used a sample of a 100 observations. Our estimate, $\bar{x}$, of the mean, μ, depends on the particular sample we have used. Naturally, the average time spent per day by a salesperson to serve pseudo customers calculated from a sample of 100 randomly observed times of EE salespersons will be different from the $\bar{x}$ calculated from a different sample of 100 randomly selected service times of EE salespersons. The value of the sample mean, $\bar{x}$, varies from sample to sample. The source of the variation in the value of the sample mean is the potential variation in the sample drawn from the population. In other words, since many samples could be drawn from a population, there are correspondingly many values of the sample mean $\bar{x}$. Thus, we can view the sample mean as a variable having a probability distribution. This distribution is called the **sampling distribution of the sample mean**.

In general, any estimator based on a sample will have a sampling distribution. There are sampling distributions for the sample variance, the sample standard deviation, as well as for the sample mean. Sampling distributions are important since they give us an idea about the accuracy of an estimator. The estimators that we commonly consider are all unbiased. An estimator is unbiased if the mean of the sampling distribution of the estimator is equal to what is being estimated. For example, the mean of the sampling distribution of $\bar{x}$ is μ, the population mean. Thus, $\bar{x}$ is an unbiased estimator of μ. Unbiased estimators are desirable because, on average, they are right. They are not consistently too high or too low.

A sampling distribution tightly concentrated around the mean tells us that the estimator is likely to be much more accurate (i.e., closer to the true value) than one that has a sampling distribution widely dispersed around the average. This is evident if one looks at Figure 1.15. Estimator 1 is more accurate than estimator 2 since estimator 1 has a higher probability of falling within any given distance from the true population value than estimator 2. This occurs because the standard deviation of the former is less than that of the latter. An unbiased estimator with a smaller standard deviation of its sampling distribution will be more accurate than one with a larger standard deviation.

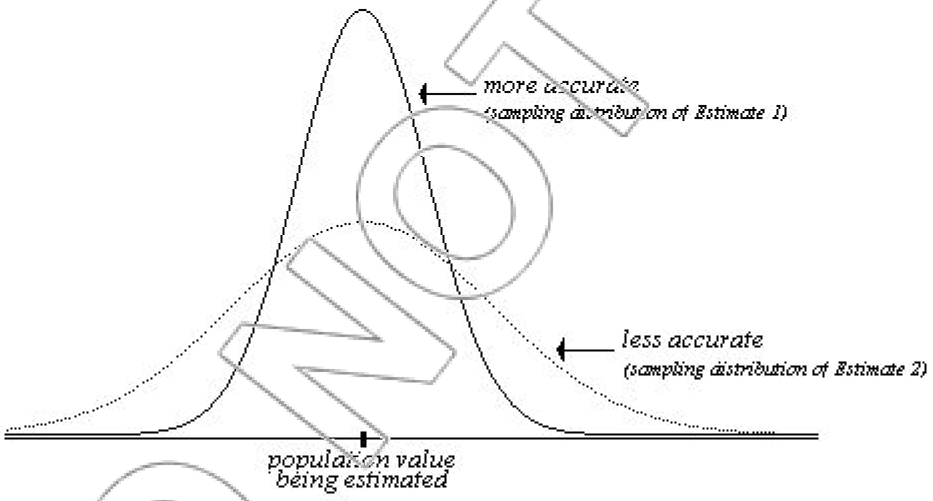


Figure 1.15. The sampling distributions of the two estimators show that estimator 1 is more accurate than estimator 2.

At this point, you might be thinking we have to draw all possible samples from the population to get a sampling distribution of an estimator. Fortunately, statistics tells us that a single sample is enough to allow us to approximate the sampling distribution of most estimators. We will make use of this fact whenever we want to determine the sampling distribution of an estimator.

HOW ACCURATE AN ESTIMATOR IS THE SAMPLE MEAN?

The accuracy of $\bar{x}$ is determined by its sampling distribution. What is the sampling distribution of $\bar{x}$? Since $\bar{x}$ is an unbiased estimator of μ, its sampling distribution has a mean of μ, the population mean. The standard deviation of the sampling distribution of $\bar{x}$, denoted $\sigma_{\bar{x}}$, is equal to the population standard deviation divided by the square root of the sample size, i.e.,

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Furthermore, as long as the sample size is not too small, the sampling distribution of $\bar{x}$ is approximately a normal distribution.[5] In sum, $\bar{x}$ has a sampling distribution that is normal with a mean of μ and a standard deviation of $\sigma_{\bar{x}}$. Equivalently,

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

has a standard normal (or z) distribution.

ESTIMATING THE SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

Since the population standard deviation is never observed, we must estimate it. The best estimator of the standard deviation of the sampling distribution of $\bar{x}$ (i.e., $\sigma_{\bar{x}}$) is denoted by $s_{\bar{x}}$, and is usually referred to as the **standard error of the mean**. $s_{\bar{x}}$ equals the sample standard deviation divided by the square root of the sample size

[5] It is exactly a normal distribution only when the population is normally distributed. However, as long as the sample size is not too small, a result known as the Central Limit Theorem tells us that the sampling distribution is approximately normal.

$$\left(\frac{s}{\sqrt{n}}\right)$$

Since the standard error of the mean is only an estimate based on the sample, it introduces some additional sampling error into our calculations. This causes

$$\frac{\bar{x} - \mu}{s_{\bar{x}}}$$

to have a t-distribution with n-1 degrees of freedom,[6] whereas as we saw above,

$$\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

has the standard normal (or z) distribution. The additional sampling error is reflected in the fatter tails of the t-distribution compared to the standard normal. This is why the t-distribution will appear so often in this text and in statistics. We will often use the notation

$$\text{t} = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

because this quantity has a t-distribution.

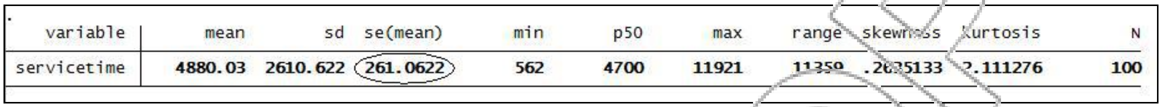
COMPUTING THE STANDARD ERROR OF THE MEAN

You can calculate the standard error of the mean, $s_{\bar{x}}$, in two different ways. Once you know the sample standard deviation, s, dividing it by the square root of the sample size ($\sqrt{n}$) yields $s_{\bar{x}}$. Proceeding in this fashion, we have the following:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{2610.622}{\sqrt{100}} = \frac{2610.622}{10} = 261.0622 \text{ seconds.}$$

[6] This is exactly true only when the population is normally distributed but is often a good approximation if it is not.

We can alternatively calculate $s_{\bar{x}}$ using the **User>Core Statistics>Univariate Statistics>Standard (ktabstat)** command. The standard error of the mean is the number in the **se(mean)** column (see Figure 1.16). Stata calculates this number to be 261.0622 seconds.



```
.
    variable |      mean        sd  se(mean)       min       p50       max     range  skewness  kurtosis         N
-------------+--------------------------------------------------------------------------------------------------------
 servicetime |   4880.03  2610.622  261.0622       562      4700     11921     11359  .2635133  2.111276       100
----------------------------------------------------------------------------------------------------------------------
```

Figure 1.16: Univariate Statistics for servicetime (standard error or the mean).

Side Comments:

In the above discussion of the sampling distribution of $\bar{x}$, we have been implicitly assuming that the sample from which $\bar{x}$ was calculated was gathered using a good sampling procedure. What makes a sampling procedure good? In a good sampling procedure, each observation should be randomly selected from the population of interest and each observation should be chosen independently of any other. Choosing observations independently means that the probability of choosing a particular observation does not depend on other observations. Such a sample is often referred to as **independently and identically distributed (i.i.d.)**.

1.9 Confidence Intervals

Having obtained an estimate, we will be interested in ascertaining its accuracy, i.e., how close the estimate is to the true value. The service manager at EE has calculated the estimated mean time spent by an EE salesperson attending to pseudo customers per day to be 4880.03 seconds. It is important for him to know how precise this estimate is. He would be happy if his estimate came within, for example, 120 seconds of the true mean. On the other hand, he might be unhappy and the estimate would be quite misleading if the estimate were 1500 seconds away from the mean. Therefore, we would like to know the probability that the estimate will be within or beyond a certain distance of the mean.

What is the probability that the estimate meets the service manager's accuracy needs? In other words, what is the proportion of samples of size n for which our estimate (the sample mean, $\bar{x}$) is within 120 seconds of the population mean, μ. In probability terms, we would like to know the probability that the sample mean is within 120 seconds of the true mean. Using the notation for probability statements, we can write this as Prob(-120 $\leq \bar{x} - \mu \leq$ 120).

From the previous sections, we know that

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

has a t-distribution with n-1 degrees of freedom. We can use this to do the following simplification of the above probability statement:

$$\text{Prob}[-120 \leq \bar{x} - \mu \leq 120]$$

$$= \text{Prob}[-120/s_{\bar{x}} \leq (\bar{x} - \mu)/s_{\bar{x}} \leq 120/s_{\bar{x}}]$$

$$= \text{Prob}[-120/s_{\bar{x}} \leq t \leq 120/s_{\bar{x}}]$$

= Area between -120/$s_{\bar{x}}$ and 120/$s_{\bar{x}}$ under a t-distribution with n-1 degrees of freedom.

In going from the first line to the second line in the above box, we divided through by $s_{\bar{x}}$. From the sample, we can calculate $s_{\bar{x}}$ by using Stata. In fact, we did compute its value previously as 261.06 seconds after rounding. Hence, in this example:

$120/s_{\bar{x}} = 120/261.06 = 0.46$

$-120/s_{\bar{x}} = -120/261.06 = -0.46$

Since n = 100, the t-distribution has 99 degrees of freedom (100-1 = 99). Therefore, the required probability is the area between -0.46 and 0.46 under a t-distribution with 99 degrees of freedom. This is the shaded area in Figure 1.17.

We can use the **ttail** command to calculate this. In the Stata Command box, type in **display 2*ttail(99,0.46)**. Stata returns the value 0.6465249. So the required probability is 0.35, i.e.,

$$\text{Prob}[-0.46 \leq (\bar{x} - \mu)/s_{\bar{x}} \leq 0.46] = 1-0.6465249 \approx 0.35.$$

This implies that the service manager's estimate of the average time spent by an EE salesperson interacting with pseudo customers per day has a probability of 0.35 of being within 120 seconds of the true average time spent with pseudo customers by a salesperson daily.

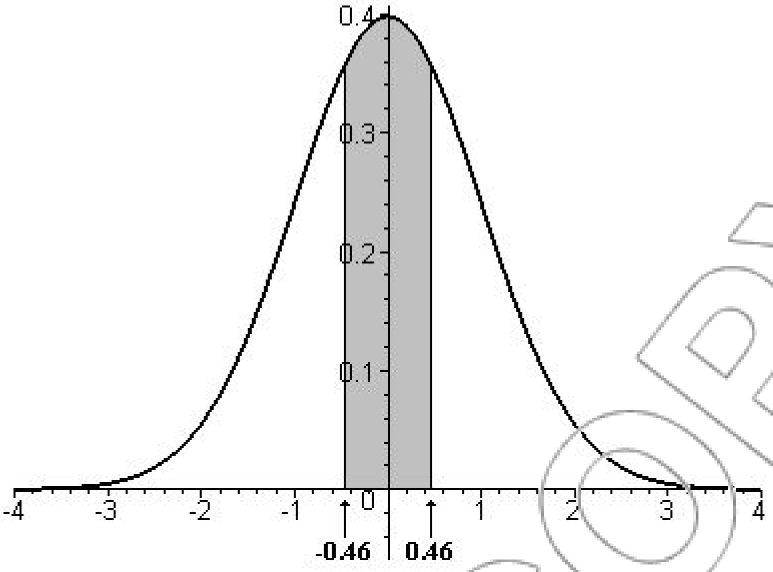


Figure 1.17: t-distribution with 99 degrees of freedom.

In the form of an equation, we have shown the following:

$$\boxed{\text{Prob}[\,\bar{x} - 120 \le \mu \le \bar{x} + 120] = 0.35}$$

In other words, we calculated the probability of selecting a sample of size 100 that gives a sample mean time within 120 seconds of the true mean. However, once we have the sample, the sample mean either is within 120 seconds of the true mean or it is not. For this reason, it would be incorrect to plug in $\bar{x}$ = 4880.03 seconds (as calculated previously) and conclude that the probability the true mean, μ, is between 4760.03 seconds and 5000.03 seconds is 0.35. Instead, we say that we are 35% [= (0.35)(100)%] *confident* the true mean is between 4760.03 seconds and 5000.03 seconds. Specifically, if our sample is one of the 35% of possible samples having a sample mean that is within 120 seconds of the population mean, then the interval we calculated for μ will contain the true mean.

Why do we say that we are 35% confident that the true mean is between 4760.03 seconds and 5000.03 seconds rather than saying the probability the true mean is between 4760.03 seconds and 5000.03 seconds is 0.35 or 35%? This distinction between confidence and probability emphasizes that the randomness lies in which elements of the population are observed in the sample and not in the value of the population mean. Informally, any given sample you observe may be more or less representative of the population as a whole. If the sample happens to be more representative, the sample mean will be close to the population mean. On the other hand, if the sample is unrepresentative, then the sample mean will lie far from the population mean. Of course, one can never tell whether a particular sample is representative. The best you can do is know the probability of obtaining such a sample.

We have just seen how to calculate how confident we are that the population mean is in a given range. We can also reverse the procedure and find the range that we have a given confidence contains the population mean. For example, what is the range within which we are 95% confident that the true mean falls? The answer to this is called a 95% confidence interval for the population mean, μ. Once the sample mean, $\bar{x}$, and the standard error of the mean, $s_{\bar{x}}$, are known, computing the confidence interval for the population mean, μ, is straightforward. However, before we proceed, it is necessary to introduce a new notation.

For α between 0 and 1, $\mathbf{t_{\alpha/2,\,(n-1)}}$ is the value such that there is a $\boldsymbol{\alpha/2}$ probability of being above that value in a t distribution with ***n-1*** degrees of freedom. In Stata, $\mathbf{t_{\alpha/2,\,(n-1)}}$ = invttail(n-1, α/2).

For example, if $\alpha = 0.05$ and n = 100, then $t_{0.05/2,\,(100-1)} = t_{0.025,\,99}$. Figure 1.18 illustrates the meaning of $t_{0.025,\,99}$ graphically. Using Stata, we can calculate $t_{0.025,\,99}$ = invttail(99, 0.025) = 1.98.

Using the above notation, a (1-α)(100)% confidence interval for the population mean, μ, is given by the following:

The (1-α)(100)% Confidence Interval for μ is [$\bar{x} - t_{\alpha/2\,,\,(n-1)}\ s_{\bar{x}}$, $\bar{x} + t_{\alpha/2\,,\,(n-1)}\ s_{\bar{x}}$]

(1-α)(100)% is called the level of confidence (or confidence level). A 95% confidence interval for μ tells us that 95% of the time a sample of size n is drawn from the population and used to calculate a 95% confidence interval that interval will contain μ. For a graphical representation, see Figure 1.19.

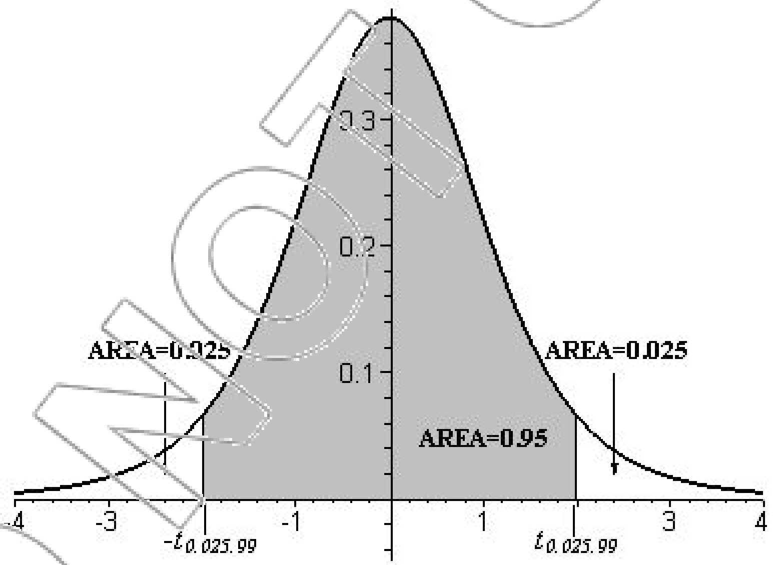


Figure 1.18: A t-distribution with 99 degrees of freedom with $t_{0.025,99}$ indicated.

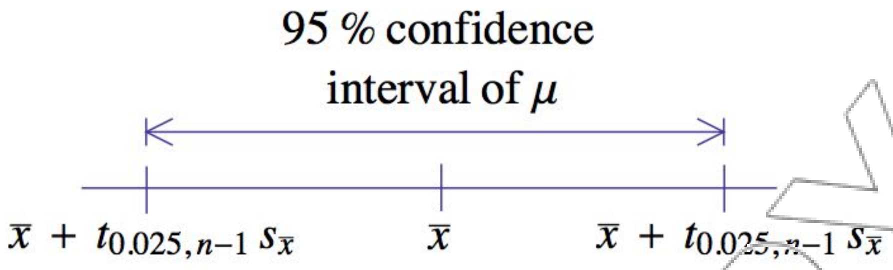

95 % confidence
interval of μ
$\bar{x} + t_{0.025,n-1}\, s_{\bar{x}}$
$\bar{x}$
$\bar{x} + t_{0.025,n-1}\, s_{\bar{x}}$

Figure 1.19: 95% Confidence interval.

We will see how to calculate confidence intervals with the help of an example. Suppose we want to find the 95% confidence interval for the mean service time for pseudo customers. As we have seen above, to calculate the confidence interval, we will need to know the values of the following three quantities: $\bar{x}$, $s_{\bar{x}}$, and $t_{\alpha/2\ ,\ (n-1)} = t_{0.05/2\ ,\ (100-1)} = t_{0.025\ ,\ 99}$.

We know that $\bar{x}$ and $s_{\bar{x}}$ are equal to 4880.03 seconds and 261.06 seconds, respectively. To calculate $t_{0.025,\ 99}$ using Excel, we could use the TINV command with 0.05 for the **Probability** and 99 for **Deg_freedom.** As above, TINV(0.05, 99) = 1.98. The value of $t_{0.025\ ,\ 99}$ can also be calculated using the **invttail** command in Stata. Typing **display invttail(99,0.025)** in the Command box will also produce the value 1.98.

Therefore, the 95% confidence interval for the mean service time for pseudo customers is the following:

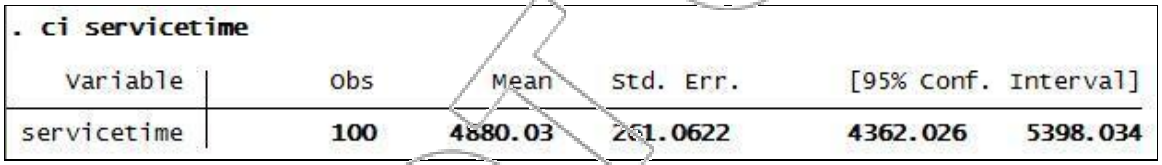
$[\ \bar{x} - t_{0.025\ ,\ 99}\, s_{\bar{x}}\ ,\ \bar{x} + t_{0.025\ ,\ 99}\, s_{\bar{x}}\ ]$

= [4880.03 – (1.98)(261.06) , 4880.03 + (1.98)(261.06)]

= [4363.13 , 5396.93]

This means we are 95% confident the average time spent by a service person interacting with pseudo customers in one day is between 4363.13 seconds and 5396.93 seconds.

In Stata, you can automatically calculate the 95% confidence interval for the population mean, μ, of a variable by using the **Confidence interval** command. Consider the previous example where we want to calculate the 95% confidence interval for the mean service time for pseudo customers. To calculate this in Stata, open **service.dta** and click **Statistics>Summaries, tables, and tests>Summary and descriptive statistics>Confidence intervals** (or type **db ci**). Choose **servicetime** as your variable and click **OK**.[7] You should get the following:



```
. ci servicetime

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
 servicetime |        100     4880.03    261.0622        4362.026    5398.034
```

Stata calculates the 95% confidence interval for the mean service time to be [4362.026, 5398.034]. The discrepancy between the Stata output and our manually calculated result is due to our rounding of $t_{0.025,99}$ to two decimal places.

Note that in Stata, you can easily calculate the confidence interval for the population mean of a variable for any confidence level. For example, to find the 90% confidence interval for the mean service time, simply type **ci servicetime, level(90)** and get [4446.565, 5313.495].

[7] Alternatively, you can directly type the command **ci servicetime** into the Command box.

The standard error of the mean plays a crucial role in determining the width of a confidence interval. This makes sense since we learned previously that the smaller the standard deviation of the sampling distribution, the more accurate an estimator is.

Confidence intervals can identify reasonable best (or worst) case scenarios regarding the mean value. For example, since the 95% confidence interval for the mean service time for pseudo customers is (4363.13 seconds, 5396.93 seconds), we can say, "We are 95% confident that salespeople spend at least an average of 4363.13 seconds interacting with pseudo customers per day and at most an average of 5396.93 seconds per day with pseudo customers." Furthermore, we can say, "We are 97.5% confident that salespeople spend at most an average of 5396.93 seconds per day." [8] Similarly, "We are 97.5% confident that salespeople spend at least an average of 4363.13 seconds per day with pseudo customers."

Now that management estimated the time spent with pseudo customers, what should its decision be? Since 5396.93 seconds is fewer than 5760 seconds (equal to the 1 hour 36 minutes cutoff that management decided on), management is 97.5% confident that average time spent by an EE salesperson serving pseudo customers is less than the threshold. Management, therefore, should conclude that pseudo customers are not a large enough drain on salespersons' resources to change policy given the costs and disruptions involved with these changes.

Confidence intervals may also be constructed for proportions and we briefly discuss them here. The special properties of proportions that we discussed earlier are useful in this regard. For

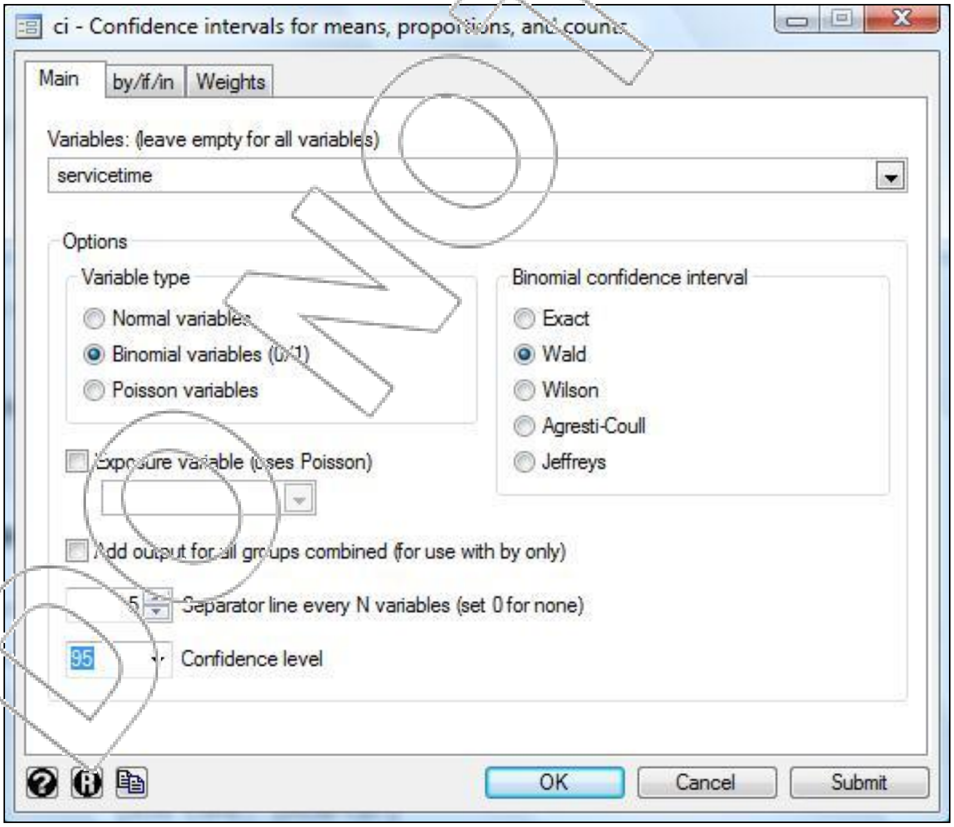
[8] How did we get 97.5% confidence when the 5396.93 seconds figure comes from a 95% confidence interval? A 95% confidence interval is constructed so the mean will be below the lower bound of the interval for 2.5% of samples, above the upper bound of the interval for 2.5% of samples and between the interval limits for 95% of samples. If we want to say how confident we are that the mean will be below the upper bound without specifying whether it is above or below the lower bound, then our confidence level is 95% plus the 2.5% below the interval to make a total of 97.5%.

instance, with a sample proportion of $\bar{p}$, the **standard error of the proportion** $s_{\bar{p}}$ is equal to

$\sqrt{\bar{p}(1-\bar{p})/n}$. $\dfrac{\bar{p}-p}{s_{\bar{p}}}$ has approximately a standard normal (or z-) distribution. A (1-α)(100)% confidence interval for the proportion is

$$[\ \bar{p} - z_{\alpha/2}\, s_{\bar{p}}\ ,\ \ \bar{p} + z_{\alpha/2}\, s_{\bar{p}}\ ].$$

Note that in Stata, you can easily calculate these confidence intervals for proportions. After loading your dataset of interest, click **Statistics>Summaries, tables, and tests>Summary and descriptive statistics>Confidence intervals** or type **db ci**. Enter the name(s) of binary variable(s) in the "Variables" field, and choose **Binomial variables** and **Wald** as your variable type and binomial confidence interval, respectively. You can specify the confidence level at the bottom of the dialog box. You should have a dialog box that looks like this:



ci - Confidence intervals for means, proportions, and counts

Variables: (leave empty for all variables)
servicetime

Options

Variable type
○ Normal variables
● Binomial variables (0/1)
○ Poisson variables

Binomial confidence interval
○ Exact
● Wald
○ Wilson
○ Agresti-Coull
○ Jeffreys

□ Exposure variable (uses Poisson)

□ Add output for all groups combined (for use with by only)

5 Separator line every N variables (set 0 for none)

95 Confidence level

Click **OK**, and Stata will report the sample proportion (displayed under the **Mean** column), standard error of the proportion (**Std. Err.**), and the (1-α)(100)% confidence interval for the proportion (**(1-α)% Conf. Interval**) of your selected variable.[9]

SUMMARY

In this chapter, we introduced several important ideas including discrete and continuous probability distributions, the mean, variance and standard deviation, proportions, and the normal and t-distributions. We worked extensively on integrating Excel and Stata into our understanding of these concepts. Later, we learned how to use Stata to estimate the mean and standard deviation and other aspects of probability distributions given a data sample. We learned how to use that same data to quantify the accuracy of these mean estimates using the standard error of the mean and confidence intervals for the mean. We also examined the special case of proportions.

NEW TERMS

Probability distribution A description of how probabilities are spread out over possible outcomes

Discrete probability distribution A distribution which can only take on a certain countable number of values

Continuous probability distribution A distribution that can take on any value within a given range or ranges

[9] Selecting the Wald binomial confidence interval uses the formula presented above. This relies on the central limit theorem to approximate the binomial with a normal distribution. Selecting Exact instead of Wald will calculate a confidence interval based on the binomial distribution itself rather than the approximation. Neither is unambiguously more correct or useful than the other.

Mean The center or average of a distribution

Variance A measure of the spread around the mean determined by averaging the squared deviations from the mean

Standard deviation A measure of the spread around the mean determined by taking the square root of the variance

Normal distribution Any of the family of common bell-shaped probability distributions

Standard normal distribution A normal distribution with mean of 0 and standard deviation of 1

t-distribution Another family of distributions similar to the standard normal but with fatter tails

Degrees of freedom A parameter used to characterize the t-distribution

Population The entire set of values of interest

Sample The portion of the population that is observed

Sample size The number of observations in the sample

Sample mean The mean or average of the values in the sample, denoted by $\bar{x}$

Sample variance The variance of the sample, denoted by s^2

Sample standard deviation The standard deviation of the sample, denoted by s

Sampling distribution of the sample mean The probability distribution of $\bar{x}$

Unbiased An estimator whose mean is equal to the parameter being estimated

Standard error of the mean An estimate of the standard deviation of the sampling distribution of $\bar{x}$, denoted by $s_{\bar{x}}$ and equal to $\frac{s}{\sqrt{n}}$.

independent and identically distributed (i.i.d.) A sampling procedure that creates a sample with desirable properties

Confidence interval A range of values that will contain the mean of the population with a certain specified level of confidence

NEW STATA AND EXCEL FUNCTIONS

STATA

User>Core Statistics>Univariate Statistics>Standard (ktabstat)

This command generates univariate statistics for all variables contained in the current Stata data file. These statistics include the sample mean, sample standard deviation, standard error of the mean, minimum, median, maximum, range and sample size. It also generates some other measures of the variables' distributions such as skewness and kurtosis that we will not make use of here.

Alternatively, you can directly type the command **ktabstat**.

User>Core Statistics>Univariate Statistics>Custom (tabstat)

This command allows you to specify up to eight statistics that you want Stata to display. The direct command is **tabstat *varlist*, s(…)**, where ***varlist*** corresponds to the name of the variables for which you want to calculate the summary statistics. You can specify the names of summary statistics in the **s(…)** portion of the command. (For the complete list of summary statistics, type **help tabstat** into the Stata Command box and refer to the Options>statistics section.) Typing **_all** instead of ***varlist*** will generate univariate statistics for all variables currently listed in Stata. Note that Stata cannot generate univariate statistics for string, or non-numeric, variables. Therefore, if there is any string variable present in your dataset, typing the direct command **tabstat _all, s(…)** will result in an error. You can still execute the **tabstat** command on numeric variables by omitting the names of string variables from ***varlist***. However, it is generally easier to use the **ktabstat** command instead, where it is programmed to convert string variables to numeric variables temporarily prior to executing the **tabstat _all, s(…)** command. Your original dataset will not be affected by this temporary conversion.

Statistics>Summaries, tables, and tests>Summary and descriptive statistics>Confidence intervals

Alternatively, you may type **db ci**. This opens the Stata **ci** dialog box, where you can choose variable(s) for which you want to calculate confidence intervals for the population mean(s).

Alternatively, you can directly type the command **ci *varlist*, level(#)**, where # corresponds to (1-α)(100)%. Omitting the **level(#)** option will generate a 95% confidence interval for the population mean of a variable by default.

To calculate confidence intervals for proportions through the **ci** dialog box, choose **Binomial variables** and **Wald** in the “Variable type” and “Binomial confidence interval” field, respectively.

Alternatively, you can directly type the command **ci *varlist*, binomial wald level(#)**.

normal(z)

Typing **display normal(z)** into the Command box will return the area to the left of z under the standard normal distribution.

invnormal(p)

Typing **display invnormal(p)** into the Command box will return the value x for which the probability of falling to the left of that value under the standard normal distribution is p.

ttail(n,t)

Typing **display ttail(n,t)** into the Command box will return the area to the right of t under a t-distribution with n degrees of freedom. You may enter a positive or negative value for t.

invttail(n,p)

Typing **display invttail(n,p)** into the Command box will return the value x for which the probability of falling to the right of that value is p under a t-distribution with n degrees of freedom.

EXCEL

AVERAGE

Typing =AVERAGE(A2:A7) into a blank cell will return the average of the numbers in cells A2:A7. You can select **Insert>Function** and choose AVERAGE from the list of statistical functions.

NORMDIST

Typing =NORMDIST(20,25,10,1) into a blank cell will return the area to the left of 20 under the normal distribution with a mean of 25 and a standard deviation of 10.

NORMINV

Typing =NORMINV(0.318,25,10) into a blank cell will return a number such that the probability of obtaining a value less than that number from a normal distribution with a mean of 25 and standard deviation of 10 will equal 0.318.

NORMSDIST

Typing =NORMSDIST(-1.91) into a blank cell will provide you with the area under the standard normal curve to the left of -1.91. This area equals the probability of having an outcome from a standard normal less than -1.91. To find the probability of an outcome greater than +2.04 (the area under the curve to the right of 2.04), use =1-NORMSDIST(2.04).

NORMSINV

Typing =NORMSINV(0.42) into a blank cell will return a number such that the probability of obtaining a value less than that number from a standard normal distribution will equal 0.42.

TDIST

Typing =TDIST(1.76,48,1) into a blank cell will return the area above 1.76 in a t-distribution with 48 degrees of freedom. Typing =TDIST(1.76, 48, 2) will return the area above 1.76 plus the area below -1.76 in a t-distribution with 48 degrees of freedom. You may not enter a negative number for the first argument. You can select **Insert>Function** and choose TDIST from the list of statistical functions.

TINV

Typing =TINV(0.05,98) into a blank cell returns the value having area 0.025 above it in a t-distribution with 98 degrees of freedom. This tells you how far in each direction one would have to go from the mean to get an area of 1-0.05 = 0.95 underneath the t-distribution.

NEW FORMULAS

The (1-α)100% confidence interval for a mean: [$\bar{x}$ - $t_{\alpha/2\,,\,(n-1)}\ s_{\bar{x}}$, $\bar{x}$ + $t_{\alpha/2\,,\,(n-1)}\ s_{\bar{x}}$]

The (1-α)100% confidence interval for a proportion: [$\bar{p}$ - $z_{\alpha/2}\, s_{\bar{p}}$, $\bar{p}$ + $z_{\alpha/2}\, s_{\bar{p}}$]

CASE EXERCISES

1. Return to me

A Hawaiian hotel chain is interested in studying tourists who travel to the state. One question they are investigating is whether or not tourists who return to the islands stayed at the same hotel as in their previous trip. The data file **return** lists the responses of 1,000 tourists who were involved in the study. A one (1) indicates they did return to the same hotel whereas a zero (0) indicates they did not. Calculate the proportion of tourists in the study who stay at the same hotel as they had on their previous trip. Using the formulas in section 1.4 and the proportion you just calculated, calculate the variance and standard deviation of the responses in the study.

2. EE TV sales

The weekly sales of flat panel televisions at one EE store (store A) follow a normal distribution with mean of 12 and standard deviation of 4. Store B usually has lower sales normally distributed but with mean of 9 and standard deviation of 3. If the two stores currently have 18 and 14 flat panel televisions in stock, respectively, and neither will receive a new shipment for the next week, determine which store has the higher probability of running out of stock.

If the company has declared that each store should stock enough inventory so the chances of running out of stock are at most 2%, determine the minimum number of flat panel televisions each store should keep in its weekly inventory to comply with the rule.

3. EE job applications

Certain data from EE's 4,000 stores are not entered into its electronic data base. For instance, employment applications are typically handwritten on paper forms and never re-entered into their computer system. EE would like to learn more about the acceptance rate for entry-level employees. Specifically, it feels that if stores are accepting more than half of their applicants, then the quality of the typical employee may suffer. Since entering these data for its hundreds of thousands of applicants would be expensive and time consuming, EE has decided to use sampling to learn about this issue. Access the data in the file **EESample**, which contains information from a random sample of 55 EE stores.

a. Determine the sample mean, sample standard deviation, and the standard error of the mean.
b. Construct a 95% confidence interval for the true mean acceptance rate of entry-level job applicants at EE stores.
c. Construct a 90% confidence interval for the true mean acceptance rate of entry-level job applicants at EE stores.
d. Assuming the true mean acceptance rate of entry-level jobs was 50%, determine the chances that the sample mean could have been as low as it is or even lower.
e. What does your answer to part d tell you about the feasibility of the assumption about the true mean?

4. EE stores

The management at EE wants to investigate the consistency in hiring practices across all of its stores. Rather than learning whether the mean acceptance rate for all EE stores is less than 50%, it wants to know the probability that any given store has an acceptance rate above 50 percent.

Access the data in the file **EESample**, which contains information from a random sample of 55 EE stores.

Create an additional column of data called **half_plus** which is equal to one (1) if the acceptance rate is greater than 50 percent.[10]

a. Determine the sample proportion for the fraction of EE stores which hire more than half of their applicants.

b. Provide a 95% confidence interval for the true proportion.

c. Provide a 70% confidence interval for the true proportion.

d. Assuming the true proportion of stores that accept over half of their applicants is 0.50, determine the chances that our sample proportion would have been as low as it is or even lower.

e. What does your answer to part d tell you about the feasibility of the assumption about the true proportion?

5. Cashing out

A local mortgage bank in New Jersey is interested in knowing more about its customers. Specifically, it would like to understand how much home equity customers who refinance their homes are likely to cash out. A sample of 65 loans is contained in the file **njbank**.

a. Determine the sample mean, sample standard deviation, and the standard error of the mean for the amount of home equity cashed out.

[10] To do this in Stata, first load the **EESample.dta** file. Then, you can type the following commands: 1) **generate half_plus=1 if Acceptance_Rate>0.5**, and 2) **replace half_plus=0 if half_plus==.** (make sure to include the period after ==). Open the Data Browser to verify that the new data are generated correctly. See the Appendix for general instructions on how to generate and/or manipulate variables in Stata.

b. Construct a 95% confidence interval for the true mean cash out value for customers at the bank

c. Construct an 82% confidence interval for the true mean cash out value for customers at the bank.

The bank is interested in the proportion of customers who did not take any cash out when they refinanced. Make a new column of data titled **No_Cash** that equals one (1) if the customer took no cash out and zero (0) for all other amounts. [11]

d. Determine the sample proportion of customers who did not take any cash out when they refinanced.

e. Construct a 95% confidence interval for the true proportion of customers who did not take any cash out when they refinanced.

f. If the true proportion of customers who did not take any cash out when they refinanced is equal to 0.5, determine the chances that the bank would have discovered a sample proportion as low as or lower than it did in its sample.

Problems

1. Given z follows a standard normal distribution, determine the following:

a. Prob(z < 2.8)

b. Prob(z < 1.8)

c. Prob(z < 0.8)

d. Prob(z < -0.2)

[11] To do this in Stata, first open **njbank.dta**. Then, you can type the following commands: 1) **generate No_Cash=1 if Cash_Out==0**, and 2) **replace No_Cash=0 if No_Cash==.** (make sure to include the period after ==). Open the Data Browser to verify that the new data are generated correctly.

e. Prob(z < -1.2)

2. Given that z follows a standard normal distribution, determine the following:

a. Prob(z > 2.3)

b. Prob(z > 1.3)

c. Prob(z > 0.3)

d. Prob(z > -0.7)

e. Prob(z > -1.7)

3. Given that z follows a standard normal distribution, determine the following:

a. Prob(2.9 > z > 2.1)

b. Prob(1.9 > z > 1.1)

c. Prob(0.9 > z > 0.1)

d. Prob(-0.3 > z > -1.1)

e. Prob(-1.3 > z > -2.1)

4. Given that x follows a normal distribution with mean of 55 and standard deviation of 12, determine the following:

a. Prob (x < 90)

b. Prob (x < 71)

c. Prob (x < 57)

d. Prob (x < 42)

e. Prob (x < 25)

5. Given that x follows a normal distribution with mean of 7 and standard deviation of 20, determine the following:

a. Prob (x > 30)

b. Prob (x > 9)

c. Prob (x > 2)

d. Prob (x > -12)

e. Prob (x > -29)

6. Given that x follows a normal distribution with mean of 800 and standard deviation of 350, determine the following:

a. Prob (1000 < x < 1200)

b. Prob (800 < x < 1000)

c. Prob (600 < x < 800)

d. Prob (400 < x < 600)

e. Prob (200< x < 400)

7. Given that z follows a standard normal distribution, determine the value of z for the following examples:

a. The area to the left of z equals 0.50

b. The area to the left of z equals 0.18

c. The area to the left of z equals 0.025

d. The area to the right of z equals 0.29

e. The area to the right of z equals 0.10

f. The area to the right of z equals 0.05

8. For a t distribution with 24 degrees of freedom, determine the following:

a. Prob (t > 1.25)

b. Prob (t > 0.92)

c. Prob (t > 0.58)

d. Prob (t > 0.21)

e. Prob (t > -0.25)

f. Prob (t > -2.05)

9. For a t distribution with 64 degrees of freedom, determine the following:

a. Prob (t < 1.55)

b. Prob (t < 0.72)

c. Prob (t < 0.18)

d. Prob (t < 0.04)

e. Prob (t < -0.75)

f. Prob (t < -1.99)

10. A Gallup Poll (*Will Investors Jump on the Optimism Bandwagon?* October 27, 2003) noted that 57% of investors say the economy has hit bottom. The article also states that the survey included a random sample of 802 adult investors. Determine a 95% confidence interval for the true proportion of investors who would say that the economy has hit bottom.

11. In response to concern by many of its clients, Nucleus Research reported findings from a recent study on spam and employee productivity (*Spam: The Silent ROI Killer* September 24, 2003) The article noted that the average employee in its survey of 117 workers spent 6.5 minutes per day dealing with unwanted emails or spam. Assuming the sample standard deviation, s, is 14 minutes per day, determine a 90% confidence interval for the true mean number of minutes per day that employees spend dealing with spam.

12. You are given a sample consisting of 83 data points with a sample mean of 37 and a sample standard deviation of 21.

a. Construct a 90% confidence interval for the true mean

b. Construct a 95% confidence interval for the true mean

c. Construct a 99% confidence interval for the true mean

13. A sample of 43 data points results in a sample mean of 1.15 and a sample standard deviation of 0.482.

a. Construct a 90% confidence interval for the true mean

b. Construct a 95% confidence interval for the true mean

c. Construct a 99% confidence interval for the true mean

CHAPTER 2

CONSUMER PACKAGING: CONDUCTING AND USING HYPOTHESIS TESTS

In this chapter, you will learn about one of the most important and widely applied statistical techniques: hypothesis testing. Hypothesis testing is a basic tool we will use throughout the course when we want to convince ourselves or others that our data provide evidence for some fact about the world. For example, we will use hypothesis testing to study the effectiveness of our test marketing, identify political gender gaps, and confirm stylized facts regarding stock market anomalies. We will also use it in later chapters as a central piece of the regression model.

2.1 Hypothesis Testing: How to Make Your Case with Data

In the first chapter, you learned some of the basics of how to use data to estimate important features of the world. For example, by observing sales in test markets, you can form an estimate of average sales in a full product rollout by calculating the sample average in the test markets. Similarly, by collecting data on visitors to an e-commerce web site, you can form estimates of useful quantities, such as the proportion of visitors clicking on banner ads and the proportion arriving at the site through links on third-party sites. You also learned how to use confidence interval estimates to help assess the accuracy of your estimates.

One of the primary uses of statistical estimates is to convince others (or even ourselves) that something is true. Whether you are the one looking for an advantage by using statistics to bolster your argument or you are the person whom the presenter wants to convince, you must understand how estimates can be used as proof or evidence. The method used to prove or support arguments with statistics is called **hypothesis testing**. In this section, we will learn the fundamentals of hypothesis testing and see some applications with marketing and financial data using estimators you learned about in the previous chapter. As we move through this text and learn and apply new and more sophisticated estimation techniques, hypothesis testing will continue to play a prominent role.

A good, non-technical way to understand much of the logic and terminology associated with hypothesis testing is to think of a criminal trial in a court of law. Imagine for a moment that you are a prosecuting attorney in a murder case. Your goal is to prove to the jury that the defendant is guilty of murder. In hypothesis testing, what you would like to prove is called the **alternative hypothesis** (often denoted $\mathbf{H_a}$ or sometimes $\mathbf{H_1}$**)**. All the possibilities that are not in the alternative hypothesis are called the **null hypothesis** (denoted $\mathbf{H_0}$). For example, for the lawyer, the null hypothesis is that the defendant is not guilty of murder, and the alternative hypothesis is that the defendant is guilty of murder. The null and alternative hypotheses do not overlap and, together, cover all possibilities. In other words, the null is true, or the alternative is true, but not both. The null and alternative must always be set up so this is the case.

What are the possible outcomes of the trial? Either the jury will find the evidence convincing enough to declare the defendant guilty or it will not, in which case the defendant is declared not guilty. Similarly, in a hypothesis test, either the evidence (based on the data) is strong enough for you to accept the alternative hypothesis as true, or it is not. For historical reasons, accepting the alternative is more commonly referred to as "rejecting the null hypothesis." Since at least one of the two hypotheses must be right, rejecting the null hypothesis is the same as accepting the alternative hypothesis. (Ensure you understand this.) Thus, the two possible outcomes of a hypothesis test are rejecting the null hypothesis and not rejecting the null hypothesis. A hypothesis test can never result in rejecting the alternative hypothesis or, equivalently, accepting the null hypothesis. If a jury finds the defendant not guilty, that means the evidence was not strong enough to prove the defendant guilty. It does not mean the evidence proved the defendant was innocent. Standard criminal trials are not set up to prove innocence. They can only prove or fail to prove guilt. The same is true of hypothesis tests. They can only reject the null or fail to reject the null. This is why you must ensure when setting up a hypothesis test that the alternative hypothesis is what you hope to prove; it is impossible to prove a null hypothesis using a hypothesis test.

What makes evidence strong or weak? In hypothesis testing, we say that evidence (in support of the alternative or, equivalently, against the null) is strong if, assuming the null hypothesis were true, the evidence would be unlikely to have been found. Two examples from the trial should make this clear. Suppose the victim had been strangled and fingerprints found on the victim's neck matched the defendant's fingerprints. Is this strong or weak evidence? To evaluate this, we must ask ourselves what the probability of a matching fingerprint appearing on the victim's neck would be if the defendant were not guilty. Assuming the defendant was not someone who had some other reason to be close to the victim (e.g., assume they were not spouses), then this probability would be small. This is what it means to have strong evidence. On the other hand, suppose we discover the murderer was wearing blue jeans. Furthermore, we discover the defendant owns a pair of blue jeans. Is this strong evidence? Well, what is the probability, assuming that the defendant is not guilty, that he or she would own at least one pair of blue jeans? This probability is high as many people who are not murderers wear blue jeans. Therefore, this is weak evidence and would be insufficient to prove guilt. The statistical measure of strength of evidence, expressed in probability terms, is called the p-value. As in the above examples, low p-values correspond to strong evidence against the null/supporting the alternative, and high p-values correspond to weaker evidence.

So, strong evidence favors rejecting the null (finding the defendant guilty) and weak evidence does not, but how strong should we require the evidence to be before we reject (or declare guilt)? Statistics, like the courts, cannot deliver perfection. Just as a jury will sometimes come to the wrong verdict, a hypothesis test will sometimes lead to an incorrect conclusion. A trial can have two types of errors: (1) The jury could find the defendant guilty when, in fact, he or she is innocent and (2) the jury could fail to find the defendant guilty when, in fact, the defendant is guilty. In hypothesis testing terms, error (1) is rejecting the null hypothesis when the null hypothesis is true. This is called **type I error**. As you might guess, errors like (2) (i.e., not rejecting the null hypothesis when the null is false) are called **type II errors**. Ideally, we would like the probability of making each of these errors to be small (in the courtroom and in hypothesis testing). In the court, we can control the probability of a type I error by setting the standard of proof required for a conviction. For example, many of you have probably heard the phrase "beyond any reasonable doubt" used in this regard. In many trials, the jury is not supposed to return a guilty verdict unless the evidence shows beyond any reasonable doubt the defendant is guilty. Of course, this verbal directive is vague and open to interpretation, but it suggests the jury should not convict unless it is convinced the probability of a type I error is small. In hypothesis testing, as in the courtroom, we have to set a standard of proof. We do this by choosing a **level of significance** (denoted by the Greek letter alpha, α) between 0% and 100% (0.00 and 1.00). The level of significance states the maximum probability of a type I error that is acceptable. So, if you conduct a hypothesis test using a small level of significance, it will take strong evidence for you to reject the null hypothesis. If you do reject the null in such a case, however, it is unlikely that you have done so in error. On the other hand, setting a higher level of significance allows you to prove your point (reject the null) more often but with a higher probability of making the point in error.

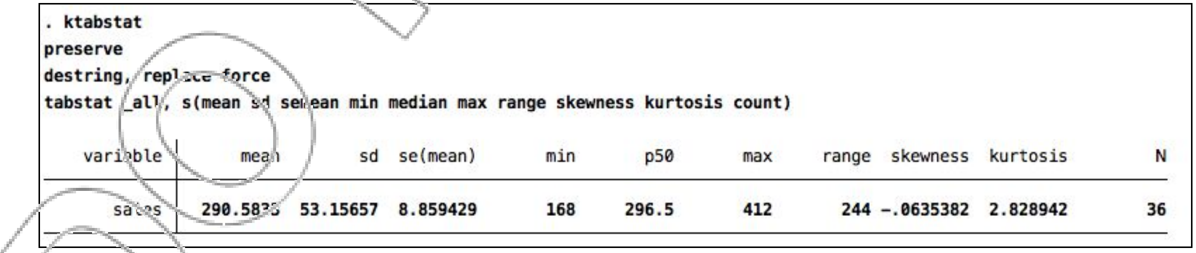
We will not say much about the type II error in this book, but you should know a few things about it. First, once the level of significance is set, the probability of making a type II error decreases as the sample size of your data increases. Therefore, the main tool in fighting against type II error is gathering more data. Second, the maximum probability of making a type II error is often denoted by the Greek letter beta (β) and 1-β is often called the **power** of a hypothesis test. So, if a test is said to be powerful, that means that the probability of a type II error is low. Conversely, a test that lacks power is one that may quite often fail to reject the null (i.e., be inconclusive) when the null is false. Again, increasing the sample size will make any test more powerful.

Now that you have learned the logic and terminology behind hypothesis testing, we turn to some examples to see how this works in practice. It may be helpful to refer back to this section if you find yourself getting confused at any point about what hypothesis tests are doing.

2.2 Test Marketing

Your company produces personal computers and is considering the introduction of new color options for the hardware in the hopes of boosting sales. Maintaining production of more than one color of computer is costly. For introducing new colors to be profitable, the company has set a sales goal of 275 units per week. The marketing department introduced and advertised the new colors in a test marketing experiment over 36 weeks. The weekly sales are given in the file **testmarket**. Based on the sales in the test market, should the company adopt the new color options?

To answer this question let us take a look at the descriptive statistics for the sample data. Loading **testmarket.dta** into Stata and then clicking **User>Core Statistics>Univariate Statistics>Standard (ktabstat)** results in the output in Figure 2.1.



```
. ktabstat
preserve
destring, replace force
tabstat _all, s(mean sd semean min median max range skewness kurtosis count)

    variable |      mean        sd  se(mean)       min       p50       max     range  skewness  kurtosis         N
-------------+--------------------------------------------------------------------------------------------------
       sales |  290.5833  53.15657  8.859429       168     296.5       412       244 -.0635382  2.828942        36
----------------------------------------------------------------------------------------------------------------
```

Figure 2.1: Univariate statistics for sales.

The sample mean of weekly sales $\bar{x}$ = 290.58, the sample standard deviation of weekly sales is s = 53.157, and the estimated standard deviation of the sample mean (called the standard error of the mean) equals $s_{\bar{x}}$ = 8.8594. We are going to need these numbers later.

We can rephrase the posed question: Do the sales in the test market indicate that the average sales per week will exceed 275 units? We are going to answer this question using hypothesis testing.

As a first step, determine the null hypothesis and the alternative hypothesis. To formulate the two hypotheses, focus on what you want to prove. The statement you want to prove should always appear as the alternative hypothesis. The way this hypothesis is established is by rejecting another hypothesis, namely the null hypothesis. Therefore, the null hypothesis is the statement you want to reject. Recalling the courtroom analogy, you prove that someone is guilty by showing that innocence can be rejected.

In our example, suppose we want to convince the management that the sales in the test market justify the introduction of the new colors. That is, we want to argue the average weekly sales if we go ahead with the color options will exceed 275 units. We define the alternative hypothesis as follows:

H_a: Average sales per week will exceed 275 units.

The opposite of the alternative hypothesis yields the null hypothesis.

H_0: Average sales per week will be less than or equal to 275 units.

Denote the average sales per week by μ. We can rewrite the hypotheses in formal terms:

$$H_0: \mu \leq 275$$

$$H_a: \mu > 275$$

The hypotheses concern the population average weekly sales, μ, rather than the sample average sales, $\bar{x}$, because μ determines sales going forward. If all we desired were to prove something about the sample average from the test market, there would be no need for hypothesis testing – the sample average is known and may be directly compared with 275. Hypotheses will always be about an unknown value or values.

The second step of hypothesis testing relates the sample data to the hypotheses. After all, we want to use the sample data to reject the null hypothesis. When would we do that? If the average sales of the new color PCs in the test market were much higher than 275, we would start to doubt that the null hypothesis is correct. On the other hand, if the average weekly sales were barely above or maybe below 275 units, we would not question the null hypothesis. By how much must sales exceed 275 units for us to reject the null hypothesis? To answer this question, we tentatively assume the null hypothesis is true with μ = 275. This value for μ will be the most difficult to reject of any in the null hypothesis since it is closest to the values in the alternative hypothesis. If we can reject this assumption, we can reject the null hypothesis.

We want to evaluate how far away the observed weekly sales in the test market are from the target value of 275. To make a probability statement, it is convenient to measure this difference in units of estimated standard deviations of $\bar{x}$:

$$\boxed{t = (\bar{x} - 275)/ s_{\bar{x}}}$$

This value measures the number of estimated standard deviations the sample mean, $\bar{x}$, is from the assumed mean, 275. This measure is called a **test statistic**. In our example, the test statistic takes on the following value:

$$\boxed{t = (290.58 – 275)/8.8594 = 1.7586}$$

The expression for the test statistic should look familiar to you. In the previous chapter,

$$\frac{\bar{x} - \mu}{s_{\bar{x}}}$$

had a t-distribution with n-1 degrees of freedom, where n is the sample size. We are tentatively assuming $\mu = 275$ and have a sample size of 36. So, in our example, the test statistic has a t-distribution with n-1 = 35 degrees of freedom. This fact is the reason we used t to denote the test statistic above.

The third step of hypothesis testing uses the test statistic to find the p-value. Assuming that the null hypothesis is true, the p-value is the probability of obtaining a sample result that is as least as unlikely as the one we have observed. In the context of our example, the p-value is the probability of obtaining a sample mean of $\bar{x}$ = 290.58 or higher assuming the true mean is $\mu = 275$. This probability is the area above 1.7586 in a t-distribution with 35 degrees of freedom as shown in Figure 2.2.

We can determine the p-value using the Stata **ttail** command. The probability of obtaining a sample mean of 290.58 or higher if μ = 275 equals ttail(35, 1.7586) = 0.0437. Therefore, the p-value equals 0.0437.

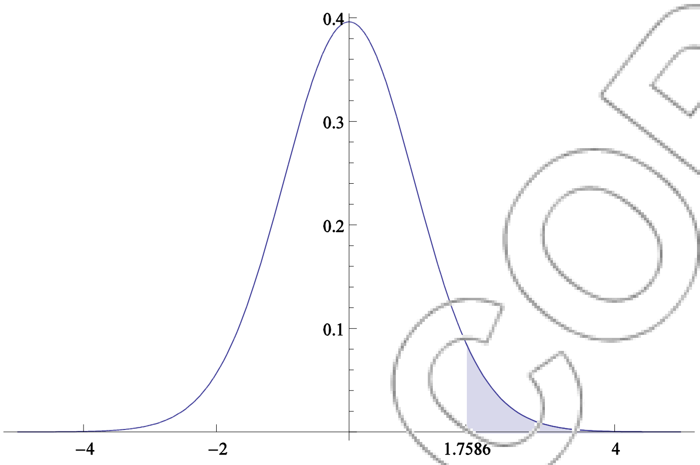


Figure 2.2: t-distribution with 35 degrees of freedom.

When the p-value is small, it is unlikely the sample results came from a population where the null hypothesis is true. The smaller the p-value, the stronger the evidence in favor of the alternative hypothesis.

The fourth step of hypothesis testing compares the calculated p-value to the level of significance (α, the maximum allowable probability of a type I error) that you have previously determined is appropriate for this test. In statistics, we can never be 100% sure when we make a conclusion based on sample data. Therefore, we have to decide on the probability with which it is acceptable to make an error.

The value for α will usually be given. So, choosing a value for α is not an issue, in particular when you perform a hypothesis test for someone else's use. Often, industry-specific standards and product-specific standards exist for α. In general, the costlier it is to claim that you have proved your claim when it is wrong, the smaller the α you should choose. Typical levels of α seen in practice will be between 0.01 and 0.1. For purposes of this text, if you need to specify α and have not been given any information to the contrary, you may assume $\alpha = 0.05$. However, the level of the p-value has its own meaning even if α is unspecified. Typically, a p-value will be clearly high or low; p-values over 0.3 would typically be considered high (and thus weak evidence for the alternative) in any application, and p-values less than 0.05 would typically be considered low (and thus strong evidence for the alternative). In between, judgment is needed.

The introduction of the color options entails much risk. If sales turn out to be mediocre, your company might face significant losses. Therefore, company policy is to be conservative in the evaluation of test data. Typically, the marketing department uses a level of significance of 5%, that is $\alpha = 0.05$.

The final step of hypothesis testing reaches a conclusion about the null hypothesis. The straightforward decision rule is this: If the p-value is smaller than or equal to the specified level of significance α, then we can reject the null hypothesis. If the p-value is larger than α, then we cannot reject the null hypothesis.

The p-value of 0.0437 is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis. Based on the sales in the test market, we are convinced that the average weekly sales will exceed 275 units. Your company should introduce the new color PCs, and the procedure of the hypothesis test is complete.

Suppose, based on new information about costs, you find the sales for the new colors must exceed 285 units per week to be profitable. What would your recommendation be in that case?

The new hypotheses are the following:

H_0: $\mu \leq 285$

H_a: $\mu > 285$

The value of the test statistic for this new scenario equals:

$t = (290.58 – 285)/8.8594 = 0.6298$

.

The resulting p-value is ttail(35, 0.6298) = 0.2665. We cannot reject the null hypothesis because the p-value is larger than α. Your company should not introduce the new colors yet. (A good strategy might be to collect more data on the test market, which might enable us to get a better idea about the potential sales of the new color PCs.)

What would your conclusion be if sales must exceed 300 units per week for the colors to be successful? The sample mean, $\bar{x}$ = 290.58, is smaller than 300. So, obviously you cannot conclude sales are going to exceed 300 units. In such a case, we do not need to perform a hypothesis test. It is clear that there is insufficient evidence to prove sales will exceed 300 units.

Before the marketing department started its test market campaign, it did extensive market research on the sales potential of the new colors. The research effort led to the projection that average weekly sales of the new color PCs would be 280 units. What do you think about the accuracy of this estimate now that you have sales data available from the test market?

We have some doubts about marketing department's claim and will try to prove them wrong. The alternative hypothesis states that average weekly sales are not equal to 280. The opposite, namely that average weekly sales equal 280, is the null hypothesis. More formally, we define the hypotheses as follows:

$$H_0\text{: } \mu = 280$$

$$H_a\text{: } \mu \neq 280$$

The test statistic equals the following:

$$t = (290.58 - 280)/8.8594 = 1.1942$$

We are going to doubt the null hypothesis if the sample mean significantly deviates from the value of 280, i.e., when the sample mean is considerably smaller or considerably larger than the prediction of the marketing department. The p-value for this test equals the sum of two probabilities, namely the sum of the probability of a deviation by at least 1.1942 standard deviations above the assumed mean and of the probability of a deviation by at least 1.1942 standard deviations below the assumed mean. This value is given by the shaded area in Figure 2.3.

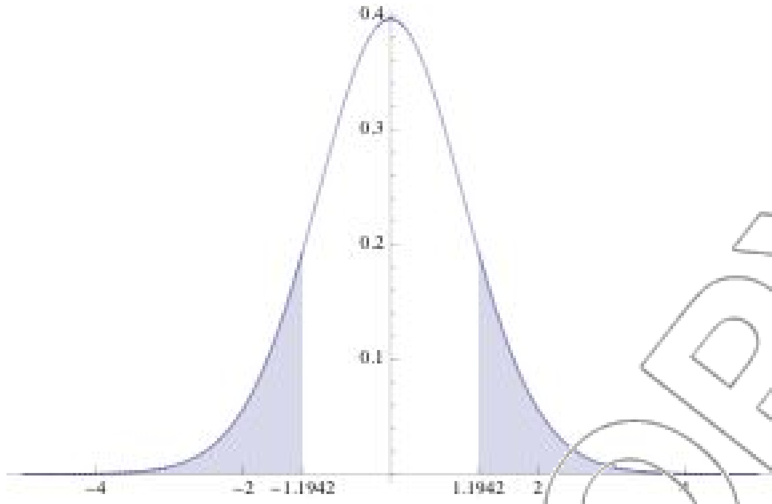


Figure 2.3: t-distribution and p-value for two-tailed test.

We can compute this p-value using the **ttail** command:

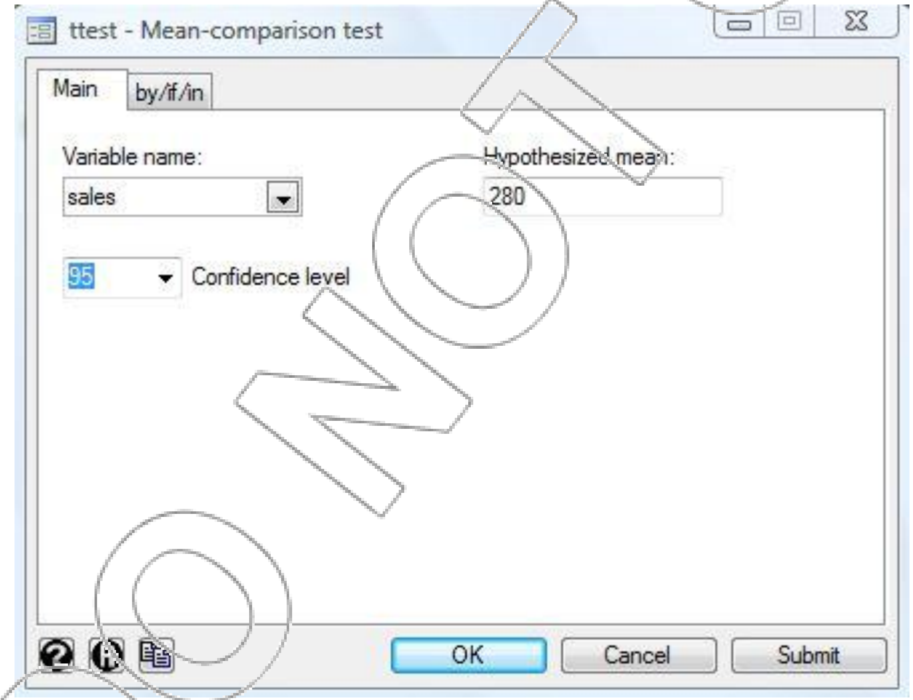
p-value = 2*ttail(35,1.1942) = 0.2404

Using the significance level α = 0.05, we conclude we cannot reject the null hypothesis that average monthly sales per district will equal 280 units. Therefore, we cannot claim on the basis of the test market data that the marketing department's forecast was wrong.

This last test differs from the previous ones since the null hypothesis is not an inequality but an equation. Tests of this form are called **two-tailed** hypothesis tests. Whenever the null hypothesis is an inequality, the hypothesis test is called **one-tailed**. The null hypothesis of a one-tailed test always contains the borderline case, that is, it contains a ≤ or a ≥ sign. The strict inequality sign (> or <) always appears in the alternative hypothesis.

The test statistics for one-tailed tests and two-tailed tests have the identical form. The main difference in the analysis is in the calculation of the p-value. For a one-tailed test, you can simply use the **ttail(n-1, t)** command (or 1-ttail(n-1, t) if you are calculating the area to the left of a test statistic). For a two-tailed test, you need to multiply the ttail value by 2 and use the absolute value of the test statistic, that is, 2*ttail(n-1, |t|), because the p-value includes the area in both the upper and lower tails of the distribution.

We can conduct a one-tailed or two-tailed hypothesis test much more quickly using Stata's **ttest** command. Consider our previous example, where we want to test the marketing department's claim that average weekly sales are equal to 280. To do this in Stata, click **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>One-sample mean-comparison test**.[1] This will open the following dialog box:



Variable name: sales

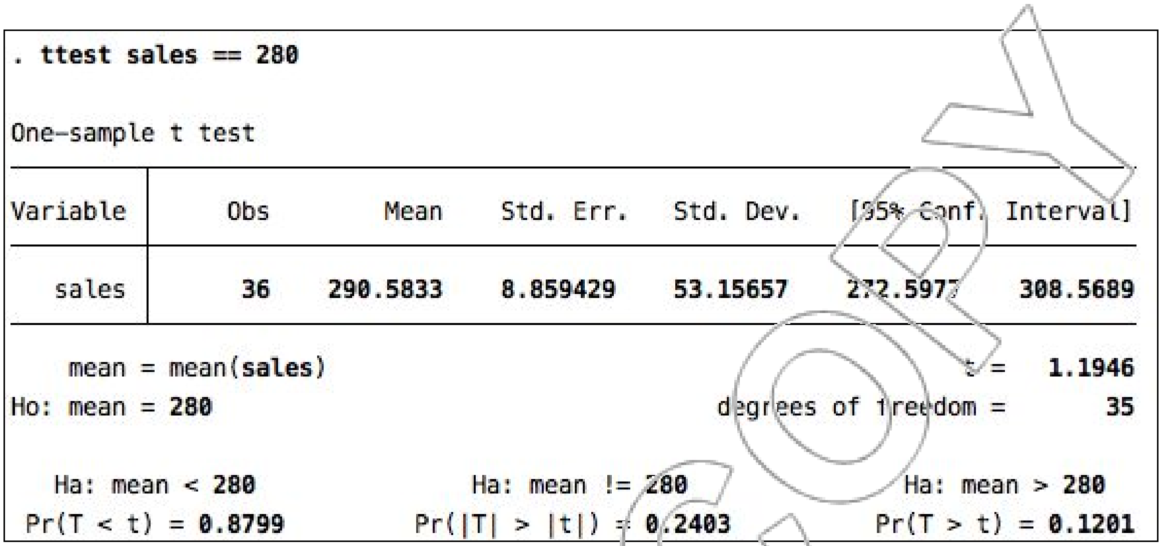
Hypothesized mean: 280

95 Confidence level

Choose **sales** from the "Variable name" list and enter **280** in the "Hypothesized mean" field. The default confidence level is 95%, and you can change it if you want, although it does not affect the

[1] Alternatively, you can directly type the command **ttest sales == 280**.

hypothesis test calculations that Stata does at all and simply determines which confidence interval for the mean Stata reports. Click **OK**, and Stata will return the following:



```
. ttest sales == 280

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   sales |      36    290.5833    8.859429    53.15657    272.5977    308.5689
------------------------------------------------------------------------------
    mean = mean(sales)                                            t =   1.1946
Ho: mean = 280                                   degrees of freedom =       35

   Ha: mean < 280              Ha: mean != 280               Ha: mean > 280
 Pr(T < t) = 0.8799        Pr(|T| > |t|) = 0.2403        Pr(T > t) = 0.1201
```

As you can see, Stata displays the sample mean (Mean), the standard error of the mean (Std. Err.), and the degrees of freedom from which you can manually calculate the test statistic and the appropriate p-value. However, Stata has already done this work for you. The test statistic is listed on the right-hand side of the output, where t = 1.1946 (the slight difference from our calculation is due to rounding). At the bottom of the output, Stata lists the respective p-values for all possible alternative hypotheses of interest (i.e., H_a: $\mu < 280$, H_a: $\mu \neq 280$, and H_a: $\mu > 280$). Since, in this example, we are interested in the alternative hypothesis that average weekly sales are not equal to 280, we look to the middle column and find the p-value to be Pr(|T| > |t|) = 0.2403, which agrees with our manual calculation (up to rounding).

2.3 Hypothesis Testing: A Formal Analysis

Now let us see what goes on behind hypothesis testing, review the mechanical calculations, and see why they really work.

The first formal step in hypothesis testing is writing down the two hypotheses. For example, in the last test marketing example the hypotheses that we developed were the following.

$$H_0: \mu = 280$$

$$H_a: \mu \neq 280$$

Hypothesis tests are always stated in terms of the true parameters we are interested in and not in terms of the estimators. Here, the parameter we are interested in is μ, the true average sales.

The estimate we derived for the average sales was $\bar{x} = 290.58$.

To evaluate the evidence in our data, we will initially assume the null hypothesis is correct. We then see if our observed result is likely or unlikely given the null. If it is likely, then it is not strong evidence in favor of the alternative, and we cannot reject the null. Conversely, if it is unlikely (less likely than the level of significance that we have set up in advance), we will reject the null hypothesis.

The null hypothesis determines the sampling distribution of our estimator, $\bar{x}$. What is this distribution? First, we make an assumption that this distribution is a normal distribution. (If our sample is large, this assumption is justified by the central limit theorem.) Any normal distribution has a mean and a standard deviation. The mean is the one given by the null hypothesis, e.g., 280. As you learned in the first chapter, the standard deviation of $\bar{x}$, which we will denote by $\sigma_{\bar{x}}$, is given by $\sigma / \sqrt{n}$. Since we do not know σ, we must use the sample standard deviation, s, to estimate it. Therefore, the estimated standard deviation of $\bar{x}$ (which we will denote by $s_{\bar{x}}$, sometimes called the standard error of the mean) is given by $s/\sqrt{n}$.

To evaluate the strength of our evidence, we want to see how far away our observed estimator is from the value we would expect if the null hypothesis were true. To do this, we look at the quantity: estimator minus the value given in the null hypothesis. Since we would like to use this difference to make a probability statement, it is convenient to convert it into a number of standard deviations by dividing by the standard deviation of our estimator. Therefore, our test statistic will have the following form:

$$\boxed{\text{test statistic} = \frac{\text{estimator - value given in the null hypothesis}}{\text{standard deviation of the estimator}}}$$

This test statistic has the following interpretation: Our estimate is (insert value of test statistic) standard deviations away from the value given in the null hypothesis. In our example, our estimator is $\bar{x}$ = 290.58, the value in the null hypothesis is 280, and the standard deviation of the estimator is $\sigma_{\bar{x}}$. Since we are using $s_{\bar{x}}$ (= 8.8594) to estimate $\sigma_{\bar{x}}$, our test statistic will have a t-distribution instead of a standard normal (or z) distribution. Finally, the degrees of freedom for this t-distribution is n-1, where n is the sample size.

In our example, the test statistic (often written t since it has a t-distribution) is t = (290.58-280)/8.8594 = 1.1942, which means that our estimator $\bar{x}$ is 1.1942 standard deviations above the value in the null hypothesis. We saw earlier that the corresponding p-value = 0.2404, which means that if the null hypothesis were true, there is about a 24% chance of getting a value of our estimator as far away as 1.1942 standard deviations (or further).

ONE-TAILED TESTS

The example above was a two-tailed test because the alternative hypothesis included values both above and below the value in the null. In general, if the null hypothesis is an equality, then the test is a two-tailed test. In other examples, we may want to prove that a parameter is above a certain value or prove that it is below a certain value instead of showing it is simply different from a certain value. This requires a one-tailed test. Such a test is called one-tailed because the values in the alternative hypothesis are all on one side of the values in the null hypothesis. For example, if we want to prove that average sales are greater than 275, we would use the following hypotheses:

$$H_0: \mu \leq 275$$

$$H_a: \mu > 275$$

Notice two things here. First, the "equals" value appears in the null hypothesis as, by convention, it always will. Second, when forming our test statistic we have to know what number to plug in for the value in the null hypothesis. The rule is we always use the equals value. In this example, the value of the test statistic is $t = (290.58\text{-}275)/8.8594 = 1.7586$. We used the equals value of 275 for the value in the null hypothesis. Since our alternative hypothesis has a greater than (>) sign, only positive values of the test statistic will provide evidence against the null hypothesis. Thus, the one tail we care about when calculating the p-value in this example is the upper tail or the one with positive values. This p-value is the area above 1.7586 in a t-distribution with 35 (= n-1) degrees of freedom. As you saw in the test marketing example, we can find this area using Stata's **ttail** command as follows:

p-value = ttail(35,1.7586) = 0.0437

Similarly, if we wanted to prove that average sales were less than 275, we would use these hypotheses:

H_0: $\mu \geq 275$

H_a: $\mu < 275$

Here, the test statistic is again t = (290.58 - 275)/8.8594 = 1.7586 (the same as above!) Since the alternative hypothesis has a less than sign, however, only negative values of the test statistic will provide evidence against the null hypothesis. Therefore, when calculating the p-value, the one tail we care about is the lower tail, or the one with negative values. So, the corresponding p-value is the one which gives the area below 1.7585 in a t-distribution with 35 (= n-1) degrees of freedom. Since the **ttail** command always gives the area above a given number, we can find the area below 1.7586 by using p-value = 1- ttail(35, 1.7586) = 0.9563. The p-value came out large, indicating weak evidence against the null (or in favor of the alternative). We could have seen this without any calculation. Whenever you do a one-tailed test and the estimated value is on the wrong side of the equals value in the null (i.e., above the null value if the alternative looks at the lower tail or below the null value if the alternative looks at the upper tail), you automatically know the p-value is larger than 0.5. Since this is higher than any level of significance you would ever want to use, you know you cannot reject the null (or accept the alternative) using these data. In such a case, calculating the test statistic and exact p-value is not necessary.

Suppose we want to show that average sales are below 310. The appropriate hypotheses are the following:

H_0: $\mu \geq 310$

H_a: $\mu < 310$

The test statistic is t = (290.58 - 310)/8.8594 = -2.192. The correct p-value is the area to the left of -2.192 in a t-distribution with 35 degrees of freedom. Using Stata to calculate the p-value for this example, you can either type **display 1-ttail(35, -2.192)** or use the symmetry of the t-distribution and type **display ttail(35, 2.192)**. It may help you to draw a picture (see Figure 2.4) to understand why these areas are the same. In either case the answer is p-value = 0.0176.

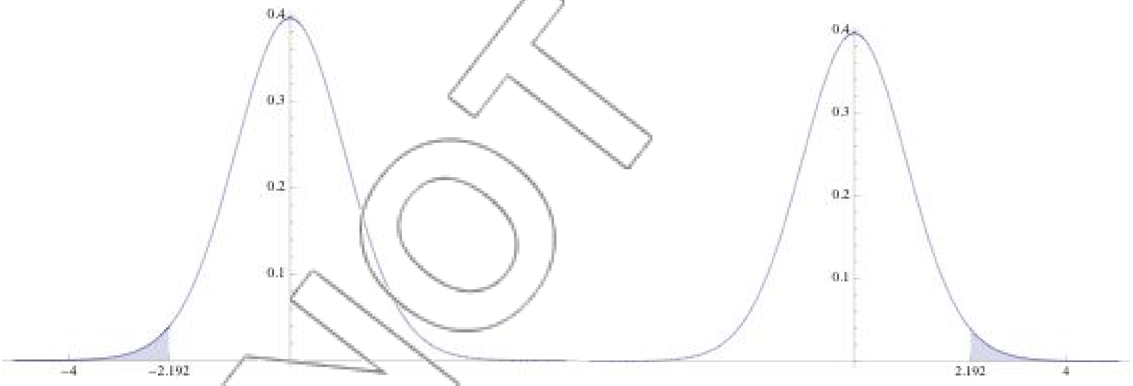


Figure 2.4: Symmetry of t-distribution.

MECHANICS OF TESTS CONCERNING A POPULATION MEAN

Step 1: Choose the appropriate hypothesis test:

One-tailed tests		Two-tailed test
H_0: $\mu \geq \mu_0$	H_0: $\mu \leq \mu_0$	H_0: $\mu = \mu_0$
H_a: $\mu < \mu_0$	H_a: $\mu > \mu_0$	H_a: $\mu \neq \mu_0$

Step 2: Calculate the test statistic:

We have the same test statistic whether we face a one-tailed test or a two-tailed test.

The test statistic is computed using the following formula:

$$\boxed{t = \frac{\text{estimator} - \text{value in the null hypothesis}}{\text{standard deviation of the estimator}} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}}$$

It has a t-distribution with n-1 degrees of freedom.[2]

Step 3: Calculate the p-value:

One-tailed test, less than sign in alternative: p-value = 1 - ttail(n - 1, test statistic).

One-tailed test, greater than sign in alternative: p-value = ttail(n - 1, test statistic).

Two-tailed test: p-value = 2*ttail(n - 1, $|\text{test statistic}|$).

$|\text{test statistic}|$ means the absolute value of the test statistic. That is, it is equal to the test statistic if the test statistic is positive, and it is equal to -test statistic if the test statistic is negative.

Step 4: Final decision:

Suppose our designated level of significance is α (e.g. 0.05 = 5%).

[2] Rarely, you may be given a value for σ, the population standard deviation. In this case, use $\sigma_{\bar{x}}$ in place of $s_{\bar{x}}$, and use the standard normal (z) distribution in place of t.

If p-value $\leq \alpha$, we reject the null hypothesis (and accept the alternative hypothesis).

If p-value > α, we cannot reject the null hypothesis (and cannot accept the alternative).

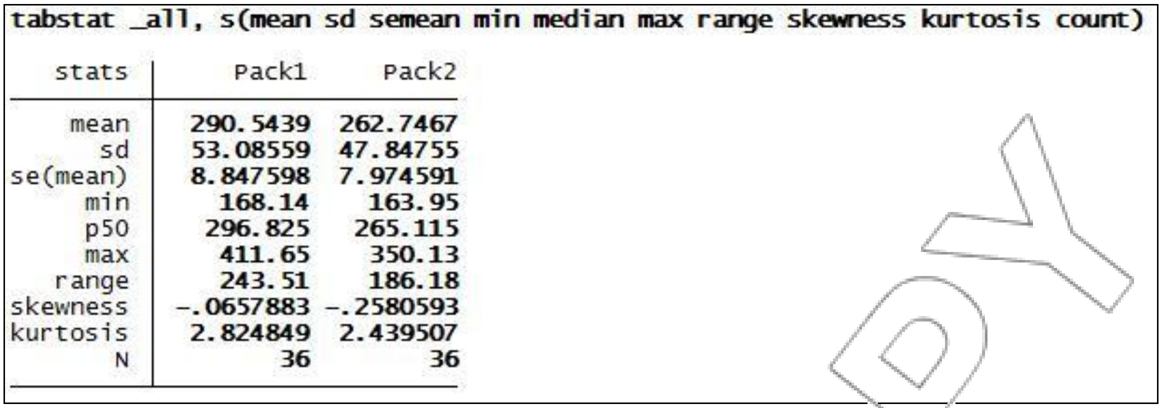
TESTS CONCERNING THE POPULATION PROPORTION

Just as we have done hypothesis tests where the parameter is the population mean, we can do tests about the population proportion. We form the test statistic in the same way as above. However, in this case, since our estimator is the sample proportion, $\bar{p}$, instead of the sample mean, we need a different formula for the standard deviation of the estimator. We will not make use of tests concerning proportions until the next section on two population problems.

2.4 Consumer Packaging

The marketing department at a large consumer products firm is considering changing the packaging of one of its primary sales items. Two alternatives are being considered. To assess the relative strengths of these two alternatives, the marketing research department is directed to test which package sells better. Accordingly, a collection of 72 sales districts (similar in terms of demographic characteristics) is selected; 36 are assigned for testing package 1, and the other 36 are used to test package 2. Sales figures for a one-month test period are collected (in the file **package**). The variables pack1 and pack2 contain the observations on sales for the districts assigned to packages 1 and 2, respectively. Each variable has 36 observations. First, we will look at the descriptive statistics.

User>Core Statistics>Univariate Statistics>Standard (ktabstat)



```
tabstat _all, s(mean sd semean min median max range skewness kurtosis count)

   stats |     Pack1      Pack2
---------+--------------------
    mean |  290.5439   262.7467
      sd |  53.08559   47.84755
se(mean) |  8.847598   7.974591
     min |    168.14     163.95
     p50 |   296.825    265.115
     max |    411.65     350.13
   range |    243.51     186.18
skewness | -.0657883  -.2580593
kurtosis |  2.824849   2.439507
       N |        36         36
------------------------------
```

Figure 2.5: Univariate statistics for pack1 and pack2

Now think conceptually for a moment. What are our two populations here? One is any store where the product is sold in package 1, now or in the future, and the other is stores where it is sold in package 2, now or in the future. The variable of interest for each population is sales, and specifically we want to compare average monthly sales from the two populations, i.e., average monthly sales if we adopt package 1, to average monthly sales if we adopt package 2. Call these numbers μ_1 and μ_2, respectively. The first 36 districts in our experiment give us a sample from population 1, and the next 36 districts give us a sample from population 2. We can use the sample from each population to estimate its population parameters. Mean sales from the first 36 stores (written $\bar{x}_1$ = 290.54) give our estimate of μ_1, and, using the other 36 stores, $\bar{x}_2$ = 262.75 is our estimate of μ_2.

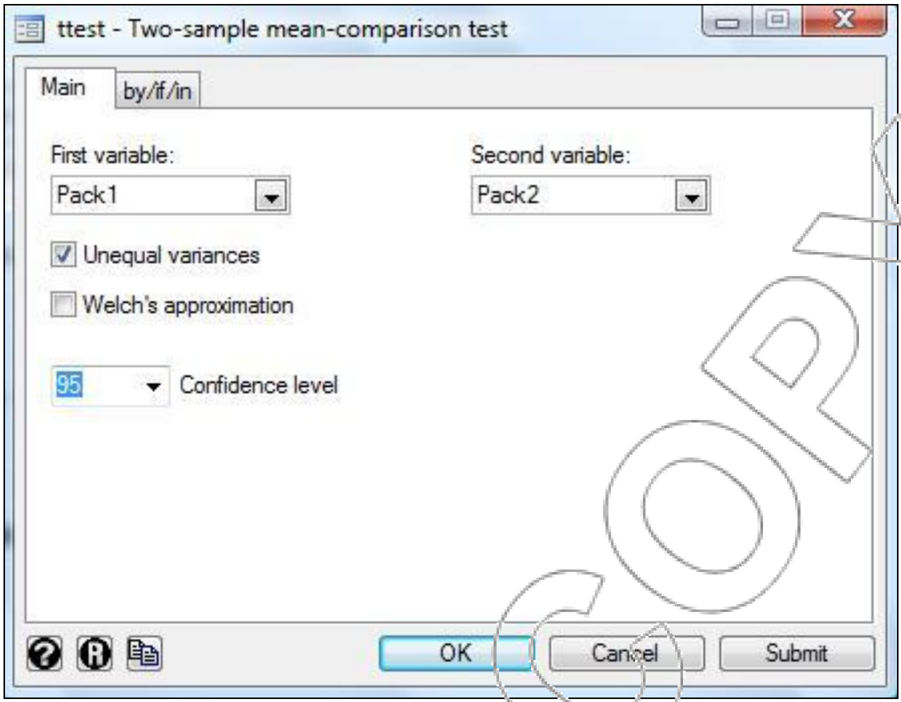
Obviously, our estimates suggest that sales will be higher on average with package 1 since we can estimate the difference μ_1-μ_2 by $\bar{x}_1 - \bar{x}_2 = 27.79$. So, if you had to make the choice right now between the two packages, the rational decision (assuming that the packages cost the same to produce, etc.) would be to go with package 1. However, you have other options. You could choose to continue or expand the marketing experiment, postponing your final decision until you have more data. So, it is worth asking how confident you are that package 1 is the better of the two. After all, a month is not a long time, and 36 stores might not be a big enough sample. In other words, it might be that package 1 is inferior, and unfortunately, you hit an atypical sample. Hypothesis testing can help by telling you how strong the evidence you have is for a particular proposition. In this case, since you have the option of continuing the experiment, you want to be fairly certain of the superiority of package 1 before concluding that it is the better one. You make the alternative hypothesis the statement that packaging 1 is better in terms of average monthly sales. (Recall that the alternative hypothesis is the one you want to prove – here you want to see if the data convincingly show that package 1 is better).

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_a: \mu_1 - \mu_2 > 0$$

How do we perform this test? For the purposes of this example, we will use Stata's **ttest** command to do it. (You can see it done "by hand" in the next section, which explains the statistical theory of two-sample tests.)

After loading **package.dta** into Stata, click **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample mean-comparison test** to open the **ttest** dialog box. Select **Pack1** and **Pack 2** from the "First variable" and "Second variable" lists, respectively. Check the box next to "Unequal variances." Your dialog box should look like this:



ttest - Two-sample mean-comparison test

First variable: Pack1

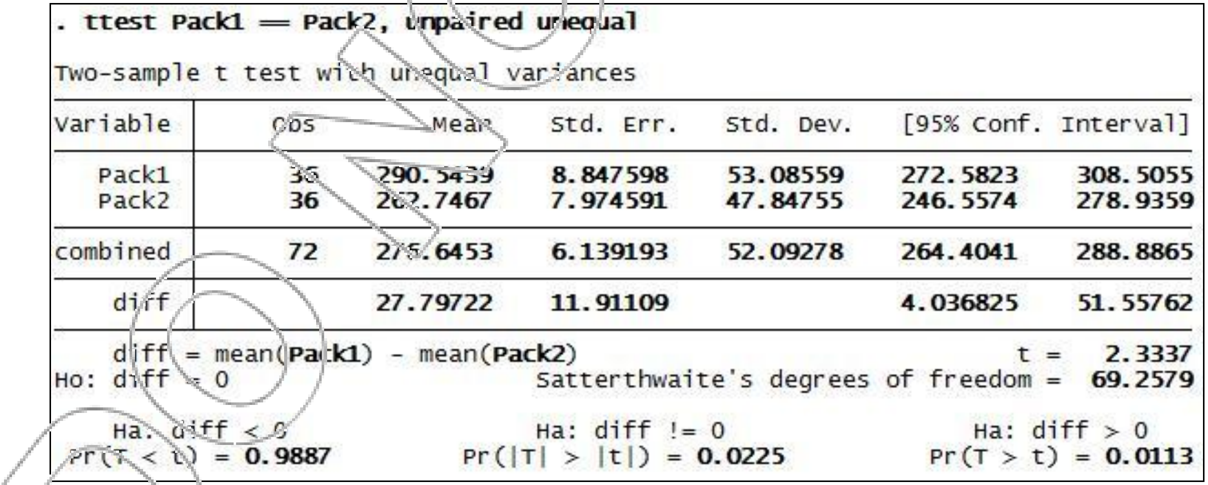
Second variable: Pack2

☑ Unequal variances

□ Welch's approximation

95 Confidence level

The analogous command is **ttest Pack1 == Pack2, unpaired unequal**.[3] Execute the command, and Stata will return the following:



```
. ttest Pack1 == Pack2, unpaired unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   Pack1 |      36    290.5439    8.847598    53.08559    272.5823    308.5055
   Pack2 |      36    262.7467    7.974591    47.84755    246.5574    278.9359
---------+--------------------------------------------------------------------
combined |      72    276.6453    6.139193    52.09278    264.4041    288.8865
---------+--------------------------------------------------------------------
    diff |            27.79722    11.91109                4.036825    51.55762
------------------------------------------------------------------------------
    diff = mean(Pack1) - mean(Pack2)                              t =   2.3337
Ho: diff = 0                     Satterthwaite's degrees of freedom =  69.2579

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9887         Pr(|T| > |t|) = 0.0225          Pr(T > t) = 0.0113
```

[3] Typing "Pack1 == Pack2" tells Stata that we are testing equality of means between the variables Pack1 and Pack2. "Unpaired" indicates that we are not assuming any special meaning to the order of the observations. In particular, the kth observation of pack1 is not assumed to be any more or less related to the kth observation of pack2 than to any other observation of pack2. Finally, we type in "unequal" since we do not assume equal variances for the two populations.

Since our alternative hypothesis is H$_a$: $\mu_1 - \mu_2 > 0$, we will refer to the rightmost alternative hypothesis. Stata gives us the p-value (p = 0.0113) associated with this one-tailed test. It tells us that if package 1 is no better than package 2 (i.e., if the null hypothesis is true), there is at most a probability of .0113 of seeing as big a difference favoring package 1 in the sample averages as we have obtained. Thus, we may be highly confident that package 1 is better than package 2. For any significance level, α, above 1.13%, we can say that package 1 has (statistically) significantly greater average sales than package 2.

A final important point here is that you should distinguish between statistical significance and economic significance. That the difference in average sales across the two kinds of packaging is statistically significant means we have strong evidence of a difference. It does not tell us how important that difference is, i.e., whether it is economically significant. In this case, the estimated difference does seem economically significant: Going from package 2 to package 1 is estimated to increase sales on average by (290.54-262.75)/262.75 = 10.58 percent. However, think about the following scenario: Imagine you must choose between two alternative packages and suppose that you are currently using package 1, so you will incur some costs if you switch to package 2. Suppose further you conduct a marketing experiment as above (but with a larger sample size), and find that sales with package 2 are higher by an estimated 0.3%, and this difference is statistically significant. In that case, you would likely choose not to change over (at least for the time being) because the estimated difference, though statistically significant, may not be economically significant. It may be too small to justify incurring the costs of switching over.

2.5 Two Populations

This section expands on the example above and explains the statistical techniques used to compare two populations. This material follows from what you learned about one-population testing though the formulas may look a little more complicated. Consider the following: We have a sample from population 1, giving a sample mean of $\bar{x}_1$, and a sample from population 2, giving a sample mean of $\bar{x}_2$. We will assume both samples are not too small (say n_1 and n_2 are at least 30). For small samples, some extra issues arise (see the note at the end of this section). If population 1 has a mean of μ_1 and a standard deviation of σ_1 and population 2 has a mean of μ_2 and a standard deviation of σ_2, then the first sample mean, $\bar{x}_1$, is approximately normally distributed with a mean of μ_1 and a standard deviation of

$$\sigma_{\bar{x}_1} = \sigma_1 / \sqrt{n_1}\ .$$

The second sample mean, $\bar{x}_2$, is (approximately) normally distributed with a mean of μ_2 and a standard deviation of

$$\sigma_{\bar{x}_2} = \sigma_2 / \sqrt{n_2}\ .$$

Two properties of random variables are important to us here. If *X* and *Y* are independent random variables, the mean and variance of their difference, *X*-*Y*, are given by the following:

$$\boxed{\begin{gathered}\mu_{X-Y} = \mu_X - \mu_Y \\ \sigma^2{}_{X-Y} = \sigma^2{}_X + \sigma^2{}_Y\end{gathered}}$$

We apply these formulas to $\bar{x}_1$ and $\bar{x}_2$, giving the following:

$$\boxed{\begin{aligned}\mu_{\bar{x}_1-\bar{x}_2} &= \mu_1 - \mu_2 \\ \sigma^2_{\bar{x}_1-\bar{x}_2} &= \sigma^2_{\bar{x}_1} + \sigma^2_{\bar{x}_2} = \frac{\sigma_1^{\ 2}}{n_1} + \frac{\sigma_2^{\ 2}}{n_2}\end{aligned}}$$

So, $\bar{x}_1 - \bar{x}_2$ is (approximately) normally distributed with a mean of μ_1 - μ_2 and a standard deviation of the following:

$$\boxed{\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\left(\sigma_1^{\ 2}/n_1\right)+\left(\sigma_2^{\ 2}/n_2\right)}}$$

As in the case of one population, because σ_1 and σ_2 are unknown, we will need to estimate them using sample standard deviations s_1 and s_2 instead. Thus, we use

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\left(s_1^{\ 2}/n_1\right)+\left(s_2^{\ 2}/n_2\right)}$$ to estimate $\sigma_{\bar{x}_1-\bar{x}_2}$.

An approximate (1-α)(100)% confidence interval for $\mu_1 - \mu_2$ is given by the following:[4]

$$\boxed{\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2,n_1+n_2-2}\, s_{\bar{x}_1-\bar{x}_2}}$$

The test statistic for hypothesis tests concerning $\mu_1 - \mu_2$ is the following:

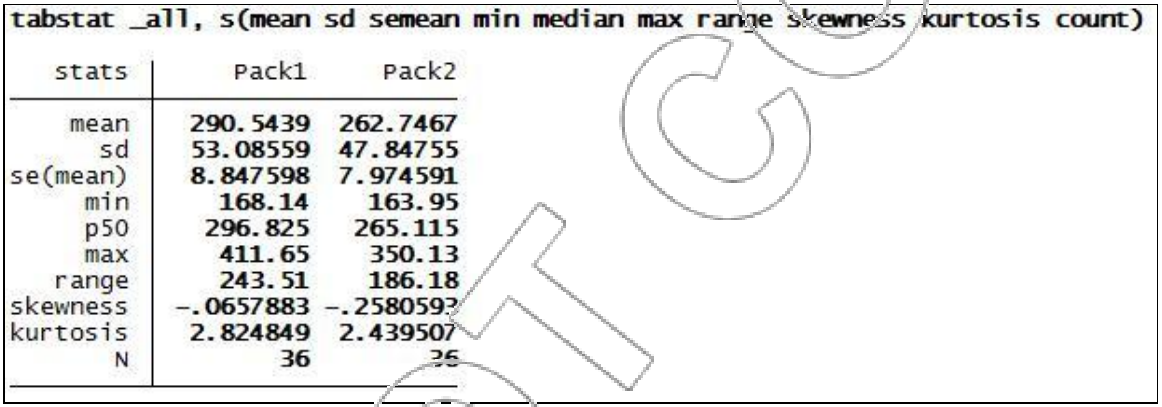
[4] The use of n_1+n_2-2 degrees of freedom for the t in the confidence interval formula is only strictly correct if the variances of the two samples are the same. If the variances differ, the approximate degrees of freedom to use is given by Satterthwaite's formula:

$$\text{[EQ]}\ df = \frac{\left(s_{\bar{x}_1-\bar{x}_2}\right)^4}{\frac{\left(s_{\bar{x}_1}\right)^4}{n_1-1}+\frac{\left(s_{\bar{x}_2}\right)^4}{n_2-1}}.$$

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_0}{s_{\bar{x}_1 - \bar{x}_2}}$$

The equals value in the null hypothesis tells us what to insert for $(\mu_1-\mu_2)_0$.

Recall the consumer packaging example of the previous section. The univariate statistics were the following:



```
tabstat _all, s(mean sd semean min median max range skewness kurtosis count)

    stats |     Pack1      Pack2
----------+--------------------
     mean |  290.5439   262.7467
       sd |  53.08559   47.84755
 se(mean) |  8.847598   7.974591
      min |    168.14     163.95
      p50 |   296.825    265.115
      max |    411.65     350.13
    range |    243.51     186.18
 skewness | -.0657883  -.2580593
 kurtosis |  2.824849   2.439507
        N |        36         36
-------------------------------
```

Figure 2.6: Univariate statistics for pack1 and pack2.

So, we have $\bar{x}_1$ = 290.54, s_1 = 53.086, $\bar{x}_2$ = 262.75, s_2 = 47.848. Our estimate for the difference in means μ_1-μ_2 is 290.54-262.75 = 27.79. We estimate the standard deviation of $\bar{x}_1 - \bar{x}_2$ by using the equation below:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left((53.086)^2/36\right) + \left((47.848)^2/36\right)} = 11.91$$

You can verify this value by checking the Stata ttest output from Section 2.4. Stata lists the standard deviation of $\bar{x}_1 - \bar{x}_2$ in the **Std. Err.** column and the **diff** row. Now we may, for example, construct an approximate 95% confidence interval for our point estimate. It is given by $\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2, n_1+n_2-2}\, s_{\bar{x}_1-\bar{x}_2}$ = 27.79±invttail(n_1+n_2-2,α/2)(11.91) = 27.79±(1.9944)(11.91) = (4.04, 51.54). We also can do the hypothesis test that we had Stata perform for us previously. The null and alternative hypotheses were as listed below:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_a: \mu_1 - \mu_2 > 0$$

The test statistic is equal to:

$$\frac{27.79 - 0}{11.91} = 2.333$$

Calculating the area above 2.333 in a t-distribution with 70 degrees of freedom gives a p-value of ttail(70, 2.333) = 0.01126. How does this compare with the computer output? Stata's **ttest** command gave us a p-value of 0.0113. There are two reasons for the slight discrepancy. One is our use of n_1+n_2-2 = 70 as the number of degrees of freedom for the t-distribution. As explained in the footnote to the formula for the confidence interval for $\mu_1 - \mu_2$, when the variances of the populations are not equal there is a more exact formula for degrees of freedom (called Satterthwaite's degrees of freedom in the Stata output). In this example, this formula gives approximately 69 rather than 70. The second reason is numerical round-off error, as we rounded the means and standard deviations to fewer decimal places than Stata did.

POPULATION PROPORTIONS

Analogous formulas for differences in population proportions can be summarized briefly as follows. We will again assume the samples are large. (In practice, estimating population proportions from small samples is unusual.) Given sample proportions $\overline{\mathrm{p}}_1$ and $\overline{\mathrm{p}}_2$, we estimate the standard deviation of their difference using the following:

$$s_{\overline{p}_1-\overline{p}_2} = \sqrt{\frac{\overline{p}_1\left(1-\overline{p}_1\right)}{n_1} + \frac{\overline{p}_2\left(1-\overline{p}_2\right)}{n_2}}$$

A (1-α)(100)% confidence interval for $p_1 - p_2$ is given by the following:

$$\overline{p}_1 - \overline{p}_2 \pm z_{\alpha/2}\, s_{\overline{p}_1-\overline{p}_2}$$

The test statistic for hypothesis tests concerning $p_1 - p_2$ is the following:

$$z = \frac{\overline{p}_1 - \overline{p}_2 - \left(p_1 - p_2\right)_0}{s_{\overline{p}_1-\overline{p}_2}}$$

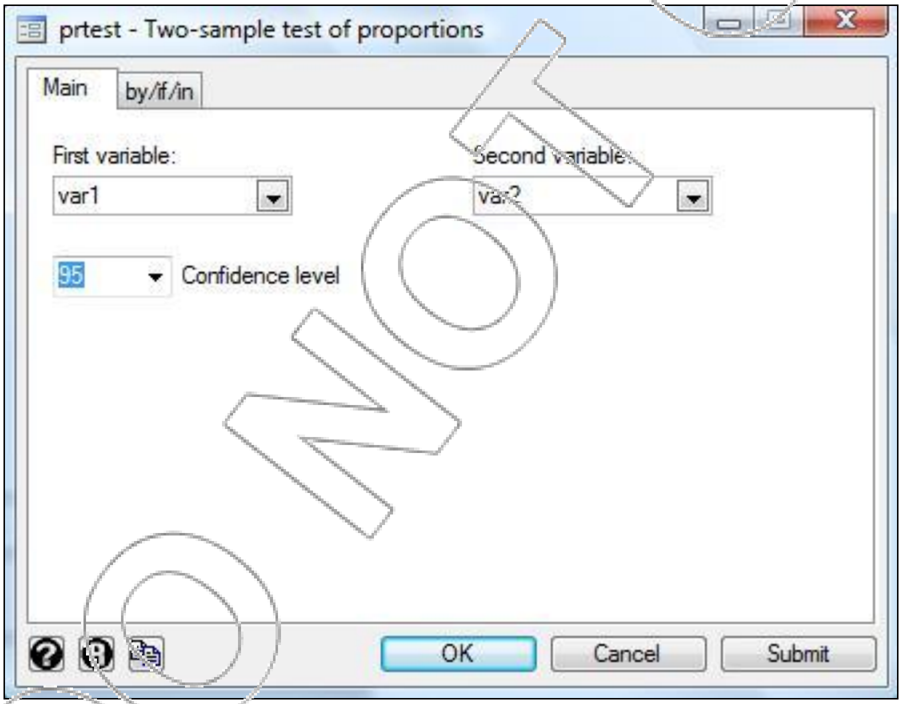
The comparison of proportions is the only type of hypothesis test or confidence interval for which we will use a standard normal (z) distribution rather than a t-distribution.

In Stata, you can conduct a one-sided or two-sided hypothesis test on the equality of proportions by using the **prtest** command. As an example, we will use the file **proportion**, which contains two binary variables, **var1** and **var2**, with 30 observations each. **Var1** has thirteen observations equal to 1, and **var2** has ten observations equal to 1. Therefore, $\bar{p}_1 = 13/30 = 0.433$, and $\bar{p}_2 = 10/30 = 0.333$. Suppose we want to conduct the following hypothesis test:

$$H_0: p_1 - p_2 \leq 0$$

$$H_a: p_1 - p_2 > 0$$

To do this in Stata, click **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample proportion test**. Select **var1** for the first variable and **var2** for the second variable, as shown in the following:

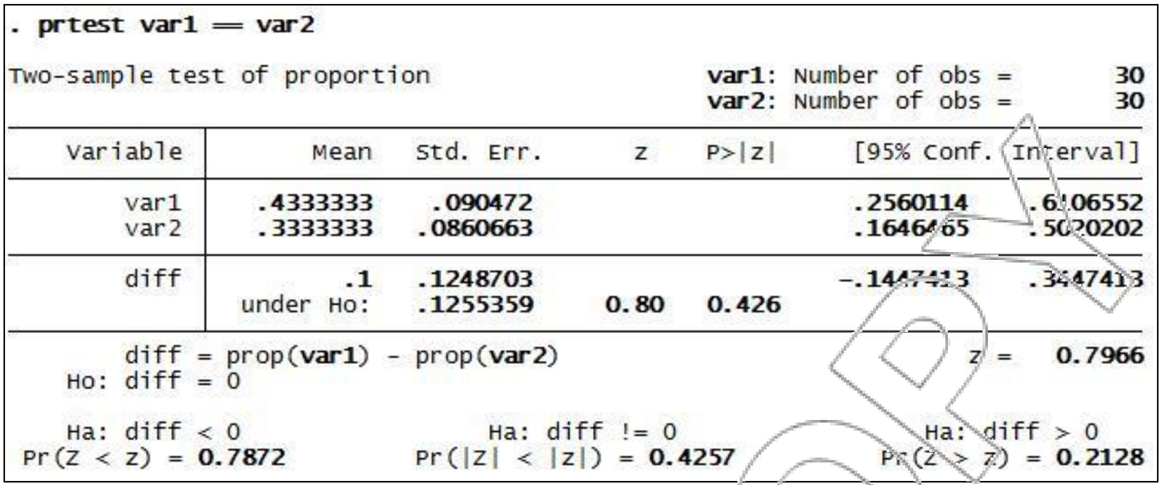


First variable: var1

Second variable: var2

95 Confidence level

The corresponding typed command is **prtest var1 == var2**. Executing the command will generate the following result:



```
. prtest var1 == var2

Two-sample test of proportion                   var1: Number of obs =       30
                                                var2: Number of obs =       30
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        var1 |   .4333333    .090472                      .2560114    .6106552
        var2 |   .3333333   .0860663                      .1646465    .5020202
-------------+----------------------------------------------------------------
        diff |         .1   .1248703                     -.1447413    .3447413
             |  under Ho:   .1255359     0.80   0.426
------------------------------------------------------------------------------
        diff = prop(var1) - prop(var2)                            z =   0.7966
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.7872         Pr(|Z| < |z|) = 0.4257          Pr(Z > z) = 0.2128
```

Given our alternative hypothesis, H$_a$: $p_1 - p_2 > 0$, we see that the p-value is Pr(Z > z) = 0.2128, or 21.28%. Therefore, we do not have strong enough evidence to show that $p_1 - p_2 > 0$ if we are using a significance level below 21.28%.

There are two things to note when using Stata's **prtest** command. First, the standard errors of the proportion for **var1** and **var2** can be found in the first two rows under the **Std. Err.** column (which are 0.0905 and 0.0861, respectively). The value for $\bar{p}_1 - \bar{p}_2$ is shown in the **Mean** column and the **diff** row (= 0.1). The value for $s_{\bar{p}_1 - \bar{p}_2}$ is shown in the **Std. Err.** column and the **diff** row (= 0.1249). Using these reported values, you can manually calculate the test statistic and the p-values. The 95% confidence interval for $p_1 - p_2$ is automatically calculated as (-0.145, 0.345).

Second, note that Stata reports an additional standard error in the **Std. Err.** column and the **under Ho:** row (= 0.1255). In fact, this is the value that Stata uses in place of $s_{\bar{p}_1 - \bar{p}_2}$ in calculating the test statistic and the p-values. This standard error is calculated using the following formula:

$$\text{Std. Err. under } H_0 = \sqrt{p_c * (1 - p_c) * (\tfrac{1}{n_1} + \tfrac{1}{n_2})}$$

Here, n_1 and n_2 denote the number of observations for **var1** and **var2**, respectively, And the pooled estimate of proportion, p_c, is calculated as:

$$\boxed{p_c = \frac{x_1 + x_2}{n_1 + n_2},}$$

where x_1 and x_2 denote the number of 1's in **var1** and **var2**, respectively. Stata uses this pooled estimator because if, in fact, the two proportions are equal, it is the best estimator of the common proportion. If you calculated the p-value for the alternative hypothesis H_a: $p_1 - p_2 > 0$ using the original standard error of the difference in proportions, $s_{\bar{p}_1-\bar{p}_2}$ = 0.1249, you would get a test statistic of z = $\frac{0.1-0}{0.1249}$ = 0.8006 and a corresponding p-value of 1-normal(1.2085) = 0.2117, which is slightly smaller than the p-value calculated by Stata's **prtest** command.

Note that to use Stata's **prtest** command, you need to have an actual dataset containing binary variables of interest. Sometimes you may only be given the respective sample sizes and sample proportions from two populations. In this case, you can still conduct a hypothesis test concerning two population proportions by using Stata's **prtesti** command. To do this, click **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample proportion calculator**. In the ensuing **prtesti** dialog box, enter the respective sample sizes and sample proportions for your two populations and specify a confidence level.[5] Click **OK**, and Stata will display an output very similar to the **prtest** output shown above. In the **diff** row, you will

[5] Alternatively, you can type the direct command **prtesti *size1 p1 size2 p2***, where size# and p# corresponds to the sample size and the sample proportion of population #.

find $\bar{p}_1 - \bar{p}_2$ and $s_{\bar{p}_1 - \bar{p}_2}$ in the **Mean** and **Std.Err.** columns, respectively, with which you can calculate the appropriate test statistic and p-values.

NOTE ON SMALL SAMPLE SIZES

When doing two population statistics when one or both samples are small (fewer than 30, say), some additional issues arise. First, as in the single population case, we cannot assume that our estimators (the sample means) are normally distributed unless we think the populations follow distributions close to normal. Second, if for some reason we believe that the two populations have the same standard deviation, then we can make use of that fact to obtain estimates that (in the case of small samples) are significantly more efficient. Though we will not cover techniques for dealing with these special cases, you should be aware these issues arise when you have small sample sizes.

Example: Political Gender Gaps

Men and women may have significantly different opinions on political candidates. One month before the 2003 California Governor's recall ballot, a Field Poll[6] noted several gender gaps among the top candidates including Cruz Bustamante, Arnold Schwarzenegger, and Tom McClintock. According to their press release, we are told that Cruz Bustamante is the first choice to replace Governor Gray Davis by 26 percent of likely male voters and 35 percent of likely female voters. Is this gender difference in support for Bustamante statistically significant? A difference is statistically significant only if we can prove it is not equal to zero using a hypothesis test. To try to do so, we use the following hypotheses (where p_m and p_w are the true proportions of men and women, respectively, supporting Bustamante):

$$H_0: p_m - p_w = 0$$

$$H_a: p_m - p_w \neq 0$$

To carry out this test, we need to know the sample sizes. The last page of the press release tells us that the total sample size was 505, so assume 252.5 men and 252.5 women. (This should be approximately right since they were sampled randomly.) Then we get an estimated standard deviation of the difference in proportions:

$$s_{\bar{p}_m - \bar{p}_w} = \sqrt{\frac{.26(1-.26)}{252.5} + \frac{.35(1-.35)}{252.5}} = 0.041$$

and a test statistic:

[6] The Field Poll, Tuesday, Sept 9th, 2003.

$$z = \frac{.26 - .35 - 0}{.041} = -2.207$$

The above test statistic gives a p-value of 0.027 (=2*normal(-2.207)), i.e., there is only a 2.7% chance that a difference this large could be due to sampling error rather than a genuine difference in the proportions of men and women supporting Bustamante. If we were using a 5% level of significance, we would conclude that the gender gap in support for Bustamante was significant.

The exercises at the end of this chapter should give you plenty of practice in using these techniques.

Further information from the Field Poll, Tuesday, Sept 9th, 2003:

Replacement candidate preferences by subgroup

…There is a significant gender gap in voter preferences in the replacement election. Bustamante holds a thirteen-point advantage over Schwarzenegger among women voters, 35% to 22%, while men are slightly favoring Schwarzenegger (29% to 26%)….

[*table 3* reports that 16% of men and 10% of women voters prefer Tom McClintock. while *table 7* shows that in the vote to recall Governor Davis, 38% of men and 41% of women support the governor and would vote against the recall.]

About the Survey Sample Details

The findings in this report are based on a telephone survey conducted September 3–7, 2003, in English and Spanish among a random sample of likely voters in California. A representative sample of [505 likely voters was selected]…. According to statistical theory, results from the overall likely voter sample have sampling error of ±4.5 percentage points at the 95 percent confidence level. Results from subgroups have somewhat larger sampling error ranges. There are other possible sources of error in any survey in addition to sampling variability. Different results could occur because of differences in question wording, sampling, sequencing, or through omissions or errors in interviewing or data processing. Extensive efforts were made to minimize such potential errors.

2.6 Asset Returns

Another interesting application comes from finance. The data set here consists of 20 years of monthly data (1926–1945) on the returns for various different asset classes: the S&P500, portfolios of small stocks (the bottom 20% of market capitalization of the New York Stock Exchange (NYSE)), of corporate bonds, of government bonds, and of Treasury bills. (The data can be found in the file **capm**.) Investment decisions are often based in part on past performance, so a natural question to ask is whether performance has been stable over time. In this example, we will try to determine if the average return on an asset class changed over the period.

This will be a hard question to answer. For example, could one ever reject a theory that said that every month is unique with a different average return? Furthermore, if you define the asset class closely enough, it is highly likely that the characteristics of the return distribution change across time due to, for example, industry-specific changes in regulations or technical innovations.

Because of this, we will start with a simpler idea. We take our 20-year sample and ask if the data suggest that average returns are stable over the period for the broad asset classes about which we have data, by comparing average returns in the first 10 years with average returns in the second 10 years.

We begin by taking a closer look at the data set. We can graphically examine the performance of one of these portfolios, the S&P500:

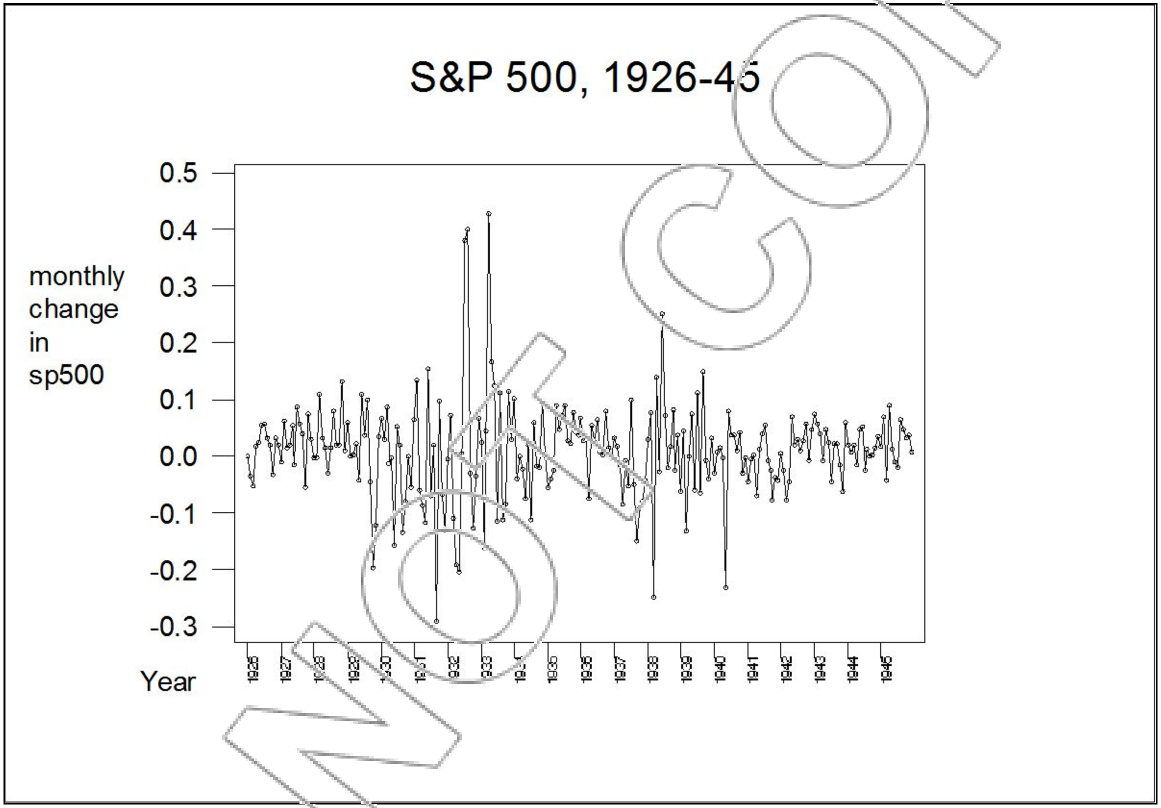
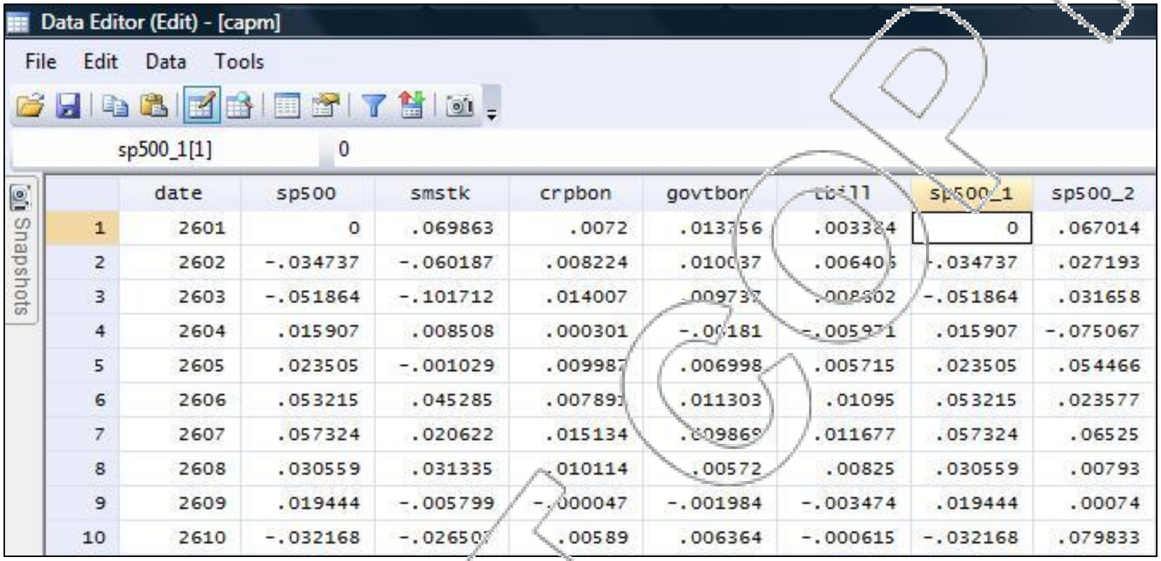


Figure 2.7: S&P 500 monthly returns (0.1 = 10%).

Market returns in this period displayed extraordinarily high variance compared with today.

To carry out our test, we first need to create two new variables, **sp500_1** and **sp500_2**, where **sp500_1** contains the returns for the S&P500 in the first 10 years, while **sp500_2** contains the returns for the S&P500 in the second 10 years. To do this in Stata, you can open the Data Editor and directly copy the first 120 observations from the **sp500** column to the **sp500_1** column. Then, copy the next 120 observations from the **sp500** column to the **sp500_2** column. Your dataset should look like this:

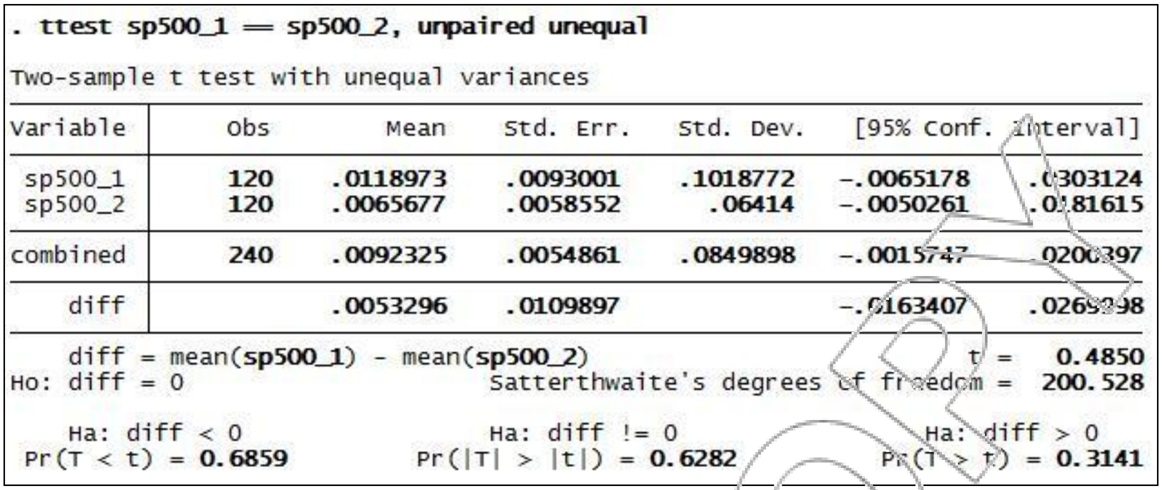


sp500_1[1] 0

	date	sp500	smstk	crpbon	govtbon	tbill	sp500_1	sp500_2
1	2601	0	.069863	.0072	.013756	.003384	0	.067014
2	2602	-.034737	-.060187	.008224	.010037	.006405	-.034737	.027193
3	2603	-.051864	-.101712	.014007	.009737	.008802	-.051864	.031658
4	2604	.015907	.008508	.000301	-.00181	-.005971	.015907	-.075067
5	2605	.023505	-.001029	.009987	.006998	.005715	.023505	.054466
6	2606	.053215	.045285	.007891	.011303	.01095	.053215	.023577
7	2607	.057324	.020622	.015134	.009869	.011677	.057324	.06525
8	2608	.030559	.031335	.010114	.00572	.00825	.030559	.00793
9	2609	.019444	-.005799	-.000047	-.001984	-.003474	.019444	.00074
10	2610	-.032168	-.026507	.00589	.006364	-.000615	-.032168	.079833

Now that we have created the new variables, we can conduct our test by clicking **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample mean-comparison test**. Choose **sp500_1** and **sp500_2** as your first and second variable, and check the box next to "Unequal variances."[7] Click **OK**, and Stata will return the following:

[7] Alternatively, you can directly type the command **ttest sp500_1 == sp500_2, unpaired unequal**.



```
. ttest sp500_1 == sp500_2, unpaired unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 sp500_1 |     120    .0118973    .0093001    .1018772   -.0065178    .0303124
 sp500_2 |     120    .0065677    .0058552      .06414   -.0050261    .0181615
---------+--------------------------------------------------------------------
combined |     240    .0092325    .0054861    .0849898   -.0015747    .0200397
---------+--------------------------------------------------------------------
    diff |            .0053296    .0109897               -.0163407    .0269998
------------------------------------------------------------------------------
    diff = mean(sp500_1) - mean(sp500_2)                          t =   0.4850
Ho: diff = 0                     Satterthwaite's degrees of freedom =  200.528

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6859         Pr(|T| > |t|) = 0.6282          Pr(T > t) = 0.3141
```

As shown in the output, the average monthly return for the first 10 years of the S&P500 is 1.19%, and the average monthly return for the last 10 years is 0.66%. Notice the substantial difference between the two sample average returns. A monthly return of 1.19% gives 15.25% annually, and 0.66% per month gives 8.21% a year. Nonetheless, since the p-value for the test with the null hypothesis that the two means are identical is large (p = 0.6282), we cannot reject the hypothesis that the mean monthly return is the same in both halves of the sample. That may seem like a surprising conclusion, but the lesson is that with so much variation in the month-to-month performance, as shown in the graph above, drawing any conclusions is difficult. Mathematically, the variation in returns makes the standard error, $s_{\bar{x}_1-\bar{x}_2}$, larger, which, in turn, makes the test statistic closer to zero and the p-value larger.

If we do the same hypothesis test for the small stock portfolio, we get a p-value of 0.6694. The average monthly return for the first 10 years of the small stock portfolio is 1.2%, and the average monthly return for the last 10 years is 2.0%. Again, despite our large estimate of the difference, we conclude that it is not statistically significant. That is, though the average returns in the first decade seemed to be lower, there is no strong evidence that this difference was real, so you would not want to rely on this difference as a basis for decision making.

SUMMARY

In this chapter, we learned how to support or reject a claim with data. Hypothesis testing allows us to ascertain the strength of the evidence provided by our data in support of an alternative hypothesis (against a null hypothesis). After learning how to structure and conduct one-tailed and two-tailed tests for a population mean or proportion, we learned how to conduct the same types of tests for the difference between two means or proportions. We learned how to use Stata to handle much if not all of the computational aspects of hypothesis testing. When we apply hypothesis testing to regression analysis later on, the computer will anticipate our interest in conducting certain important tests and will report back information about these tests making the computational aspects of testing almost effortless. Therefore, understanding how to interpret key numbers such as test statistics and p-values and how to choose appropriate hypothesis tests will be central to our study.

NEW TERMS

Hypothesis testing	The method used to prove or support arguments with statistics
Null hypothesis (H_0)	The default assumption; the opposite of the alternative hypothesis
Alternative hypothesis (H_a)	The statement you are trying to prove or show is true
Type I error	Rejecting the null hypothesis when it is true
Type II error	Failing to reject the null hypothesis when it is false
Level of significance (α)	The maximum acceptable probability of making a type I error

Test statistic The number of standard deviations that our estimator is away from the equals value in the null hypothesis

P-value The maximum probability of obtaining a test statistic value that is at least as unlikely as the observed one if the null hypothesis is true; used to determine the strength of the data's support for the alternative hypothesis

One-tailed test A hypothesis test where the alternative hypothesis uses a > or < sign.

Two-tailed test A hypothesis test where the alternative hypothesis uses the ≠ sign

NEW FORMULAS

Generically, the test statistic is computed using this formula:

$$\boxed{\frac{\text{estimator} - \text{value in the null hypothesis}}{\text{(estimated) standard deviation of the estimator}}}$$

Specifically, we learned the test statistics for the following circumstances:

Test statistics having a t-distribution

For a test concerning a population mean when the standard deviation must be estimated:

$$\text{t} = \frac{\bar{x} - \mu_0}{s_{\bar{x}}}$$

follows a t-distribution with n-1 degrees of freedom if $\mu = \mu_0$

For a test concerning the difference of two population means when the standard deviations must be estimated:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)_0}{s_{\bar{x}_1 - \bar{x}_2}}$$

follows a t-distribution with approximately n_1+n_2-2 degrees of freedom if

$$\mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$$

Test statistics having a standard normal distribution (assuming a large sample size)

For a test concerning a population proportion:

$$z = \frac{\bar{p} - p_0}{s_{\bar{p}}}$$

where

$$s_{\bar{p}} = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

For a test concerning the difference of two population proportions:

$$z = \frac{\bar{p}_1 - \bar{p}_2 - \left(p_1 - p_2\right)_0}{s_{\bar{p}_1 - \bar{p}_2}}$$

where

$$s_{\bar{p}_1 - \bar{p}_2} = \sqrt{\frac{\bar{p}_1\left(1-\bar{p}_1\right)}{n_1} + \frac{\bar{p}_2\left(1-\bar{p}_2\right)}{n_2}}$$

NEW STATA FUNCTIONS

Statistics>Summaries, tables, and tests>Classical tests of hypotheses>One-sample mean-comparison test

This opens the **ttest - Mean-comparison test** dialog box, where you can choose the variable for which you want to conduct a one- or two-tailed test for the population mean. Stata will return the test statistic as well as the p-values. The leftmost p-value corresponds to the alternative hypothesis that the population mean is less than the hypothesized mean; the middle p-value corresponds to the alternative hypothesis that population means is not equal to the hypothesized mean; the rightmost p-value corresponds to the alternative hypothesis that the population mean is greater than the hypothesized mean.

Alternatively, you can directly type the command **ttest *varname* == #, level(#)**. Omitting the **level(#)** option will tell Stata to use the default 95% confidence level for calculating the confidence intervals in the output.

Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample mean-comparison test

This opens the **ttest - Two-sample mean-comparison test** dialog box, where you can choose the two variables for which you want to conduct a one- or two-tailed test with the null hypothesis that the population means are equal. Checking the box next to “Unequal variances” specifies that the two populations are not assumed to have equal variances. Stata will return the test statistic as well as the p-values corresponding to the alternative hypotheses that the difference in population means is less than, not equal to, or greater than 0. Stata also lists the standard deviation of $\bar{x}_1 - \bar{x}_2$ in the **Std. Err.** column and the **diff** row. Note that Stata’s p-values, which are calculated using Satterthwaite’s degrees of freedom, may be slightly different from p-values calculated manually using n_1+n_2-2 degrees of freedom.

Alternatively, you can directly type the command **ttest *varname1 == varname2*, unpaired unequal level(#)**.

Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample proportion test

This opens the **prtest - Two-sample proportion test** dialog box, where you can choose the two variables for which you want to conduct a one- or two-tailed test with the null hypothesis that the population proportions are the same. Note that in conducting such a test, Stata calculates $s_{\bar{p}_1-\bar{p}_2}$ differently from the formula specified in this textbook, as under the null hypothesis the variances of the two populations should be equal and Stata takes this into account in its calculation. This is the reason for slightly different test statistic and p-values than the ones you would get using the formulas in the text. However, you can find the value for $s_{\bar{p}_1-\bar{p}_2}$ as in the text in the **Std. Err.** column and the **diff** row.

Alternatively, you can directly type the command **prtest *varname1 == varname2***. To specify the confidence level to use for confidence intervals, add the command **, level(#)**.

Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample proportion calculator

This opens the **prtesti - Two-sample proportion test calculator** dialog box, where you can enter the respective sample sizes and sample proportions of two populations of interest to conduct a one- or two-tailed test with the null hypothesis that the population proportions are the same. The

prtesti command is useful when you do not have an actual dataset. Note that you must enter integer values for sample sizes.

Alternatively, you can type the direct command **prtesti *size1 p1 size2 p2***, where ***size#*** and ***p#*** correspond to the sample size and the sample proportion of population #.

CASE EXERCISES

1: The gender gap

Look at the Field poll numbers in the Gender Gap example of Section 2.5.

a. Justify the claim in the last paragraph that "According to statistical theory, results from the overall likely voter sample have sampling error of ±4.5 percentage points at the 95 percent confidence level."

b. The last paragraph notes that "Results for subgroups have somewhat larger sampling error ranges." Estimate the "larger sampling error range" for the approval ratings of Arnold Schwarzenegger among likely women voters.

c. Test using a 5% level of significance if a gender gap exists in the approval ratings of Arnold Schwarzenegger.

d. Test using a 5% level of significance if a gender gap exists in the approval ratings of Tom McClintock.

e. Do the same for Gray Davis. In his case, would the gap have been significant if the sample proportions were the same but the sample had included 1,000 likely voters? What about if it had included 10,000 likely voters? What lesson do your answers suggest?

2. The January effect

To carry out this exercise, you need to access the **capm** dataset. Look for a "January effect" in small stocks, i.e., test if the average returns on a portfolio of small capitalization companies are different in January than in the rest of the year. Finance experts are particularly interested in looking for this kind of effect. (In finance, the efficient markets hypothesis suggests that any such anomaly is a profit opportunity.) To carry out this test you can use the **ttest** command in Stata. An easy way to do this is to first create a "dummy variable" for January, i.e., in the data editor, you will need to make a new column that contains a 1 whenever the cell is in a row which corresponds to January (look for the date in column 1) and a 0 for any other month. One way to do this is to type the 1 and the eleven zeros for the first year, and then cut and paste all the other years.[8] After creating the dummy variable (you can call it **January**), click **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-group mean-comparison test**. Choose **smstk** as your variable name, choose **January** as your group variable name, and check the box next to "Unequal variances."[9] This tells Stata to conduct a hypothesis test with the null hypothesis that the average returns in January (i.e., **January** = 1) are the same as the average returns in the rest of the year (i.e., **January** = 0) for small stocks. Report the p-value and explain what it suggests about the existence of a January effect for small stocks. Repeat the exercise for the S&P500. Finally, test to see if the return on T-bills was different in U.S. presidential election years than in other years. To do this in Stata, you need to create a new dummy variable for the election years, and conduct your hypothesis test using the new dummy variable as your group variable.

3: Fast food nation

[8] See the Appendix for more detail on generating a variable with repeated patterns.

[9] Alternatively, you can directly type the command **ttest smstk, by(January) unequal**.

A recent Gallup Poll (July 7–9, 2003) addressed the idea of holding the fast food industry responsible for the social costs of obesity in the United States. One question divided those surveyed into people who thought that fast food was good for you and those who disagreed. Two hundred thirty-six of the 1,006 people surveyed believed that fast food was good for you, and 770 of the 1006 surveyed thought that fast food was not good for you.

The survey examines if people should accept responsibility for their dietary behavior. The poll asked people how frequently they ate at fast food restaurants. Half of those who believed that fast food was not good for them ate fast food at least once a week. That is, 50% of the “not good for you” group ate fast food at least once per week. This compares with 62% for those who think that fast food is good for them.

a. Does this data show that people who believe that fast food is good eat fast food more often than those who believe that it is not good? Justify your answer.

The same survey asked infrequent fast food diners (less than once per month) if they would be more likely to eat at fast food restaurants if the restaurants offered new healthier menu options. A major fast food company has decided to go ahead with such a plan because it believes at least half of the infrequent diners would respond Yes to that question. In the Gallup Poll, only 84 of the 204 infrequent fast food diners surveyed answered Yes.

b. Is this enough evidence to convince the company to change its mind? Justify your answer.

4: Pro bowling for dollars

Each year, the Hawaiian State Government pays the NFL about $5 million for the rights to host the Pro Bowl.[10] In return, the state gets to showcase its warm weather to about six million viewers in the depth of winter. Additionally, about 18,000 mainlanders who come to Hawaii to watch the game help boost the local economy. Assessing the impact of their spending is critical for the government that spends almost 10% of its annual tourism budget on the event. One important question is if these Pro Bowl tourists spend more or less time in the state during their stay than typical mainlanders who spend an average of 10.1 days per visit.

In 2003, the Hawaiian Tourism Authority conducted a poll of 260 Pro Bowl visitors and learned that the average stay was only 8.6 days. The sample standard deviation, s, was 5.7 days. Is this strong evidence that the average Pro Bowl Visitor stays fewer than 10.1 days?

PROBLEMS

For problems 1–3, you will need to access the file **bigmovies**[11] that contains data on major films released in 1998.

1. Studios believe that one important predictor of movie revenues is the release date. Since many young people have more free time when school lets out for the summer, more big films might be released during the summer months to take advantage of the surge in demand. Of course, studios might choose to release their movies at other times when there might be less competition. Another good time might be the holidays when more people have time off to go to the movies.

[10] All data from *Survey Adds Up Return on Pro Bowl* in the Honolulu Advertiser, 2/13/03.
[11] From Internet Movie Database at http://www.imdb.com

a. If summer months are more popular for film releases than the rest of the year, then the proportion of films released during the three months of summer should be more than 3/12 or 0.25. Define p_s to be the true proportion of films released during the summer months. Set up a hypothesis test to prove that summer months are more popular as release dates for big movies.

b. Use the data in the column titled “Summer Release” to carry out the test you set up in part a.

c. Conventional wisdom states that about 10% of all movies are released during the holidays, but you disagree. Define p_h to be the proportion of films released during the holidays. Set up a hypothesis test to show the conventional wisdom is untrue.

d. Use the data in the column “Holiday Release” to carry out the test you set up in part c.

2. Another variable to consider is a movie's Motion Picture Association of America (MPAA) rating. An R rating, for instance, might prevent many younger moviegoers from seeing the film which can reduce its revenue potential.

a. Calculate the sample average Total Domestic Gross (TDG) for each of the four MPAA rating categories (R, PG-13, PG, and G.) To do this in Stata, you can use the command **tabstat TotalDomesticGross, statistics(mean) by(MPAArating)** directly or build it through the **tabstat** dialog box (type **db tabstat** or use a menu).

b. Calculate the sample standard deviation of TDG for each MPAA rating category.

c. Set up hypothesis tests to determine if a statistically significant difference in population average TDG exists between each pair of categories. You will need to set up six separate tests (R vs. PG-13, R vs. PG, R vs. G, etc).

d. Use the formulas from Section 2.5 to calculate the test statistic for each of the six tests.

e. Use the test statistics from part d to compute p-values for each of the six tests.

f. Repeat the calculations for each test directly using Stata's **ttest** command. Ensure your answers resemble the ones you found in part e. Some rounding in the hand calculations will give you slightly different answers.

3. Another important factor in determining movie revenues is genre. Certain film types like comedies might have a broader appeal than other types, e.g., horror films.

a. Calculate the sample average Total Domestic Gross (TDG) for the following four types of films: Action, Comedy, Drama, and Horror.

b. Calculate the sample standard deviation of TDG for each of these four genres.

c. Set up hypothesis tests to determine if a statistically significant difference in population average TDG exists between each pair of categories. You will need to set up six separate tests.

d. Use the formulas from Section 2.5 to calculate the test statistic for each of the six tests.

e. Use the test statistics from part d to compute p-values for each of the six tests.

f. Repeat the p-value calculations for each test directly using Stata's **ttest** command. Ensure your answers resemble the ones you found in part e. Some rounding used in the hand calculations will give you slightly different answers.

4. The file **Hawaiipercapita**[12] contains information about the annual per capita income for Hawaii's four county governments. This information, collected by the Hawaii Department of Business Economic Development and Tourism, is used to allocate state funds for many social services.

a. Calculate the sample mean and standard deviation for each county.

b. Set up hypothesis tests to determine if a statistically significant difference exists between each pair of counties. You will need to set up six separate tests.

c. Use the formulas from Section 2.5 to calculate the test statistic for each of the six tests.

d. Use the test statistics from part c to compute p-values for each of the six tests.

e. Repeat the p-value calculations for each test directly using Stata's **ttest** command. Ensure your answers resemble the ones you found in part d. Some rounding in the hand calculations will give you slightly different answers.

5. The file **bank** has data from a mid-sized local bank. The bank has recently begun offering online banking services to its clients and is curious about the level of interest in the new product. The two columns contain data on the number of online banking brochures distributed on a sample

[12] See http://www2.hawaii.gov/DBEDT/.

of weekdays and Saturdays. Management has claimed that about 330 people are taking brochures about the new service every day.

a. Calculate the sample mean and standard deviation for each column of data.

b. Test the management's claim for Weekdays using α = 0.05.

c. Test the management's claim for Saturdays using α = 0.05.

d. Use Stata's **ttest** command to test if a difference exists in the number of brochures distributed on weekday and Saturdays using α = 0.05. Do these results make sense given your answers to parts b and c?

6. The file **restaurantstocks** contains monthly data on the excess returns of five publicly traded restaurant stocks from 1984–1994. The excess returns measure the difference between the stock's performance and the government T-bill rate. We would like to know if each stock performs significantly better, on average, than the government T-bill rate over time. This would be true if their average excess returns were positive.

a. Calculate the sample mean and standard deviation of excess returns for each stock.

b. Calculate the test statistic for each stock appropriate for proving average excess returns are positive

c. Test if each restaurant stock performs better on average than the government T-bill rate (i.e., has positive average excess return) using an α = 0.05.

d. Which of the five stocks has performed the best over the 11-year period?

e. Which stock has the smallest p-value in the tests from Part c?

f. Given that the sample size is the same for each stock, how can the stock which has the highest average return be different from the one with the smallest p-value?

7. The file **forbeswealth**[13] contains data on the wealthiest 100 Americans in 2001 and 2002 from a list compiled by *Forbes* magazine. Due to the sagging stock market, the wealth of many Americans declined between 2001 and 2002. We would like to know if the decline was experienced by the wealthiest Americans.

a. Compute the mean and standard deviation of the net worth of the wealthiest Americans in both years.

b. Did the average value of the net worth of the top 100 Americans decline from 2001 to 2002?

c. Was the change you observed in part b statistically significant? Use $\alpha = 0.05$.

8. The file **forbeswealth** from problem 7 contains data on the age of the 100 wealthiest Americans. An interesting question is if the average age of the wealthy is increasing, decreasing, or remaining constant. A decrease in the average age tends to correlate with new wealth being created, whereas an increasing age tends to be associated with less turnover and fewer new members on the list.

a. Compute the mean and standard deviation of the age of the top 100 wealthiest Americans in 2001 and 2002.

b. Did the mean age increase, decrease, or stay the same?

c. Was the change you observed in part b statistically significant?

[13] From ***Forbes***, 10/6/2003, Vol. 172 Issue 7, p136

CHAPTER 3

THE AUTORAMA: INTRODUCTION TO REGRESSION THROUGH INVENTORY PLANNING

In this chapter, we will introduce linear regression. The Autorama case presents a situation where a manager is planning how to allocate a limited amount of inventory space in a new car dealership. The manager has access to data from another dealership, which allow us to explore the relationship between car buyers' income and the amount of money they pay for their cars. Since the income levels in the two areas where the dealerships are located are different, the optimal number of each type of car to stock might be different as well. Projecting the relationship between income and price that exists in the first dealership onto the new one using the technique of regression analysis will allow the manager to plan the best mix of inventory. The theory of regression is mostly left to the final subsection of this chapter. The next chapter will elaborate on the technique and extend its applicability.

3.1 Introduction

Imagine that you work for a chain of auto dealerships. Your company is opening a new dealership, and you are in charge of choosing inventory. To do this, you need to predict what product mix is appropriate, i.e., what kinds of cars your customers will buy. The total number of cars you may stock is fixed at 200 (owing to considerations of space), and your job is to decide how to break those 200 cars down by price bracket. You have two kinds of data to help you. One dataset consists of a sample of (accepted) credit applications for financing new car purchases. These data come from another dealership (in the file **autorama**). The credit application tells you the income of the applicant, and the price of the car each is buying. A second set of data shows the neighborhoods served by each dealership; specifically, you have obtained estimates of the income distributions in each neighborhood, i.e., for each neighborhood you know the percentage of people in each income bracket. You also know something about the auto purchase habits of the public. (Specifically, you know the percentage of people in each income bracket who buy a new car in any given year.) The data for the new neighborhood (which is the data relevant to you) are presented in Figure 3.1. The total adult population of the new neighborhood is 10,000 people.

income bracket ($000's)	<15	15-25	25-35	35-45	45-55	55-65	65-75	75-85
% in income bracket	7.7	16.1	26.25	26.25	16.1	6.05	1.4	0.2
% (per year) who buy new cars	1	3	5	5	5	5	5	5
number of customers	7.7	48.3	131.25	131.25	80.5	30.25	7	1

Figure 3.1: Income distribution and expected number of customers by income for the new neighborhood.

How was this table constructed? We divided the population up into income brackets using known information on the income distribution in the new neighborhood. This information is summarized in the first two table rows. In the third row, we state the historical percentage of people in each income bracket (nationally) who buy new cars in a given year. This enables us to calculate our expected customer base in each income bracket as a proportion of the total population. Recall, this neighborhood has a population of 10,000 adults. For example, since 16.10% of these adults fall into the $15,000–$25,000 income bracket, and each year 3% will buy a car, we arrive at the number 10,000*(16.10/100)*(3/100) = 48.3 customers.

A first approach might be to examine the mix of cars being purchased in the sample from the existing dealership (and shown in the histogram in Figure 3.2) and use that as an estimate of the percentage of cars that will be sold in each price bracket at the new dealership.

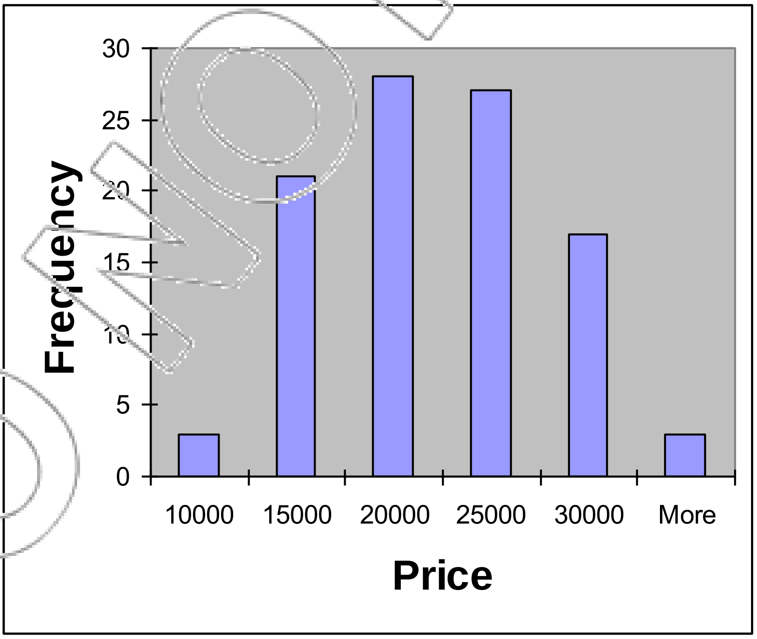


Figure 3.2: Frequency of purchases by price.

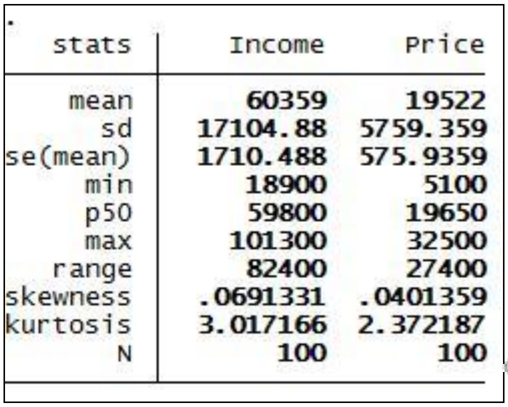
However, this approach has a problem, which is that you know the two neighborhoods have quite different income distributions. Though the average income in the new neighborhood is about $35,000, in the old one it is about $60,000. This suggests that your new customer base will be more interested in less expensive cars, so copying the product mix that is appropriate for the other dealership would be a mistake. You, therefore, decide to do something better: You will use the data from the first dealership to predict the car prices that people in a given income bracket will be interested in. You will combine this with what you know (from Figure 3.1) about the income distribution of your new customer base to get a more accurate prediction of what they will want.

3.2 Regressing Price on Income

The first thing you need to do is understand the relationship between people's income and the amount they will spend on a car. To do this, you will use the technique called regression.

Look at the data (in the **autorama** file). The data consist of 100 data points, i.e., 100 credit applications. The variable income stands for the annual income of each applicant and the variable price stands for the price of the car each is buying. Both variables are measured in dollars.

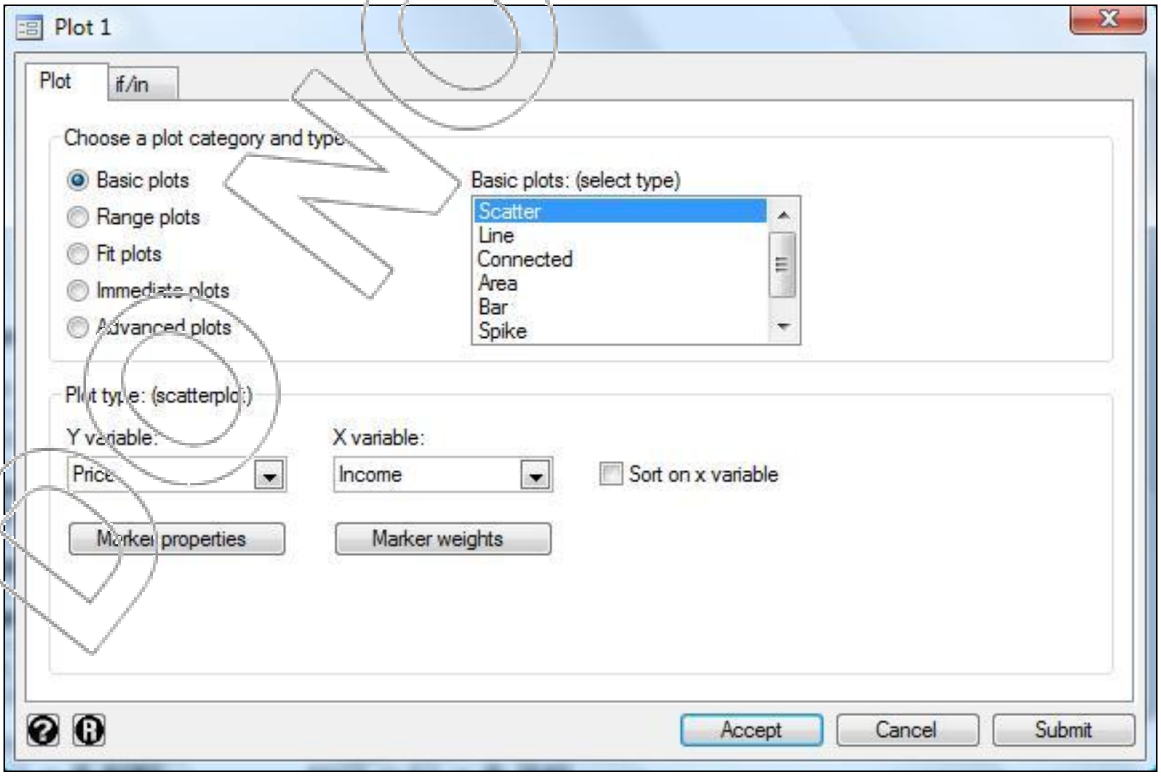
User>Core Statistics>Univariate Statistics>Standard (ktabstat)



```
.
    stats |    Income     Price
----------+--------------------
     mean |     60359     19522
       sd |  17104.88  5759.359
 se(mean) |  1710.488  575.9359
      min |     18900      5100
      p50 |     59800     19650
      max |    101300     32500
    range |     82400     27400
 skewness |  .0691331  .0401359
 kurtosis |  3.017166  2.372187
        N |       100       100
-------------------------------
```

Figure 3.3: Univariate statistics of income and price.

As you can see, the average income of applicants in our sample is $60,359 and the average price of the auto they are buying is $19,522. We get a better sense of what is in the data set by looking at a scatterplot of Price vs. Income (see Figure 3.4). You can generate this graph in Stata by clicking **User>Core Statistics>Bivariate Statistics>Bivariate Plots (twoway)** or typing **db twoway**. This will open the **twoway** dialog box. Click **Create…** and fill in the Plot 1 dialog box as shown:



Plot 1

Choose a plot category and type

● Basic plots

○ Range plots

○ Fit plots

○ Immediate plots

○ Advanced plots

Basic plots: (select type) Scatter

Plot type: (scatterplot)

Y variable: Price

X variable: Income

□ Sort on x variable

Click **Accept** and **OK**, and Stata will generate the following scatterplot:[1]

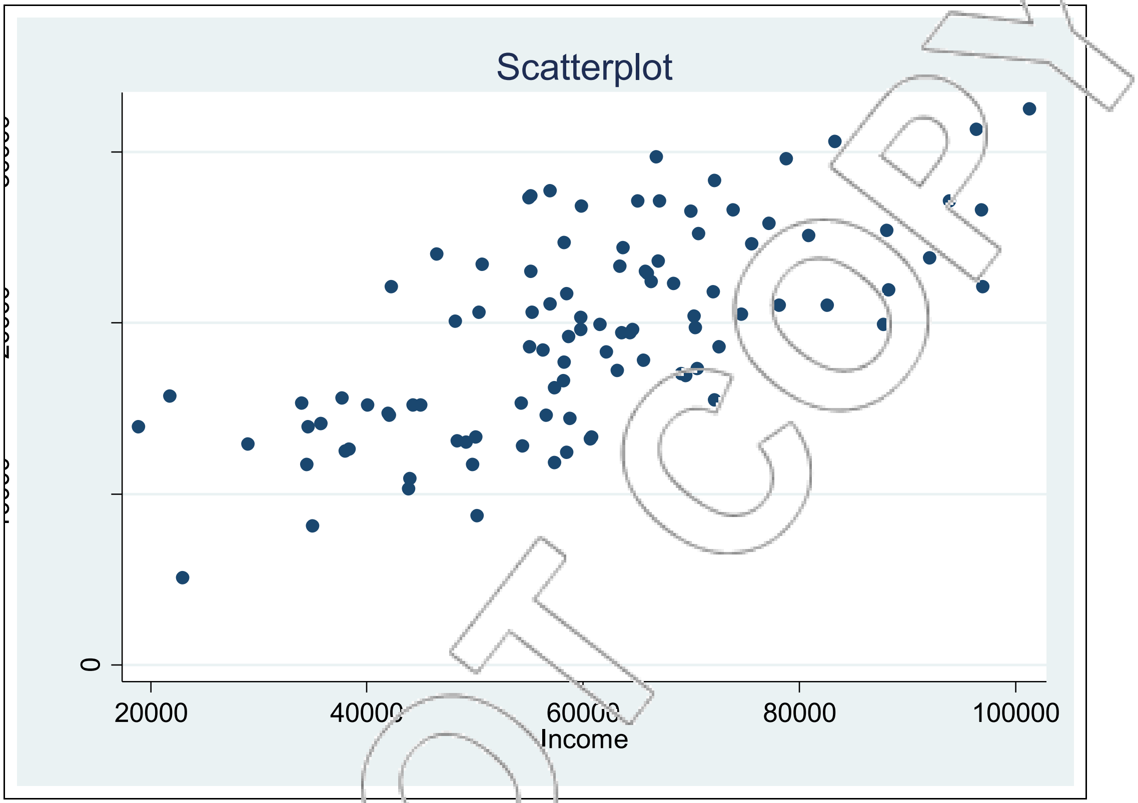
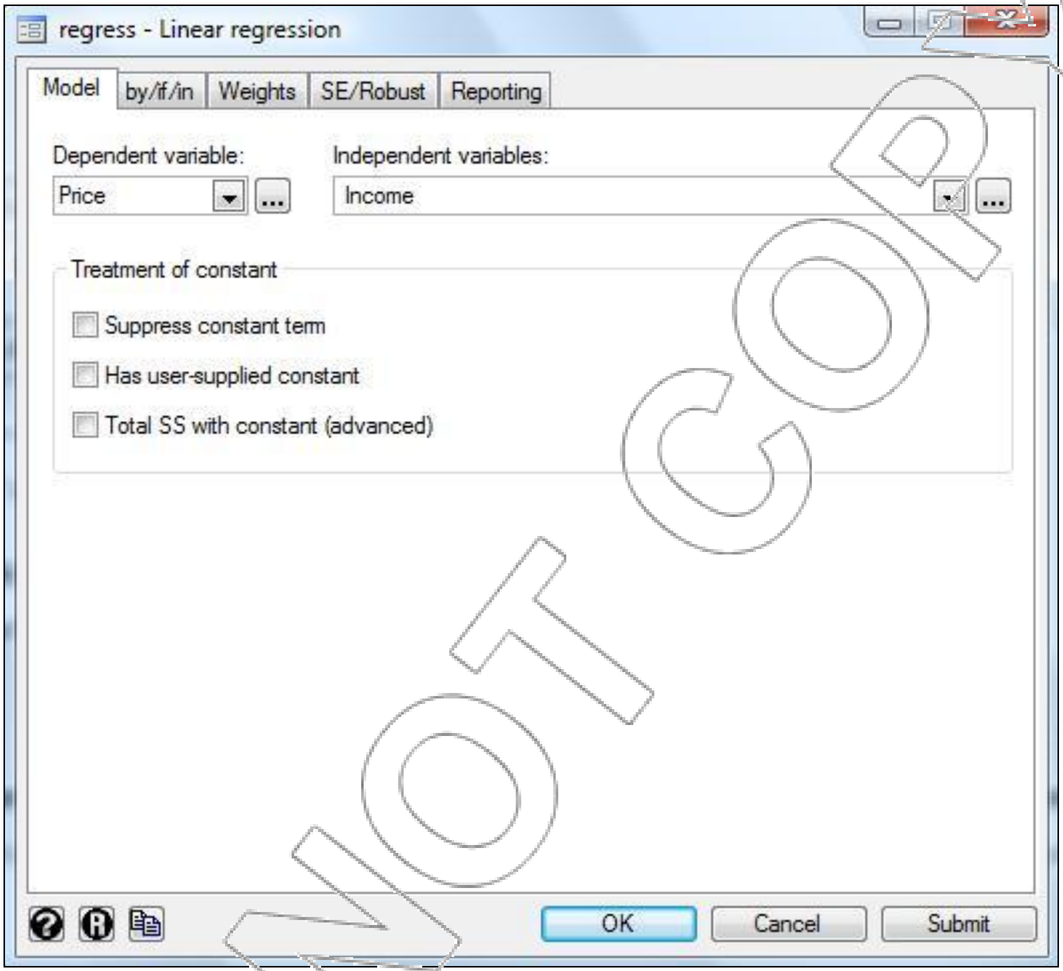


Figure 3.4. Scatterplot of price vs. income.

People seem to spend more on cars as their income rises, which is not surprising. More usefully, the relationship seems to be linear, i.e., you could draw a straight line through the scatterplot that would represent the data fairly well. But how should we choose the line, i.e., what line is going to give us the "best fit" to the data? The answer is provided by regression. We will ask Stata to produce the best-fit line by using the regression command. To do this, click **User>Core Statistics>Regression (regress)** or type **db regress**. Choose **Price** as your dependent variable

[1] Alternatively, you can directly type the command **twoway scatter Price Income**. After the graph is generated, you can click **File>Start Graph Editor** to edit your graph (such as adding titles and changing the scales of the axes). See the Appendix for more information on using the Graph Editor.

and choose **Income** as your independent variable. You should have a dialog box that looks like this:



Dependent variable: Price

Independent variables: Income

Treatment of constant

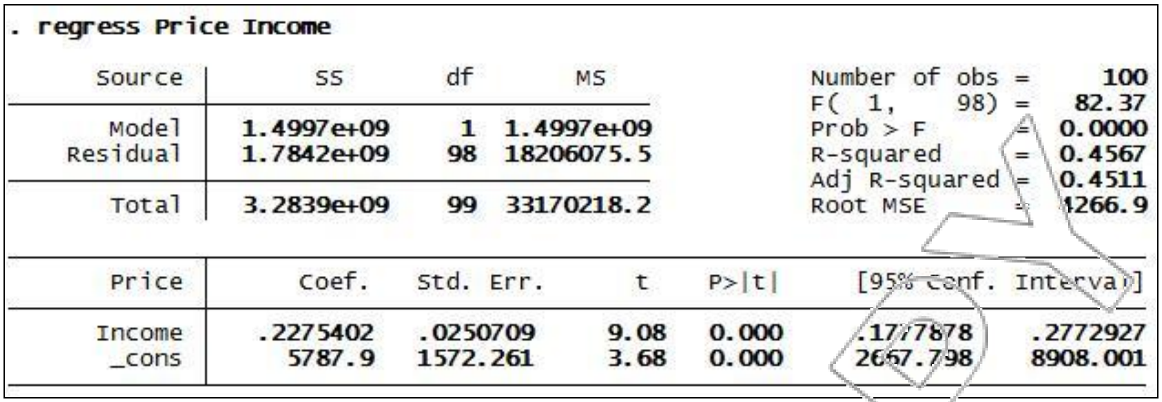
□ Suppress constant term

□ Has user-supplied constant

□ Total SS with constant (advanced)

Click **OK**, and Stata will generate the following output:[2]

[2] Alternatively, you can directly type the command **regress Price Income**. See the list of new Stata commands at the end of the chapter for more explanation.



```
. regress Price Income

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   82.37
       Model |  1.4997e+09     1  1.4997e+09           Prob > F      =  0.0000
    Residual |  1.7842e+09    98  18206075.5           R-squared     =  0.4567
-------------+------------------------------           Adj R-squared =  0.4511
       Total |  3.2839e+09    99  33170218.2           Root MSE      =  4266.9

------------------------------------------------------------------------------
       Price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Income |   .2275402   .0250709     9.08   0.000     .1777878    .2772927
       _cons |     5787.9   1572.261     3.68   0.000     2667.798    8908.001
------------------------------------------------------------------------------
```

Figure 3.5: Regression of price vs. income.

What does all this mean? First, we can write the estimated regression equation using the regression output table. In the regression we ran, **Price** is the variable on the left-hand side. On the right-hand side, we have the constant coefficient (5787.9) plus the coefficient on **Income** (0.2275) times **Income**. By equating left-hand side to right-hand side, we obtain the following equation:

price = 5787.9 + 0.2275*income

This equation represents what Stata has determined to be the best-fit line, as shown in the following diagram:[3]

[3] This graph can be generated in Stata by clicking **User>Core Statistics>Bivariate Statistics>Bivariate Plots (twoway)** or typing **db twoway** and creating two plots – a scatterplot as above and a “Fit plot” using “Linear prediction” of Price using Income and then clicking **OK**. This is equivalent to typing the command **twoway (scatter Price Income) (lfit Price Income)**. See the list of new Stata commands at the end of the chapter for more details.

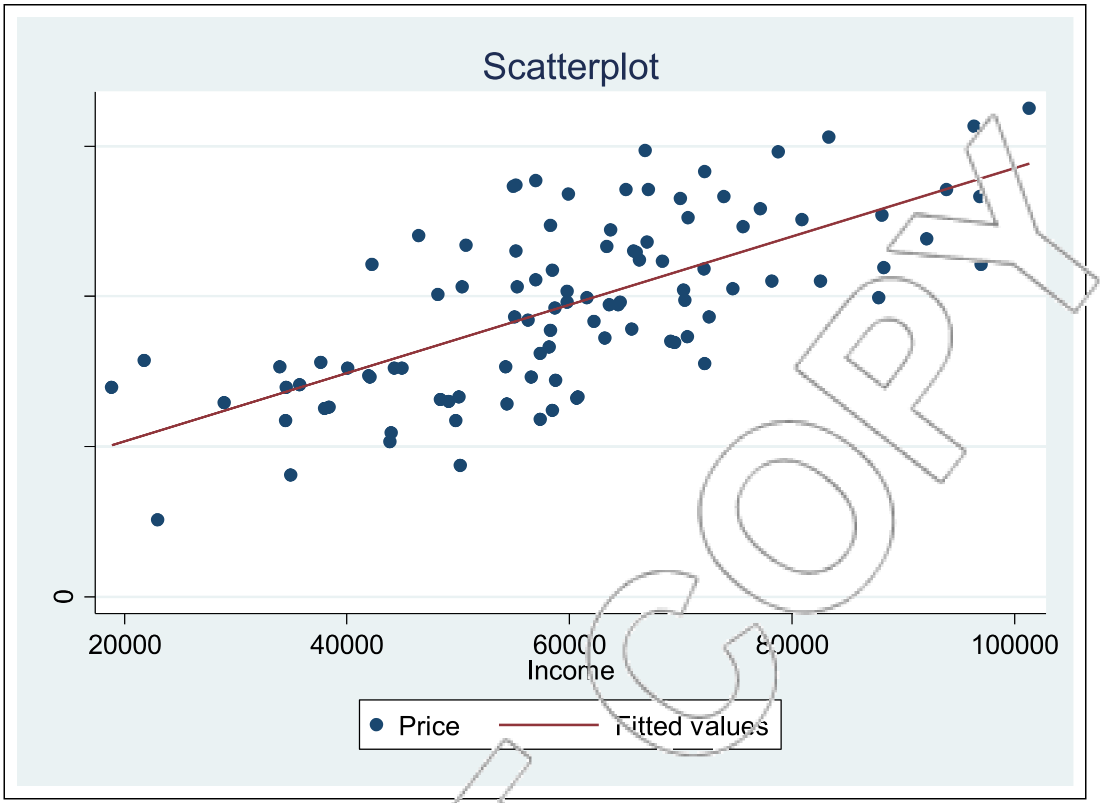


Figure 3.6: Scatterplot of price vs. income with regression line.

What it says is that the average amount spent on a new car by people with a given income is equal to, or best estimated by, $5,787.9 plus 0.2275 times their income. So, for someone earning $20,000, this estimate is $(5787.9+0.2275*20000) = $10,337.90, and for someone earning $80,000, it comes to $(5787.9+0.2275*80000) = $23,987.90.

All we have done is press a few buttons on the computer, so this may seem like magic. Before going on to use this equation, we will attempt to answer the two important questions that will allow us to understand regression better:

1. Where does this equation come from?
2. Why should we believe it provides a good estimate?

3.3 Method of Least Squares

Given any scatterplot, we would like to draw the best-fit line through the points in the diagram. To do so, we need to have some criterion for measuring what is a good fit. Intuitively, a line is a good fit if it is as close to the points as possible. So start off with a line, and see how far it is from each point. We call this distance the error, and we would like to make the errors as small as possible. We can see these errors more easily on a scatterplot with fewer points, as in Figure 3.7.

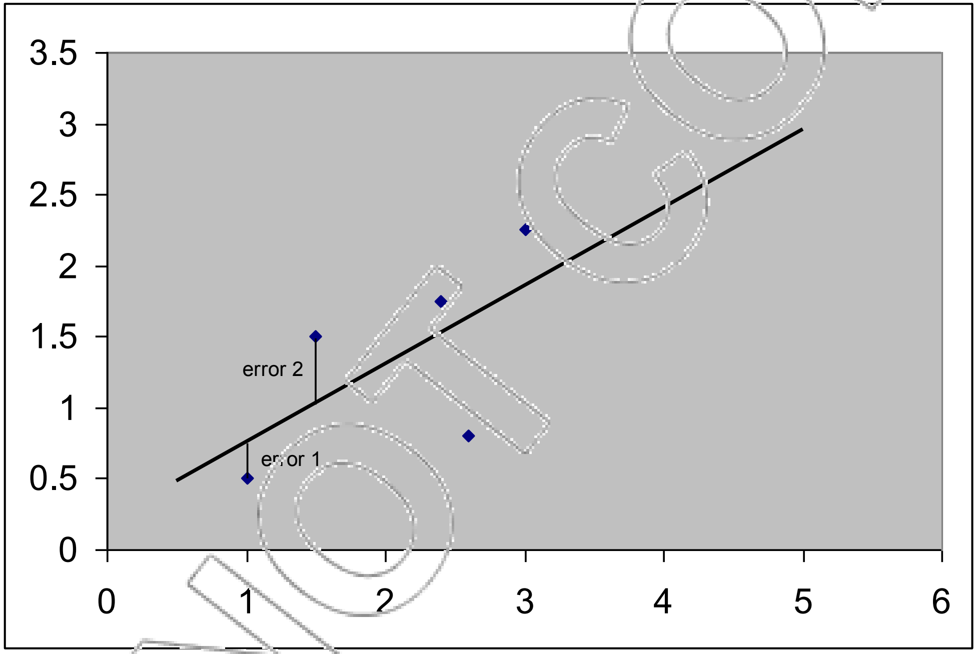


Figure 3.7: Generic scatterplot.

In this picture, we have drawn a straight line through a set of five points. The error associated with each point is the vertical distance from the line to that point. (We have marked the first two errors in the picture.) We define the sum of squared errors as the number obtained by calculating each of these distances in turn, squaring each one, and then adding all these squares. Intuitively, the number we get this way will be small if the line is close to the points, and large if it is far from them.

We can use this procedure to compare two different lines for goodness of fit. Do the calculation for each line, and then say that the one with the smaller sum of squared errors is a better fit. This suggests that we define the best-fit line as follows:

The best-fit line is the line that produces the smallest possible sum of squared errors.

Now, we can answer our first question. The equation that Stata spits out from the dataset is the equation of the best-fit line. Examine the following equation of the best-fit line:

price = 5787.9 + 0.2275 income

If we take this line, calculate the sum of squared errors, and take any other line at all and repeat the calculation, we will get a bigger number the second time.

How does Stata do this? For our purposes we really do not need to know. That is not to say that we will be using regression in a mindless or mechanical way, but what we need to understand are the underlying statistics and interpretation and not the mechanics of selecting the best-fit line. In practice and in this text, the mechanics of regression are always carried out by computer.

This approach also provides a partial answer to our second question. For example, if you look back at the summary statistics, you will see that the average price of a car ($19,522) is about one third of the average income of the people in our sample ($60,359). So, rather than running a regression, someone might suggest using the simple rule of thumb that people will buy a car whose price is about one third of their annual income. We will need to justify why the regression equation is considered a better way of estimating than this rule. One argument is that the regression equation is better than the one-third rule in the sense that it provides a better fit. We can represent the one-third rule by the following line, depicted as the 'rule of thumb' line in Figure 3.8.[4]

price = 0.333 income

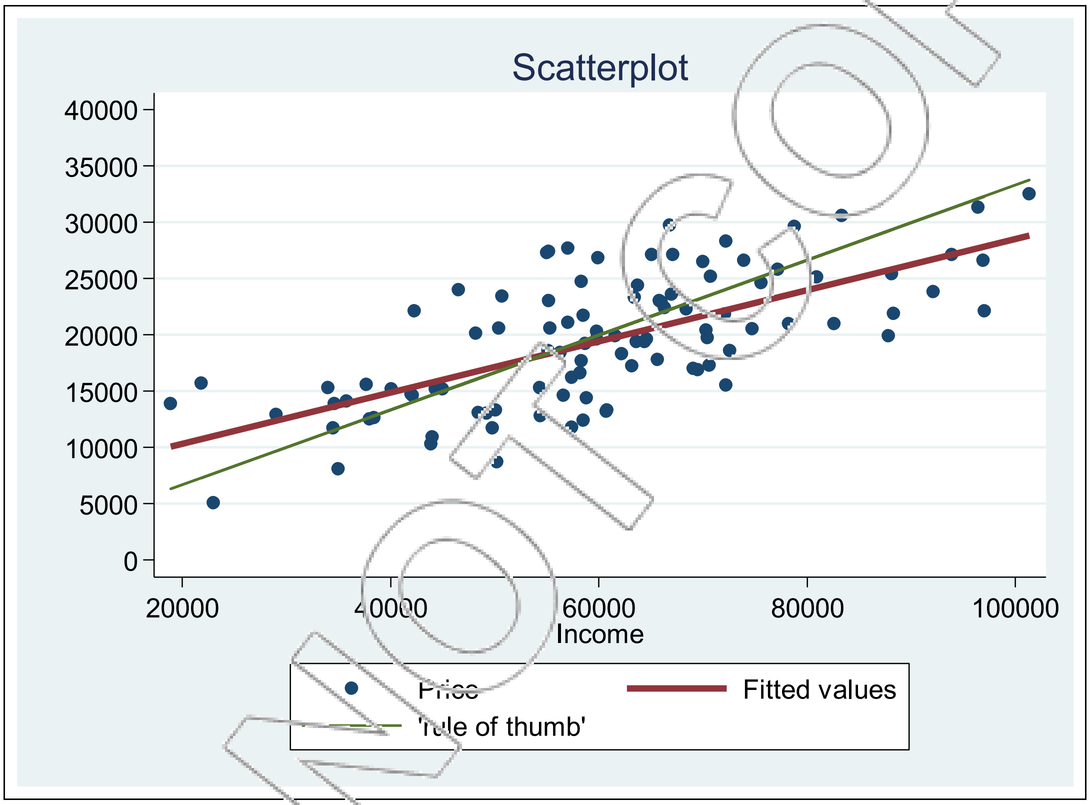


Figure 3.8: Scatterplot of price vs. income with regression and rule of thumb lines.

Using this line, the sum of squared errors is larger than the sum of squared errors from the regression line. The practical consequence of this is that estimates found using the regression line

[4] This graph can be generated in Stata by typing the following commands: 1)**generate price1 = 0.333*Income**; 2) **twoway (scatter Price Income) (lfit Price Income) (line price1 Income)** (or using the **twoway** dialog box to generate the equivalent command); and 3) using the Graph Editor to change the label in the legend to read 'rule of thumb' rather than Price1.

will be more precise (i.e., have a smaller variance) than estimates from the rule of thumb or any other line.

Now we will examine how to use the regression equation to predict demand for cars at our dealership.

3.4 Predicting Spending from the Regression Equation

Think about the people in our customer base (in statistics jargon, the population) who earn $30,000 a year. Not all of them will want to spend the same amount on a car, so what we would like to find is a distribution of their spending levels. We will make two assumptions about the distribution of spending levels for a given income.

ASSUMPTIONS

1. For each income level, spending on a car purchase is approximately normally distributed.
2. The distribution for different income levels need not have the same mean, but it does have to have the same standard deviation.

Later in this text, we will discuss the second of these assumptions in some detail. For the time being, we will ask you to take their validity on trust. They can both be tested, and in this case, the tests suggest they are reasonably correct.

Starting with our $30,000 income group, the first assumption implies we only need to know two things about the distribution of spending for this group: its mean and its standard deviation. The regression output gives us estimates of both. The mean is estimated by setting income = $30,000 in the regression equation, so it is $(5,787.9+0.2275*30,000) = $12,612.90. You can find an estimate of the standard deviation in the regression output from Figure 3.5 in the row labeled **Root MSE**. It is estimated by s = 4266.9, i.e., it is $4,266.90. So, our best guess is that, among people with annual income of $30,000, spending on a car purchase is normally distributed with a mean of $12,612.90 and a standard deviation of $4,266.90, as shown in the histogram in Figure 3.9. The estimate of the mean depends on these people having an income of $30,000, but the estimate of the standard deviation does not, which fits assumption 2 above.

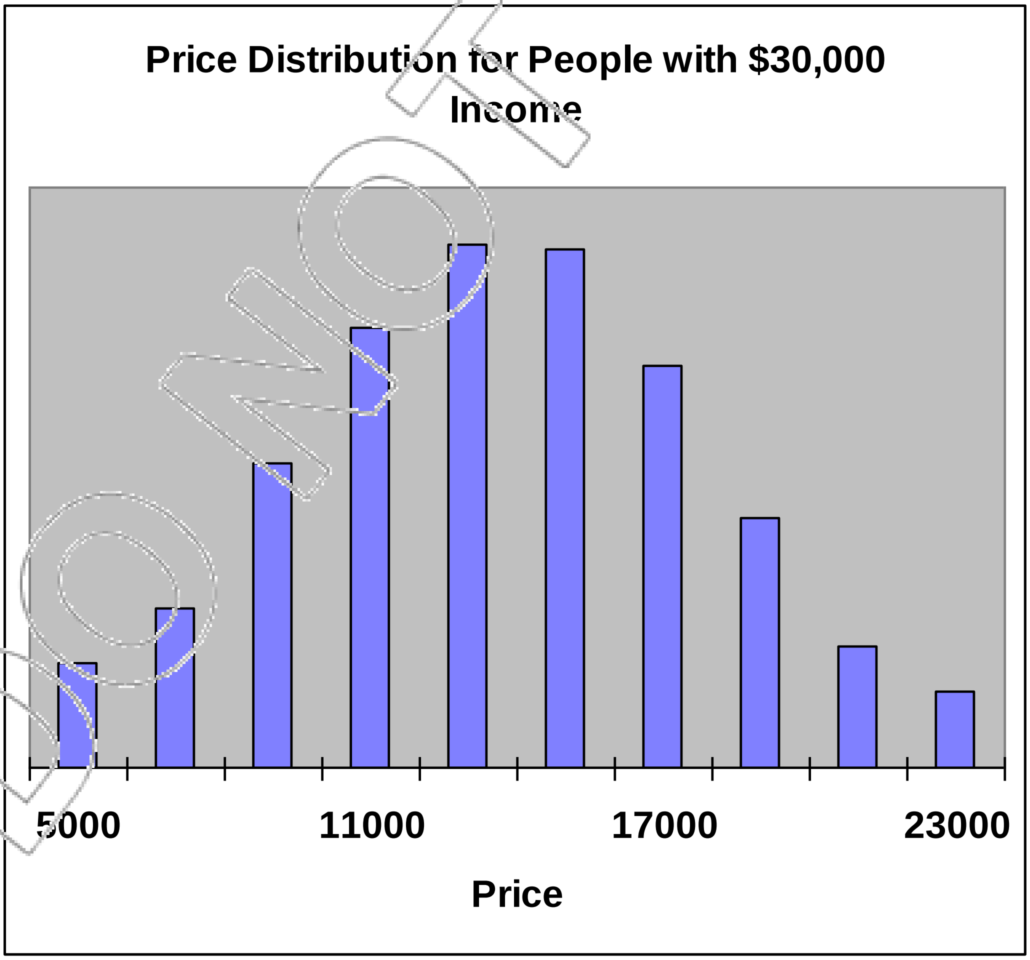


Figure 3.9: Price distribution for income level of $30,000.

What we will do now is divide our cars into a series of price brackets and use our knowledge of the normal distribution to say what proportion of these people will buy autos in each bracket. For example, we know that the proportion of prices paid by this income group, which are below $16,000, is the same as the area to the left of 16,000 in a normal distribution with a mean of 12,612.90 and a standard deviation of 4,266.9. One way to calculate this area is to use the standard normal distribution. Standardizing the value 16,000 by subtracting the mean and dividing by the standard deviation yields the following:

$$z = \frac{16{,}000 - 12{,}612.90}{4{,}266.9} = 0.7938$$

Therefore, for this income group, the proportion of prices paid that are less than $16,000 is the area to the left of 0.7938 in a standard normal distribution. Using Stata, you can calculate this area by typing **display normal(0.7938)** in the Command box. This area is 0.7863. So, this tells us that an estimated 78.63% of the population in the $30,000 income group buys cars priced below $16,000. By a similar analysis, the proportion buying cars priced below $14,000 is 62.74%, so this tells us that (.7863-.6274)*100 = 15.89% of these customers will buy in the $14,000–$16,000 price bracket. We can do the same calculations for $10,000–$12,000, $12,000–$14,000, and every other price bracket, giving a complete picture of the demand for customers with an income of $30,000. (For convenience, we have divided car prices into $2,000 price brackets.)

We now know something about the price preferences of the customers with a given income. How do we use this information to get a picture of the overall spending distribution? Well, there are several steps.

For each income bracket in the table giving the income distribution for the new neighborhood, we will assume that all individuals in a bracket behave as if they had the median income for that bracket. The median for this neighborhood happens to be the mid-point of each income range, with the exception of the lowest income bracket, for which the median is $10,000. Also, the median for the highest bracket is $120,000. Now, for each income bracket, we use the regression estimates to calculate the number of customers we expect to fall inside each price bracket.

For example, if we want to predict the number of customers in the $35,000–$45,000 income bracket who will buy a car in the $12,000–14,000 car bracket, we proceed as follows: First, we calculate, using the regression estimates and the median income for the bracket, that purchases of cars by that income bracket are normally distributed with a mean of $(5,787.9+.2275*40,000) = $14,887.90 and a standard deviation of $4,266.90. Then, we use the normal distribution to find what proportion of that demand lies between $12,000 and $14,000. You can work this out by the same technique as above. You should get an answer of about 0.1683 (or 16.83%). Then, multiply this proportion by the number of customers in that income bracket (131, from Figure 3.1) to get the number who are expected to buy in that price bracket ((131*0.1683) = 22.05, or about 22 people).

For any particular price bracket, add the number of customers from each income bracket who will want to buy a car in that price bracket. This gives the total number of cars in that bracket that would be sold in a year, given our neighborhood of 10,000 people. This gives you the demand information you need to make your decision on what mix of cars to stock.

WARNING:

This procedure is reasonably good. However, we have made one dubious approximation. For the purposes of our prediction, we are acting as if the estimates of the mean and standard deviations of prices for each income level were exact; they are not. The mean and standard deviations are estimates from our sample and, therefore, subject to sampling error. This could be taken into account by using slightly more sophisticated statistical techniques, which we will learn when we talk about prediction intervals. Meanwhile, you should be aware that we have used this shortcut. Of course, some other approximations are present as well, due to income bracketing. Additionally, you should worry about whether the sampling technique is genuinely unbiased since people who buy cars on credit are not necessarily a representative sample of all car buyers. The problem with income bracketing is not too serious since we can always use smaller brackets to reduce the degree of approximation, but we can do nothing about the sampling problem short of collecting more data from a different source.

3.5 The Regression Model

Remember the basic ideas behind statistical inference: We have a **population** of interest, and this population is characterized by some **population parameters** that we would like to know. We take a **sample** from the population, and **estimate** the parameters. Since any estimate is based on a sample, it will contain some **sampling error**, and we use probability theory to quantify that error, so we are able to produce confidence and prediction intervals and carry out hypothesis tests. For example, our population might be the adults living in Texas, and we may want to know the average amount they spend on dining out each year. The relevant population parameter is, therefore, the population mean, and we would estimate it from a sample by looking at the sample mean. For reasonable sample sizes, we know the sampling error is normally distributed around the true value, with standard deviation equal to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

where σ is the population standard deviation.

Regression analysis involves the same concepts; however, the population parameters are different, and we must be certain we understand exactly what they are. We will illustrate them using the Autorama example.

DIVIDING THE POPULATION BY INCOME LEVEL

When predicting auto purchases, we divided the population (our customers) into many sub-populations according to income. In other words, we did not think about the distribution of demand for all our customers but about the distribution for all customers with a given annual income.

Each of these sub-populations has different auto purchase patterns. For any given sub-population, a mean price exists that people in that population pay for a car. If we knew these means, we could see how the mean price varies across the different income brackets. A nice way to do so is by drawing a graph of mean price against income.

Regression Assumption 1. This graph would be a straight line.

Of course, this assumption may not be true. Later on, we will talk about how you can check the data to see whether this is a reasonable assumption for any particular data set, and what you can do if it is not.

Returning to our example, as a consequence of Regression Assumption 1, we may assume there are some constants, β_0 and β_1, such that for any given income level, the average price paid by people in that income level satisfies the following equation:

$$\boxed{\text{average price} = \beta_0+\beta_1(\text{income})}$$

β_0 is the intercept and β_1 the slope of the graph of average price against income as shown in Figure 3.10 below.

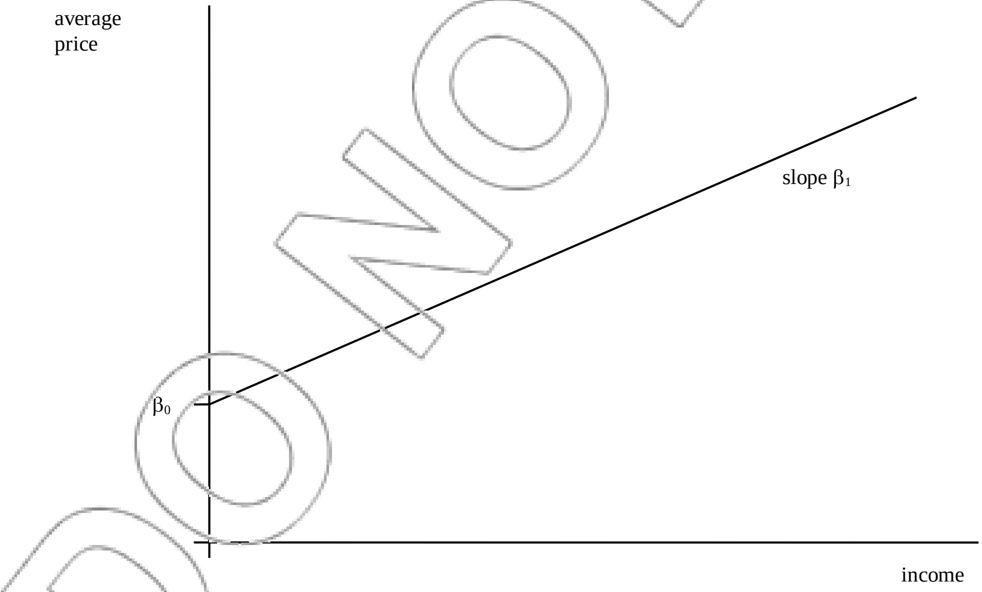
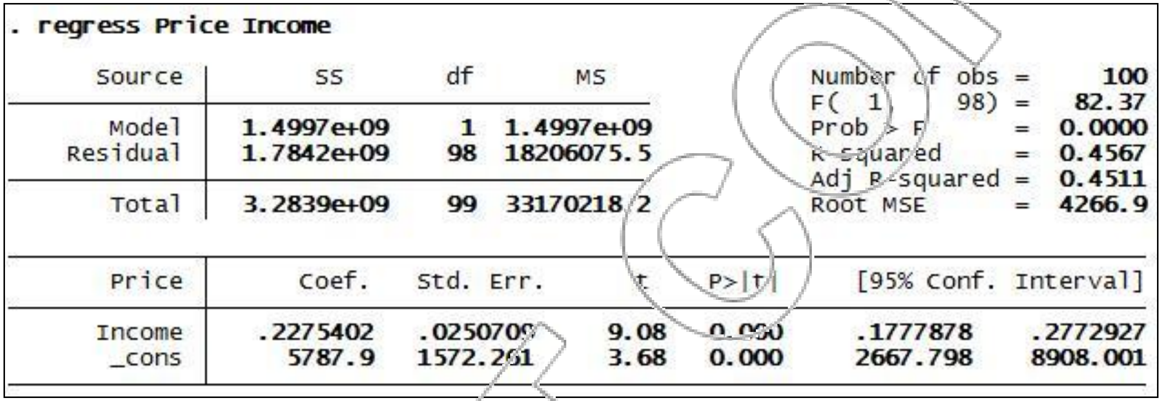


Figure 3.10: Regression line for price and income.

WHAT REGRESSION ESTIMATES

We can now talk about two of the population parameters regression estimates: They are the intercept and slope of this line, i.e., the constants β$_0$ and β$_1$. Look at the regression output in Figure 3.11.



```
. regress Price Income

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   82.37
       Model |  1.4997e+09     1  1.4997e+09           Prob > F      =  0.0000
    Residual |  1.7842e+09    98  18206075.5           R-squared     =  0.4567
-------------+------------------------------           Adj R-squared =  0.4511
       Total |  3.2839e+09    99  33170218.2           Root MSE      =  4266.9

------------------------------------------------------------------------------
       Price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Income |   .2275402   .0250709     9.08   0.000     .1777878    .2772927
       _cons |     5787.9   1572.261     3.68   0.000     2667.798    8908.001
------------------------------------------------------------------------------
```

Figure 3.11: Regression of price vs. income.

What is Stata providing here? Based on our sample, 5787.9 is the best estimate of the intercept β$_0$, and 0.2275 is the best estimate of the slope β$_1$. The constants β$_0$ and β$_1$ are the population parameters we would like to know, and the regression formulas that Stata implements give us estimates (often written b$_0$ = 5787.9 and b$_1$ = 0.2275) of the parameters. We can use these to estimate the average expenditure for any income group by substituting for income in the regression equation provided by Stata. This estimate is often written $\hat{y}$ and is referred to as a predicted value or a fitted value.

QUANTIFYING THE SAMPLING ERROR

These estimates are based on our sample, and if we had a different sample, we would get different estimates. Remember from Chapter 1, by thinking about the sample mean obtained from each possible sample we may obtain a sampling distribution, i.e., something like a histogram of all the possible sample means one would obtain from different random samples of the population. The same concept applies here, so we can talk about the sampling distributions of b_0 and b_1. We are less often interested in b_0, so we'll focus on b_1 and summarize what we have learned so far.

The idea is that when we compare two different groups of people, one of whom has an average income (say) $1,000 higher than the other, the difference between the average amount that the higher-income group will spend on an automobile and the average amount that the lower-income group will spend on an automobile is equal to $1,000 times some constant β_1. (This conclusion follows from the straight-line equation.) We do not know this constant, but we can use our sample to estimate it, which we do by taking the slope of Stata's best-fit line through our sample points. The estimate this produces is called b_1, and it is a random variable, i.e., the outcome of an experiment. If we repeated the experiment by taking a different sample, we would get a new estimate, and if we did this many times, we would get a distribution of estimates. The important things about this distribution are the following:

1. On average, b_1 is right, i.e., $E(b_1) = \beta_1$ (b_1 is called an **unbiased** estimator).
2. The distribution of b_1 has a standard deviation, written σ_{b_1}, which is estimated by Stata. We call this estimate s_{b_1}, and in this example, s_{b_1} = 0.02507. Stata reports this number in the **Std. Err.** column as seen in Figure 3.11.
3. We can generally assume that the distribution of b_1 is normal.

Regression analysis usually makes a number of other assumptions of varying importance in addition to the straight-line assumption. Later, we will discuss some of these other assumptions and talk about what happens when they are not satisfied. Nothing we have said so far in this section depends on the other assumptions, with one exception, noted below. For the time being, we will mention one of these assumptions, which you have seen in the previous section:

> Regression Assumption 2. The standard deviation of price
> for each income group has the same value, σ.

Of course, σ is another unknown population parameter. Stata produces an estimate of σ in the regression output, which is denoted by s. Stata prints the value of s in the row labeled **Root MSE** (here, s = 4266.9). The units for this estimate are the same as the units for your dependent (y) variable. Here, s = \$4266.90. The formula Stata uses to get s_{b_1} (= .02507 here) makes use of this s, so point 2 in the box above does depend on this assumption.

Do not confuse σ, the standard deviation of price for each income group, with σ_{b_1}, the standard deviation of our estimate of β_1. In Chapter 1, we made the same distinction between the population standard deviation, σ, and the standard deviation of the sample mean, $\sigma_{\bar{x}}$.

CONFIDENCE INTERVALS ON THE REGRESSION COEFFICIENTS

We can use our knowledge of the sampling distributions to make statistical inferences, i.e., to form confidence and prediction intervals and carry out hypothesis tests with our regression results.

Since b_1 is distributed normally, we know

$$\frac{b_1 - \beta_1}{\sigma_{b_1}}$$

has the standard normal distribution. So, for example, we can be 95% confident that β_1 is within ±1.96 standard deviations of b_1. In other words, $b_1 \pm (1.96)\,\sigma_{b_1}$ forms a 95% confidence interval for

β_1. We do not know σ_{b_1} , so we use our estimate s_{b_1} instead and must use a t-distribution instead of the standard normal. The general formula for a 100(1-α)% confidence interval for β_1 is the following:

$$b_1 \pm t_{\alpha/2,n-2}\, s_{b_1}$$

$t_{\alpha/2,n-2}$ is the $\alpha/2$ t-value with n-2 degrees of freedom (where n is the sample size) (**display invttail**(n-2, $\alpha/2$)). Later, you will see that Stata output tells you how many degrees of freedom to use, so you have one less thing to worry about.

For example, try to produce a 90% confidence interval for β_1. If you look at the last regression output, you will see the sample size was 100, so we have 98 degrees of freedom. (This value is given directly in the **Residual** row and the **df** column.) The 90% confidence interval is $.2275 \pm t_{.05,98}(.02507)$ = .2275±invttail(98, 0.05)(.02507) = .2275±(1.6606)(.02507) = .2275±.0416 = (.1859, .2691). The interpretation is that 90% of the time we take a sample of size 100 and use it to calculate an interval according to the formula, the interval will contain the true slope, β_1. We are therefore 90% confident (.1859, .2691) contains the true slope, β_1.

HYPOTHESIS TESTS ON THE REGRESSION COEFFICIENTS

Suppose the common industry wisdom is people will spend on average an extra $180 on their new auto for every extra $1,000 in income. In terms of our regression model, this says that the true slope, β_1, of the regression line is 0.180. (Make sure you understand why.) Our estimate seems to be higher than this, but is the difference large enough to indicate strong evidence that the true slope is higher than 0.180? If it is, then we might want to re-evaluate the common wisdom.

Therefore, we would like to know if our estimate is statistically significantly greater than 0.180. We test this by the following hypothesis test:

$$H_0: \beta_1 \leq .180$$

$$H_a: \beta_1 > .180$$

We will follow the usual hypothesis-testing procedure. Give the benefit of the doubt to the null hypothesis by assuming that $\beta_1 = .180$. Under this assumption, we know our test statistic, t, follows the t-distribution with 98 degrees of freedom:

$$t = \frac{b_1 - .180}{s_{b_1}}$$

Using Stata's numbers, we have $t = (.2275-.180)/.02507 = 1.895$. To determine the p-value, use Stata's **ttail** command remembering that we are conducting a one-tailed test and want the area in the upper tail. This command (**display ttail(98, 1.895)**) yields a p-value of 0.0305. This tells us that if the null hypothesis were true, there would only be about a 3% chance that a sample of size 100 would give an estimated slope as large as ours here. So, unless we want to be particularly careful about making a type I error (i.e., unless we want to set our level of significance, α, at less than .03), we will reject the null and conclude that our results do shed doubt on the conventional wisdom and strongly suggest that for every additional $1,000 of income, people spend more than an additional $180 when they buy a new car.

READING SIGNIFICANCE IN THE REGRESSION OUTPUT

We can now explain the **t** and **P>|t|** columns of Stata's regression output. Consider the following (two-tailed) hypothesis test:

$$H_0\text{: } \beta_1 = 0,$$

$$H_a\text{: } \beta_1 \neq 0.$$

The relevant test statistic would be $t = \frac{b_1 - 0}{s_{b_1}}$ = .2275/.02507 = 9.075. This is so large that the corresponding p-value is 0.000. If you look back at the regression output in Figure 3.11, you will see Stata has done this calculation for us: In the **Income** row that tells us about b_1, the column labeled **t** contains the test statistic, and the next column labeled **P>|t|** contains the p-value. Similarly, if we wanted to test whether or not the true intercept, β_0, is equal to 0, we can look in the **_cons** row to find the p-value for the test where $\beta_0 = 0$ is the null hypothesis (which we reject since p = 0.000).

Traditionally, people have been especially interested in testing coefficients against zero because they often use regression to test if one variable has any effect on another. In this case, saying that $\beta_1 = 0$ means that income has no effect on the price people pay for cars. Since the test is so commonly used, Stata and any other standard statistical package reports it automatically. Typically, we will be able to determine what affects what but we also need to know the effect's size. The example here illustrates that nicely. Rejecting the null in this automatic hypothesis test allows us to conclude that your income affects how much you spend on a car. This is not very profound. However, we do care that the extent of this effect is larger than the conventional wisdom. In other situations, we may be interested in small non-zero effects. (For example, in finance, tiny effects can provide arbitrage opportunities that are important.)

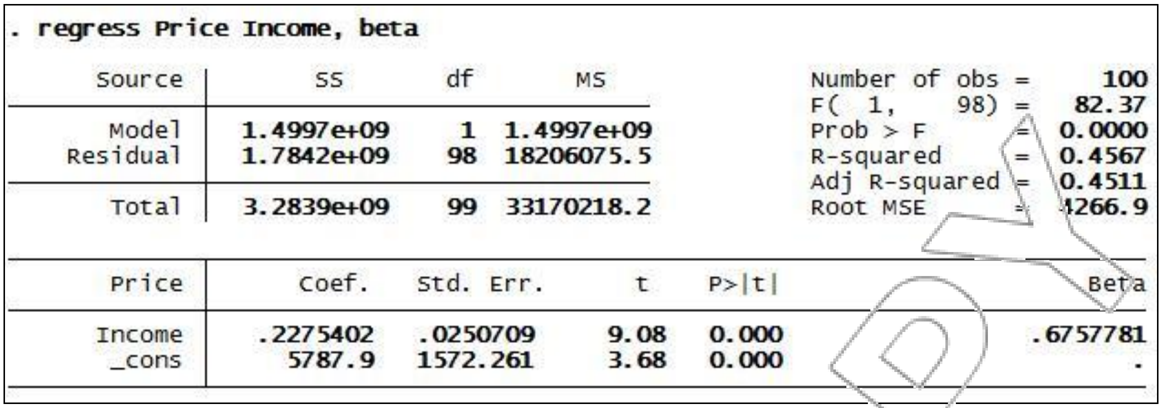
The usual terminology is to say that the estimate b_1 (or the variable income) is statistically significant at the α level if the two-tailed test of β_1 against zero leads to a rejection of the null hypothesis (at the α level of significance). Remember that all this indicates is that we have evidence that we have a non-zero coefficient. If we want to test against any other value as we did earlier with 0.18, we will have to calculate the test statistic and p-value for ourselves, as in the previous section.

Finally, you may wonder why Stata reports both the test statistic and the p-value for the test. The answer is that some people like to know the test statistic. However, the p-value contains all the information you need.

OVERVIEW OF THE REGRESSION OUTPUT TABLE

It may help you to go through part of the regression output again. After running a regression, Stata produces a table as shown in Figure 3.12.[5]

[5] We have included an option to generate an additional column labeled **Beta** so that we may explain what it means. As you can see from the output, the **Beta** column is obtained by typing the command **regress Price Income, beta**. Alternatively, navigate to **Users>Core Statistics>Regression(regress)**, click on the **Reporting** tab, and check the box next to **Standardized beta coefficients**.



```
. regress Price Income, beta

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =   82.37
       Model |  1.4997e+09     1  1.4997e+09           Prob > F      =  0.0000
    Residual |  1.7842e+09    98  18206075.5           R-squared     =  0.4567
-------------+------------------------------           Adj R-squared =  0.4511
       Total |  3.2839e+09    99  33170218.2           Root MSE      =  4266.9

------------------------------------------------------------------------------
       Price |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      Income |   .2275402   .0250709     9.08   0.000                 .6757781
       _cons |     5787.9   1572.261     3.68   0.000                        .
------------------------------------------------------------------------------
```

Figure 3.12: Partial regression output.

We will go through this table now. Recall that the regression estimates the coefficients of a straight line. These coefficients are the intercept β_0, and the coefficient on the income variable, β_1. The row labels tell us which of these coefficients each row concerns. Thus, the **_cons** row is concerned with the constant coefficient or intercept, β_0, and the **Income** row is concerned with the coefficient on the income variable, β_1. The **Coef.** column contains the actual estimates of these coefficients (b_0 = 5787.9, b_1 = 0.2275). The **Std. Err.** column is more interesting. Each of the coefficient estimates is subject to sampling error and has a distribution whose standard deviation we can estimate. Those estimates are found in this column. For example, we know that b_1 is normally distributed, its expected value is the true slope β_1, and we can estimate its standard deviation to be s_{b_1} = .02507. Similarly, the estimated standard deviation of b_0 is denoted by s_{b_0} (= 1572.261). The next two columns tell us the results of specific hypothesis tests. There is one test for each estimator. The null hypothesis is that the true value of the parameter we want to estimate is zero. The **t** column tells us the test statistic value we obtain from this test, and the **P>|t|** column tells us the corresponding p-value. The **Beta** column tells us the beta weight corresponding to income. The beta weights are coefficients of a regression where, instead of the variables themselves, standardized versions of the variables are used. Looking at the regression output table in Figure 3.12, we see the beta weight on income is 0.6758. This tells us that, on average, for a one standard deviation increase in the income, price will increase by 0.6758 standard deviations of price.

SUMMARY

We looked at how Stata chose the best-fit line through a scatterplot and how to use the equation of that line to make predictions. We applied this to predict the average price of a car bought by customers with a given income.

We assumed that, for any given income level, the amount spent on a car is normally distributed and the standard deviation of that distribution is the same for each income level. The mean of that distribution is the estimate provided by the regression equation, and the regression also provides the estimated standard deviation.

We used this information, together with some demographic data on our customer base, to predict the overall distribution of car purchases. We divided our customers into income brackets and our cars into price brackets. For each income bracket, we worked out how many of our customers would come from that bracket and how their purchases of cars would fall among the different price brackets. This told us how many cars would be sold in each price bracket by adding up how many cars would be sold in that bracket to people in the lowest income bracket, the second lowest, etc.

We examined the regression model and learned that regression studies how one variable (e.g., auto price) varies across different populations indexed by another variable (e.g., income). It assumes that this relationship is linear on average and estimates the linear relationship. We can use this estimate to make predictions and use statistical theory to perform inferences about those estimates, including confidence intervals and hypothesis tests.

NEW TERMS

Best-fit line The line generated by the least squares method that produces the smallest possible sum of squared errors

Unbiased estimator An estimator whose expected value is equal to the parameter it estimates

Residual degrees of freedom The number of data points in a regression minus the number of coefficients (including the constant). This is used to calculate the proper t-statistic to use in confidence intervals and is used in calculating p-values of hypothesis tests

Standard error of the regression (s) An estimate of σ, the standard deviation of the dependent variable (y) given (or conditional on) a fixed value of the independent variable (x)

NEW FORMULAS

$100(1-\alpha)\%$ confidence interval for β_1:$b_1 \pm$ **invttail**(n-2, $\alpha/2$)* s_{b_1}

$100(1-\alpha)\%$ confidence interval for β_0:$b_0 \pm$ **invttail**(n-2, $\alpha/2$)* s_{b_0}

Hypothesis test to see if coefficient k is statistically significant:

$$H_0: \beta_k = 0,$$

$$H_a: \beta_k \neq 0.$$

NEW STATA FUNCTIONS

User>Core Statistics>Bivariate Statistics>Bivariate Plots (twoway)

Equivalently, you may type **db twoway**. This command opens the Stata **twoway** dialog box, where you can create various types of graphs including scatterplots and best-fit lines.

To generate a scatterplot, click **Create…**. Choose **Basic plots** as your plot category and choose **Scatter** as your plot type. Select the appropriate X and Y variables and click **Accept>OK**. Once Stata generates the graph, you can open Stata's Graph Editor to make revisions to your graph.

Alternatively, you can directly type the command **twoway scatter *varY varX***.

To graph a best-fit line, click **Create…**. Choose **Fit plots** as your plot category and choose **Linear prediction** as your plot type. Select the appropriate X and Y variables and click **Accept>OK**.

Alternatively, you can directly type the command **twoway lfit *varY varX***.

If you want to add a best-fit line on top of a scatterplot for variables X and Y, you can click **Create…** again to create Plot 2 (Plot 1 is your scatterplot) for your best-fit line by following the steps above. The direct command is **twoway (scatter *varY varX*) (lfit *varY varX*)**. For more graphing options, type **help graph** into the Command box.

User>Core Statistics>Regression (regress)

Equivalently, you may type **db regress**. This command opens the regression dialog box asking you to select a dependent variable from a list of all variables in the current data worksheet. You are asked to choose one (or more) independent variables from the “Independent variables” list. Clicking **OK** will produce the regression output. Stata reports the estimated coefficients (under the **Coef.** column), estimated standard deviations of the coefficients (**Std. Err.**), test statistics for the coefficients (**t**), and p-values (**P>|t|**) for a two-tailed test with the null hypothesis that the true value of the parameter of interest is zero. You can find the appropriate degrees of freedom with which to manually calculate confidence intervals for the parameter of interest in the **Residual** row and the **df** column. In addition, you can find the **standard error of the regression** in **Root MSE**.

Alternatively, you can directly type the direct command **regress *depvar indepvars***, where ***depvar*** corresponds to the name of the dependent variable, and ***indepvars*** correspond to the name(s) of the independent variable(s).

If you want Stata to report the beta-weight(s) for independent variable(s), you can either check the “Standardized beta coefficients” box under the Reporting tab in the regress dialog box, or you can type the direct command **regress *depvar indepvars*, beta**.

CASE EXERCISES

1. A little peek at the Autorama

Consider the 80.5 expected car buyers in the $45,000–$55,000 income group from the Autorama case. Using the same assumptions we made in the chapter, determine the expected number of sales in each price bracket from this group. Hint: Use the normal distribution to determine the probability that someone in this group would buy a car in each price bracket. You may wish to do your calculations in a spreadsheet.

2. Autorama: The big picture

Take the entire set of potential car buyers at the Autorama described in Figure 3.1 and complete the objectives of the case. That is, using a spreadsheet, determine the number of cars to stock in each price bracket at your new dealership to match demand. You can do this in three steps:

a. For each income group, determine the expected number of sales within each price bracket.
b. Sum the expected sales with each price bracket to determine the total expected sales within each price bracket.
c. Multiply the fraction of total expected sales in each price bracket times 200 to determine how many of each type of car to stock in your inventory.

3. Shore Realty

Shore Realty sells real estate in Oklahoma. The company would like to be able to predict the selling price of new homes based on the home's size. It has collected data on size ("sqfoot" in square feet) and selling price ("price" in dollars) which are stored in the file **shore**. Use the data in that file to answer the following questions:

a. Use a computer to construct a scatterplot for these data with size on the horizontal axis.
b. Determine the estimated regression equation.

c. Predict the selling price for a home with 2,600 square feet.

PROBLEMS

For problems 1–3, you will need to access the **bschools2002** file, which contains data regarding the top 30 business schools based on the 2002 *Business Week* ratings.

1. Many business school surveys including this one report **mean base salaries** and **median base salaries**. These two statistics tend to be similar. Stata can help us find a relationship between the two for this dataset.

a. Construct a scatterplot with mean base salary on the vertical axis and median base salary on the horizontal axis.

b. Does this relationship appear linear?

c. Perform a regression of mean base salary vs. median base salary. Write out the estimated regression equation.

d. Use your regression equation to estimate the mean base salary for a school with a median base salary of $77,000.

e. Use your regression equation to estimate the mean base salary for a school with a median base salary of $88,000.

2. Students from better schools might command a higher salary. Comparing a school's mean base salary to its rank might help us understand this relationship.

a. Develop a scatterplot for these variables with mean base salary as the dependent variable.

b. Does this relationship appear linear?

c. Perform a regression of mean base salary vs. rank. Write the estimated regression equation.

d. Use your regression equation to estimate the mean base salary for a school ranked eighth.

e. Use your regression equation to estimate the mean base salary for a school ranked 25th.

f. Use the coefficient on the rank variable to estimate the expected increase in mean base salary from a one-unit improvement in a school's rank. Provide a 95% confidence interval for your estimate.

g. How confident are you that the true slope, β_1, is significantly different from zero?

3. Schools with larger enrollments might have more resources, making their students better prepared and more valuable to employers and, subsequently, commanding a higher salary. Of course, smaller schools may give students more personal attention, which develops better skills and could yield a higher salary for smaller schools. Studying the relationship between mean base salary and enrollment might help us understand this relationship better.

a. Develop a scatterplot for these variables with mean base salary as the dependent variable.

b. Perform a regression of mean base salary vs. enrollment. Write the estimated regression equation.

c. Use your regression equation to estimate the mean base salary for a school that enrolls 800 students.

d. Use your regression equation to estimate the mean base salary for a school that enrolls 1,800 students.

e. Use the coefficient on the enrollment variable to estimate the expected increase in mean base salary as enrollment increases by one. Provide a 95% confidence interval for your estimate.

f. Is the true slope, β_1, significantly different from zero? Use a 5% level of significance.

g. Is the true slope, β_1, significantly greater than zero? Use a 5% level of significance.

4. The top-selling beer in the world is Budweiser, which is produced by Anheuser-Busch. The company's annual reports provide the data in the file **budsales**, which presents 12 years of combined sales (in 31-gallon barrels) of all Anheuser-Busch beers. The file also contains information on the U.S. population (US Pop) based on census estimates.

a. Develop a scatterplot for these data with **barrels sold** as the dependent variable and **US Pop** as the independent variable.

b. Perform a regression of barrels sold vs. US Pop. Write the estimated regression equation.

c. What does the coefficient of the variable US Pop represent in this regression equation? Be specific and clear in your answer.

5. Access the file **taxfranchise**. The data come from a regional tax preparation company with 19 locations across the Midwest. The first variable measures the Output per Worker in terms of customers' tax forms completed per month, and the second is the annual Computer Budget per employee at that location.

a. Construct a scatterplot of Output per Worker vs. Computer Budget.

b. Perform a regression of Output per Worker vs. Computer Budget and write the estimated regression equation.

c. Use the regression equation to estimate the Output per Worker at a location with a Computer Budget of $2,500 per employee.

d. Use the regression equation to estimate the Output per Worker at a location with a Computer Budget of $3,500 per employee.

e. Use the coefficient on the Computer Budget variable to estimate the additional number of tax forms completed per month for each one-dollar increase per employee in the Computer Budget.

f. Provide a 90% confidence interval for your answer to part e.

g. Using $\alpha = 0.05$, determine if the Computer Budget is significantly related to Output per Worker.

CHAPTER 4

BETAS AND THE NEWSPAPER CASE: USING THE REGRESSION EQUATION

In this chapter, we will learn more about regression and how to use it to make predictions. We will see one of the major applications of regression in finance, the estimation of asset betas, which are numbers measuring the riskiness of different investments. Then, we will explore a new product start-up problem in the newspaper industry. Along the way, we will learn to do statistical inference with regression: In addition to producing estimates, we will be able to say something about the accuracy of our predictions through the use of confidence and prediction intervals and hypothesis tests.

4.1 Capital Budgeting and Risk

How to deal with risk in capital budgeting is one of the central issues in modern finance theory. Some of you may have encountered, at work or in finance classes, many of the concepts covered in this section. We will concentrate on the use of regression to measure asset betas. These numbers measure the riskiness of different assets, forming the basis for the most widely used approach to capital budgeting under uncertainty, the Capital Asset Pricing Model (CAPM). This section should explain enough to enable you to appreciate the importance of asset betas and how to use them in simple examples. You may wish to supplement this section by reading the relevant sections of any standard finance text, such as *Principles of Corporate Finance*, by Brealey and Myers.[1] There, you will learn about some factors we have ignored here, most notably the relation between capital structure and the cost of capital.

CAPITAL BUDGETING AND THE OPPORTUNITY COST OF CAPITAL

Suppose your company has the opportunity to begin a new project. The project will take place over two years. In year 1, you will have to invest $10 million, and in year 2, you expect average returns of $11.5 million, after which the project will end. Should you undertake the project?

The answer is you should undertake the project if it has positive net present value (NPV). If you are unfamiliar with the concept of NPV, Brealey and Myers or any other finance text will cover it in detail. For a given interest rate or cost of capital, r, the net present value is given by the following:

[1] *Principles of Corporate Finance*, 7/e. Richard A. Brealey and Stewart C. Myers. McGraw-Hill, 2003.

$$-10,000,000+\frac{11{,}500,000}{1+r}$$

You can check that the NPV will be positive if *r* is smaller than 15% (.15) but will be negative otherwise. This means, you should invest if your cost of capital is less than or equal to 15 percent. This makes intuitive sense: If you make the investment, you will get a return of 15% in exchange for having your capital tied up for one year. The cost of capital refers to how much it costs you to have your capital tied up for one year, which is the rate of return you could achieve with it. Since this investment pays 15%, you should make it if you cannot earn more than 15% elsewhere. Here, “you” means your shareholders, since as a corporate manager it is your job to maximize shareholder value.

Risk and return

However, things are more complicated once we recognize the role of risk. Assume that this investment is somewhat risky, so the return of $11.5 million is uncertain and is merely your best estimate. Your shareholders need to be compensated for bearing that risk. To determine their cost of capital, we need to see how much they could get for bearing the same risk in a different investment. Again, this makes sense: Risky investments pay more on average than safe ones, but that does not mean that you should automatically choose the riskiest investments you can find. In practice, what it means is that you need to know how high a return your shareholders need to be compensated for bearing the risk your project represents.

The CAPM formula

This brings us to the bottom line of the CAPM: What it says is that the riskiness of a project or asset can be measured by a single number, known as the beta (β), and the required rate of return satisfies the following formula:

$$r-r_f = \beta(r_m-r_f)$$

Here, r_f is the risk-free interest rate, i.e., the return on a totally safe asset, and r_m is the return on a market portfolio. We usually think of r_f as the return on US Treasury bills (T-bills), which historically have paid about 3.5%, and we often think of r_m as the return on the S&P500 index, which has earned a much higher return (around 12%.) The difference between the return on any asset and the risk-free return is called the excess return, so the formula says that the excess return on any asset should be proportional to the excess return on the market as a whole, with the constant of proportionality equal to β. If we know the beta of our project, we can use this formula to learn the correct cost of capital, r, calculate the NPV, and decide whether to make the investment.

Measuring Risk I: unique vs. market risk

When we think about a project's riskiness, we have to distinguish between two different kinds of risk: unique risk and market risk. Rational investors will hold a well-diversified portfolio of investments, with money in the stocks of many companies. This may be justified by the principle of not putting all your eggs in one basket. On a more technical level, as we saw in Chapter 1, when you take the average of a number of independent random variables, the standard deviation of that average becomes low. For instance, if you have *n* independent risks, each with the same standard deviation σ, then their average has a standard deviation of $\sigma/\sqrt{n}$. What this means is that investors do not have to worry much about the risk of any single investment, provided that risk is independent of their other investments' risks. For example, one risk facing Hewlett-Packard is if Dell will continue to steal market share away. However, whether or not that happens is mostly independent of anything else that might happen in the world, which suggests that a well-diversified investor should not have to worry about it. That is an example of a unique risk, also known as a specific risk, unsystematic risk, or diversifiable risk.

On the other hand, suppose that the economy slides into a recession. Hewlett-Packard's sales will fall and so will those of most other corporations in the United States. So, the risk of a recession is undiversifiable because any companies in which we invest will face the same risk. The companies' risks are not independent and their eggs are all in the same basket. This kind of risk is called market risk, systematic risk, or undiversifiable risk. Since investors cannot avoid this risk by diversification, they have to be compensated for bearing it. Because some companies face more such risk than others, they must offer a higher return to interest people in investing in them. For example, during recessions, people often put off buying cars, but such economic conditions do not greatly affect their use of the telephone, so auto companies have more market risk associated with them than do telephone companies.

Measuring Risk II: defining beta

Now that we know the kind of risk to measure, what remains is learning how to measure it. We can get at the right measurement by thinking about how the share price of an auto company like Ford will vary with the market as a whole, compared to that of a telecommunications company like AT&T. It turns out that when the market is doing well, Ford's shares do extremely well, but when the market is doing badly its shares do very badly. This is what you would expect: When the economy is booming, the markets are up and many people buy new cars, but when things are slow, few people do. Between 1984 and 1989, Ford's shares went up/down by 1.3% for every 1% change up/down in the market as a whole (on average and after subtracting the risk-free rate). This number (1.3) gives us a measure of how much risk is involved in holding Ford's shares. By comparison, AT&T shares were safer than the stock market as a whole during this time, with a change of only 0.76% for each 1% change in the market. These numbers are what we call the betas of the assets.

Beta measures the amount the stock price changes for a 1% change in the market as a whole.

In the next section, we will see how regression is used to calculate/estimate betas. What this section has explained is how to use the beta to make investment decisions.

Summary

We reviewed the following procedure for deciding when to make a risky investment.

1. Obtain a numerical measure of the riskiness called its beta. (We defined the beta, but did not explain how to measure it.)
2. Use the CAPM formula, $r\text{-}r_f = \beta(r_m\text{-}r_f)$, to obtain the appropriate cost of capital figure r.
3. Use that value of r to calculate the NPV.
4. Make the investment if it has positive NPV.

Implementing step 1 will be discussed in the next section.

4.2 Estimating Betas

The firm you work for owns a chain of upscale pizza restaurants in New England. Your CEO believes the lower end of the market has room for expansion and wants to start up a large chain of fast food pizza places to compete in the fast food market. You are asked to make a preliminary study of the advisability of this investment, based on an initial investment of $8 million and projected average annual profits starting at $1 million and increasing to $2 million after the first two years. You write out the NPV formula for these figures (all in $millions):

$$\mathrm{NPV} = -8 + \frac{1}{(1+r)} + \frac{1}{(1+r)^2} + \frac{2}{(1+r)^3} + \frac{2}{(1+r)^4} + \frac{2}{(1+r)^5} + \blacksquare$$

Before you can calculate this sum, you need to know the relevant discount rate, *r*. The projected profits are estimates since the true profits are uncertain, so this is a risky investment and the discount rate must reflect this riskiness. As you know from the previous section, the correct rate for discounting uncertain cash flows is given by the CAPM formula:

$$r\text{-}r_f = \beta(r_m\text{-}r_f)$$

Suppose the current risk-free rate is 4.0%, and the expected excess return of the market is 8.0%; so, $r = .04+.08\beta$. All you need is a figure for beta, which measures the riskiness of this kind of investment.

ESTIMATING BETA

Fortunately, you have data on the share performance of some other companies, which operate in the fast food market. Since they are in the same business as this project, their riskiness should be a good guide to the proposed investment's riskiness. You decide to use regression to estimate the beta from these data, which consist of the monthly excess returns of (among others) the shares of McDonald's Corporation and of a portfolio representing the market as a whole. These data are contained in the **stocks** file;[2] the data on monthly excess returns are reported in percentages.

How does regression help us here? The definition of the beta tells us how much the share price moves (relative to the risk-free rate), on average, compared with a 1% move (relative to the risk free rate) in the market as a whole. So, if we draw a scatterplot of the monthly excess returns of McDonald's (stored in the MACS column) against the monthly excess returns of the market portfolio (stored in the column MARKET) and plot a best-fit line, the slope of the line should tell us the beta.[3]

As you can see, the line is fairly steep. The regression equation (MACS = 0.253+1.458*MARKET, from Figure 4.2) tells us that the beta is estimated to be about 1.458; so, on average, a 1% change in the market is associated with a change of almost 1.5% in the value of McDonald's shares.[4] We can use this estimate to get the discount rate: r = .04+.08(1.458) = .1566, which is, about 15.7 percent. Substituting this value into the NPV calculation (and doing

[2] Derived using data from *The Center for Research in Security Prices* at http://gsbwww.uchicago.edu/research/crsp/.

[3] We also obtain the intercept of the best-fit line, usually called the asset's alpha. The estimated alpha shows our best estimate of the excess return of the given asset if the market excess return were 0. According to the CAPM equation, the intercept should be 0 (verify this for yourself).

[4] The estimated intercept is about .25% monthly, or 3% annually. In practice (finance), the intercept is usually omitted when computing the asset's expected excess return. That is, the estimated beta is plugged into the CAPM formula as if the constant estimate were zero. We do it the same way below.

some algebra), we find this gives a profit of about $3.1 million, so this investment seems to be a good idea. Before jumping to conclusions, we should check the accuracy of our beta estimate since if it is rough we cannot be too confident that the NPV is positive. The true beta might be a lot higher than our estimate, leading to a much higher discount rate, which could tip the project into unprofitability.

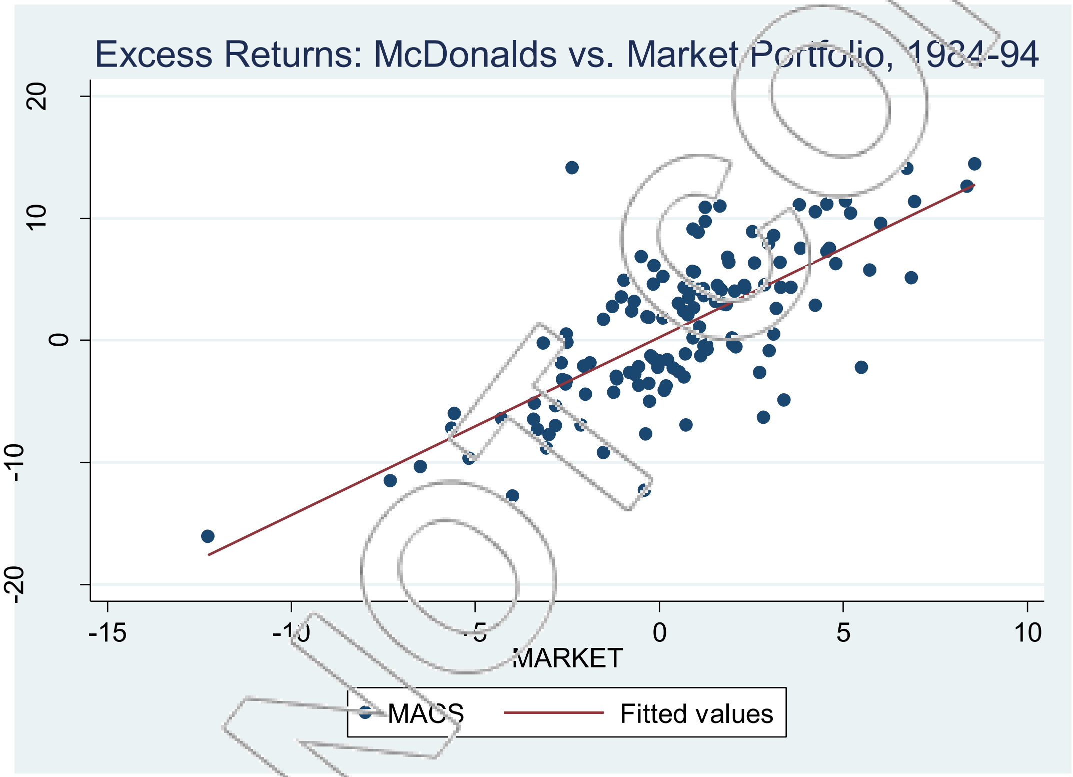
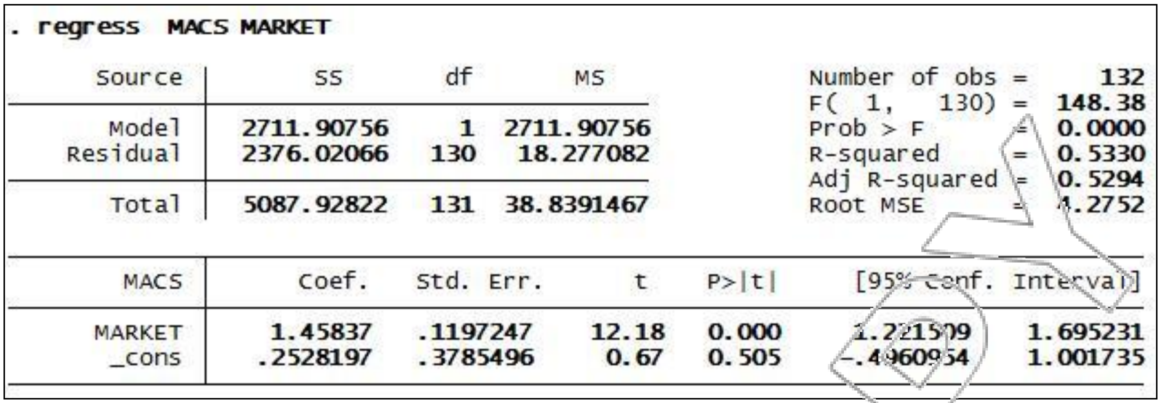


Figure 4.1: Excess returns of McDonald's vs. the market portfolio.

CONFIDENCE INTERVAL FOR BETA

To produce a confidence interval, we need an estimate (commonly called the **standard error of the coefficient**) of the standard deviation of our estimate of beta, which we can find in the regression output from Stata.



```
. regress  MACS MARKET

      Source |       SS       df       MS              Number of obs =     132
-------------+------------------------------           F(  1,   130) =  148.38
       Model |  2711.90756     1  2711.90756           Prob > F      =  0.0000
    Residual |  2376.02066   130   18.277082           R-squared     =  0.5330
-------------+------------------------------           Adj R-squared =  0.5294
       Total |  5087.92822   131  38.8391467           Root MSE      =  4.2752

------------------------------------------------------------------------------
        MACS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      MARKET |    1.45837   .1197247    12.18   0.000     1.221509    1.695231
       _cons |   .2528197   .3785496     0.67   0.505    -.4960954    1.001735
------------------------------------------------------------------------------
```

Figure 4.2: Regression of MACS vs. market.

The estimated standard deviation of b_1, which is listed under the **Std. Err.** column, is 0.1197, and we know that our sample size is 132 (11 years of monthly data). Therefore, a 90% confidence interval is given by 1.4584±(0.1197)$t_{0.05,130}$, and $t_{0.05,130}$ is about 1.657 (using the command **display invttail(130, 0.05)**). This is 1.4584±0.1984 = (1.26,1.6568). We are interested in using beta to determine the discount rate using the CAPM equation $r = .04+.08\beta$, so we can turn this into a confidence interval for r. That is, we have a beta between 1.26 and 1.6568 with 90% confidence, so we can say with 90% confidence that r is between .04 + .08(1.26) and .04 + .08(1.6568), i.e., between .1408 and .1725.

We can do a kind of worst-case analysis using this interval as follows. Suppose we have seriously underestimated beta. The true discount rate will be much higher than the 15.7% figure we used above. We can use the confidence interval to produce a sort of upper bound on the true discount rate's size: We do not know the exact value, but we are fairly (95%) confident that it is no more than 17.25%. If we repeat the NPV calculation using r = 17.25%, we will get a figure of $2.01 million, which is still positive by a wide margin. The project will be profitable even under a worst-case assumption where the appropriate discount rate is much higher than the one used. The confidence interval enables us to choose a number we may treat as our worst-case scenario.

You might wonder why we did not do a hypothesis test here. The answer is that we could have done that, but you would have needed to work out the relevant test. Finance theory suggests that the appropriate thing to prove (i.e., the thing you should use as the alternative hypothesis) is whether the true discount rate is less than the internal rate of return (IRR). You could calculate this IRR with Excel or a financial calculator (it turns out to be about 21%) and carry out the hypothesis test.

To prove the true discount rate r is less than 21%, the appropriate hypothesis test is the following:

$$H_0: r \geq 0.21$$

$$H_a: r < 0.21$$

The next step is to use the CAPM formula, $r\text{-}r_f = \beta(r_m\text{-}r_f)$, to rewrite the hypotheses in terms of beta. Using $r_f = 0.04$ and $r_m = 0.12$, the alternative hypothesis becomes $0.04+0.08\beta < 0.21$ or, rearranging, $\beta < 0.17/0.08 = 2.125$. So, the hypothesis test is the following:

$$H_0: \beta \geq 2.125$$

$$H_a: \beta < 2.125$$

Using the results from the regression, we can calculate a test statistic of t = (1.458-2.125)/0.1197 = -5.57. Therefore, the p-value =**1-ttail(130, -5.57)** ≈ 0, and we can reject the null hypothesis. We are extremely confident the true discount rate is less than the IRR, meaning this is a profitable project. This is the same conclusion arrived at using confidence intervals above.

Summary

We used regression to estimate the beta of McDonald’s from a sample of excess returns on its shares and on the market as a whole. We used this beta to estimate the correct discount rate for a capital budgeting problem. We used a confidence interval approach to get a range of possible values for the beta and did a worst-case analysis to check whether the proposed investment would be profitable under rather pessimistic assumptions about sampling error. We also carried out a similar analysis using hypothesis testing.

4.3 Predicting Circulation

A newspaper in a large metropolitan area is thinking about issuing a Sunday edition. Management estimates that this would involve a start-up cost of $2 million and fixed annual operating costs of $1 million. Once the project is up and running, it figures a profit (net of the marginal costs of printing and distribution) of $5 per reader per year. Therefore, if the newspaper gets X thousand readers, it will realize an annual profit of $(5,000X-1,000,000) in perpetuity. The cost of capital for this industry is 15%, so the NPV of this profit stream is the following:[5]

$$\begin{aligned}
&(5{,}000X - 1{,}000{,}000)\left(\frac{1}{(1+.15)} + \frac{1}{(1+.15)^2} + \frac{1}{(1+.15)^3} + \dots\right) \\
&= \frac{(5{,}000X - 1{,}000{,}000)}{.15} \\
&= \$(33{,}333X - 6{,}666{,}667)
\end{aligned}$$

If readership is low, this value will be negative, but even if it is positive, it has to outweigh the initial cost of $2 million. In other words, the project is a good one if the following occurs:

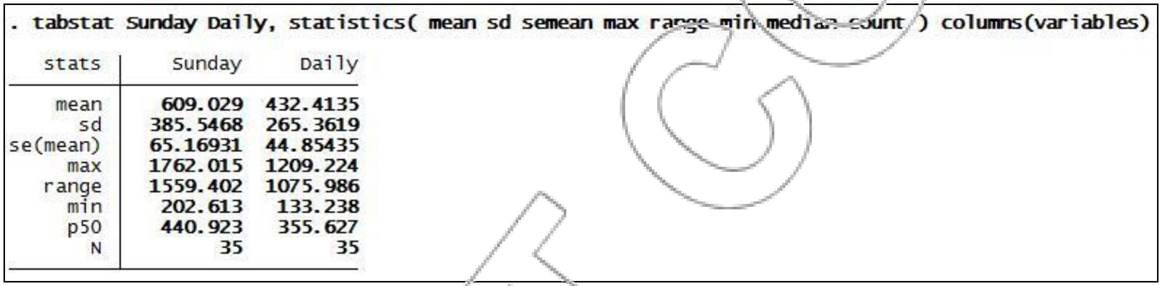
$$33{,}333X - 6{,}666{,}667 > 2{,}000{,}000$$

This is true whenever X is greater than 260. So, the break-even figure is a circulation of 260,000.

[5] Using the perpetuities formula, which says that the value of $1 per year in perpetuity is $(1/*r*). If you have not seen this, you can read about it in any standard finance text.

THE DATA

This projection is useless unless you can forecast what circulation will be. We will use regression to produce such a forecast, based on a data set called **newspapers**, which consists of the daily and Sunday circulation figures for 35 newspapers in other cities around the country.[6] The daily circulation of the paper is 190,000. We will use these data to forecast Sunday sales and to assess the forecast's accuracy. We begin with a preliminary look at the data, first via descriptive statistics and then graphically on a scatterplot.[7] All figures are in thousands.



```
. tabstat Sunday Daily, statistics( mean sd semean max range min median count ) columns(variables)

   stats |    Sunday     Daily
---------+--------------------
    mean |   609.029  432.4135
      sd |  385.5468  265.3619
se(mean) |  65.16931  44.85435
     max |  1762.015  1209.224
   range |  1559.402  1075.986
     min |   202.613   133.238
     p50 |   440.923   355.627
       N |        35        35
------------------------------
```

Figure 4.3: Univariate Statistics for Sunday and daily circulation.

Examine the data in the scatterplot shown in Figure 4.4:

[6] Derived from Hedblad, Alan, ed. ***Gale** Directory of Publications and Broadcast Media*, Gale Research, 1992.

[7] To generate univariate statistics for Sunday and daily only, click **User>Core Statistics>Univariate Statistics>Custom (tabstat)** or type **db tabstat**, select **Sunday** and **Daily** as your variables, and choose the appropriate statistics in the "Statistics to display" field.

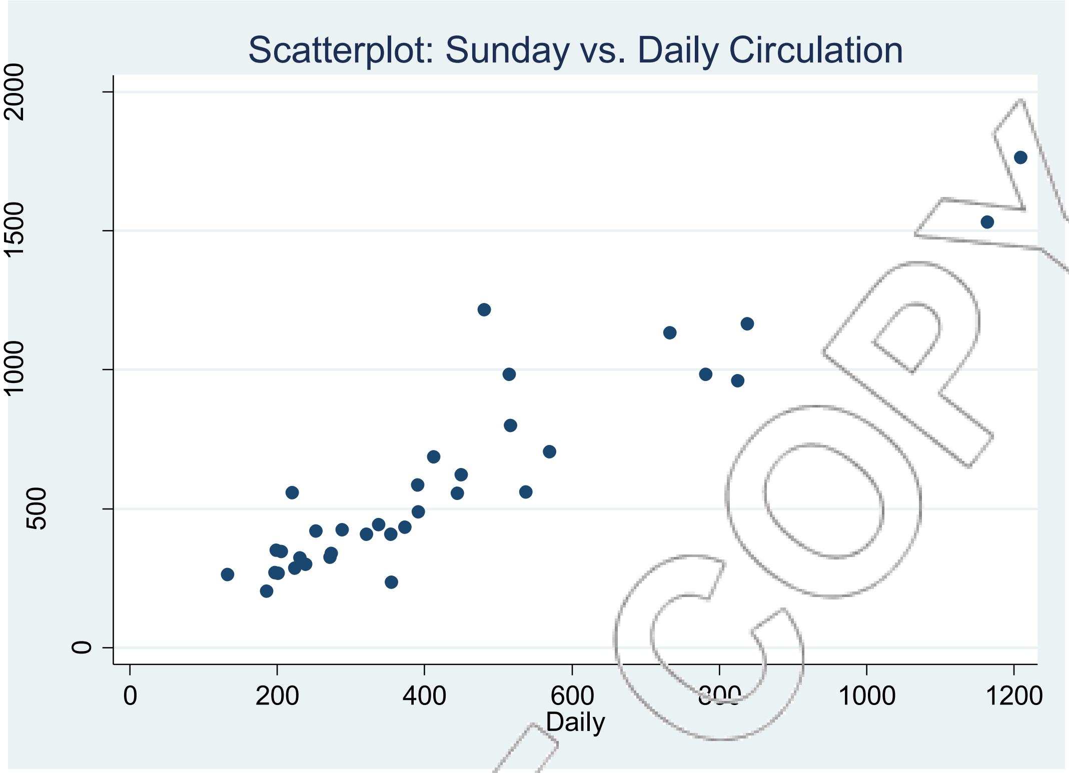
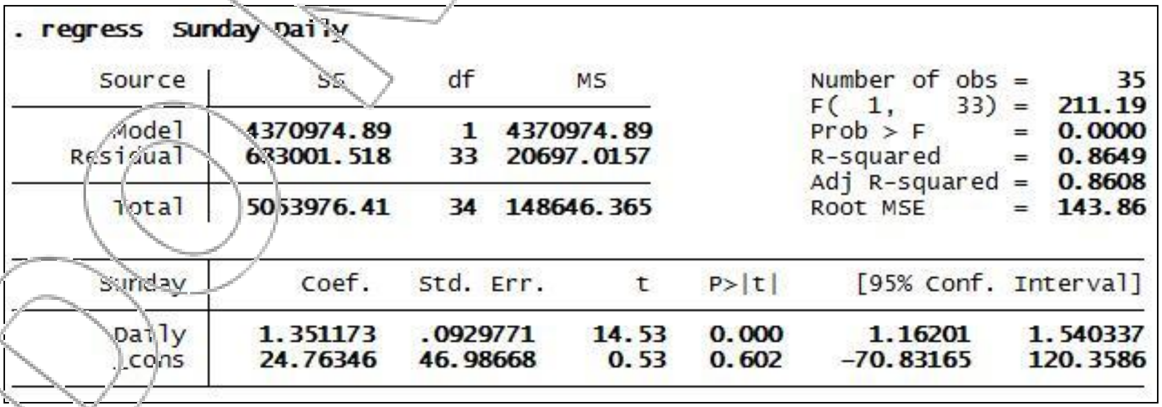


Figure 4.4: Scatterplot for Sunday vs. daily.

It looks as if the relationship is close to linear. Now, we will do a regression to see what the estimated relationship is and check to see if a strong relationship exists.



```
. regress  Sunday Daily

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  1,    33) =  211.19
       Model |  4370974.89     1  4370974.89           Prob > F      =  0.0000
    Residual |  683001.518    33  20697.0157           R-squared     =  0.8649
-------------+------------------------------           Adj R-squared =  0.8608
       Total |  5053976.41    34  148646.365           Root MSE      =  143.86

------------------------------------------------------------------------------
      Sunday |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Daily |   1.351173   .0929771    14.53   0.000      1.16201    1.540337
       _cons |   24.76346   46.98668     0.53   0.602    -70.83165    120.3586
------------------------------------------------------------------------------
```

Figure 4.5: Regression of Sunday vs. daily.

From Figure 4.5, our estimated regression equation is Sunday = 24.76+1.35 Daily. We may use the regression equation to produce an estimate of Sunday circulation for our newspaper. Substituting the daily circulation of 190 gives the following:

$$\text{Sunday} = 24.76+1.35(190) = 281.26$$

As the units are in thousands, this equation tells us the estimated Sunday circulation is 281,260. So, it looks as if we are saying that circulation will exceed our break-even figure of 260,000. But we have to be more careful than that. Regression is a statistical procedure: We are estimating the true relationship between Sunday and daily circulation, and the estimate is based on our sample, so it will contain some sampling error. In other words, there is a true line describing that relationship, which we do not know exactly but have estimated. Our best estimates of its intercept and slope are 24.76 and 1.35 respectively, but those are only estimates. We need to take this into account and quantify the sampling error in our estimate of Sunday circulation.

SAMPLING DISTRIBUTION OF THE FITTED VALUE $\hat{y}$

Earlier, we talked about the estimator b_1 and its sampling distribution. Most of what we said about b_1 also applies to the estimator b_0. Now, however, we are interested in a third estimator. The parameter we want to estimate is the average (or expected) Sunday circulation of a paper with a daily circulation of 190,000. If the regression model is right, then this is given (in thousands) by $\beta_0+\beta_1(190)$, and we estimate this average by using the following:

$$\hat{y}_{190} = b_0 + b_1(190)$$

We need to know the sampling distribution of this quantity, which is called the **fitted** or **predicted value** corresponding to *x* = 190. It makes sense to talk about the sampling distribution as this estimate is the outcome of an experiment. If we repeated the experiment by taking a different sample, we would get a different outcome, i.e., different estimates.

The sampling distribution of this estimator has the following properties:

The fitted value is normally distributed, with mean equal to the true value, i.e.,

$$\mathrm{E}(\hat{\mathrm{y}}_{190}) = \beta_0 + \beta_1(190)$$

so it is an **unbiased estimator**. Its standard deviation is written $\sigma_{\hat{y}_{190}}$ and can be estimated from the sample. This estimate, which we call the standard error of the estimated mean, is denoted by $s_{\hat{y}_{190}}$.

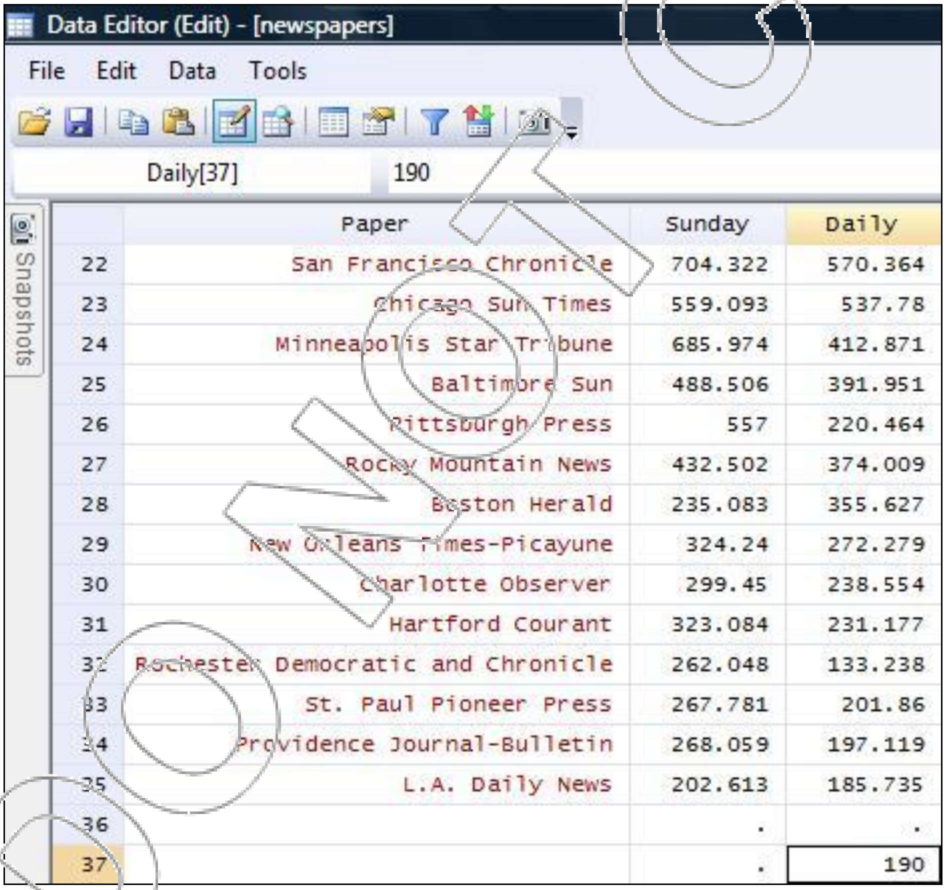
CONFIDENCE INTERVALS WITH THE FITTED VALUE

Since we know the distribution of our estimator, we can use it to produce confidence intervals as we have done with every other estimator. By now, you should know what the formula will be. To get a 100(1-α)% confidence interval for $\beta_0 + \beta_1(190)$, use the following:

$$\hat{\mathrm{y}}_{190} \pm t_{\alpha/2,n-2} s_{\hat{y}_{190}}$$

Here, $t_{\alpha/2,n-2}$ is the $\alpha/2$ t-value with n-2 degrees of freedom, and n is the sample size (i.e., $t_{\alpha/2,n-2}$ = invttail(n-2,α/2)). Next, we present an example showing how to use Stata to calculate $s_{\hat{y}_{190}}$ and this confidence interval.

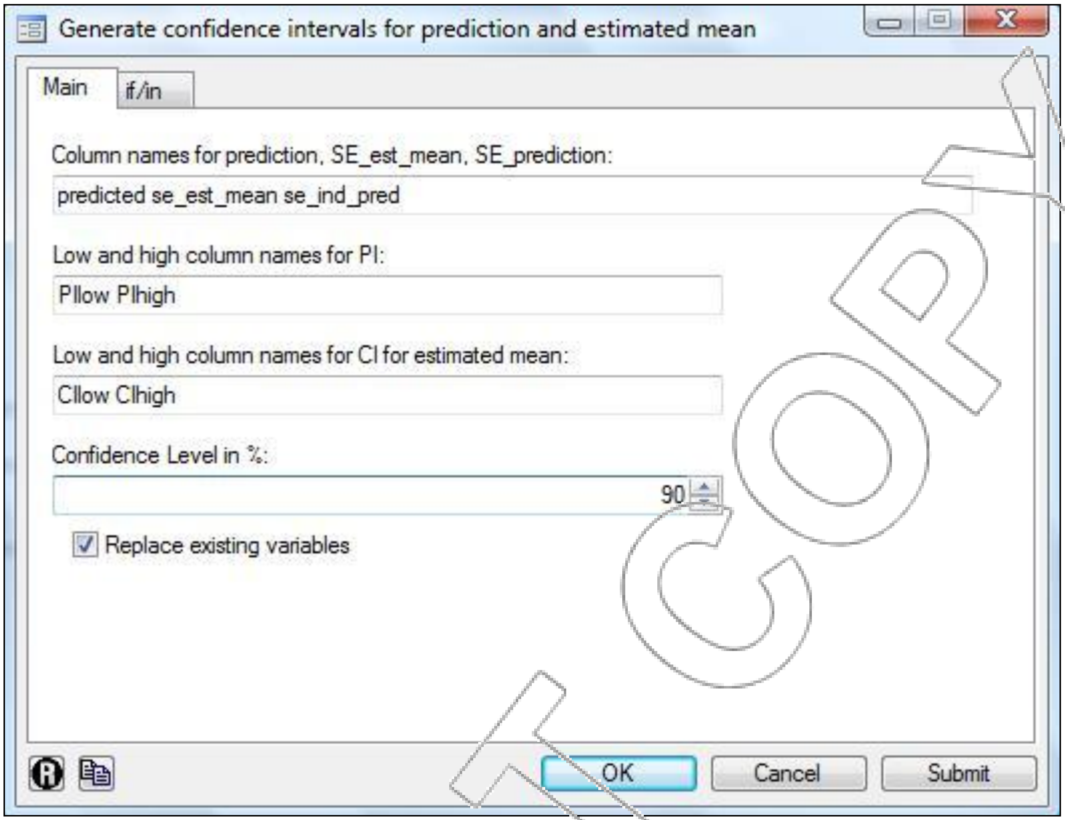
As an example, let's produce a 90% confidence interval for $\beta_0+\beta_1(190)$. First, run the regression of Sunday versus Daily in Stata. Then, open the Data Editor and enter **190** for the 37th observation under the **Daily** column (we leave a blank row to remind ourselves where the original data ends). Your data should look something like this:



Daily[37] 190

	Paper	Sunday	Daily
22	San Francisco Chronicle	704.322	570.364
23	Chicago Sun Times	559.093	537.78
24	Minneapolis Star Tribune	685.974	412.871
25	Baltimore Sun	488.506	391.951
26	Pittsburgh Press	557	220.464
27	Rocky Mountain News	432.502	374.009
28	Boston Herald	235.083	355.627
29	New Orleans Times-Picayune	324.24	272.279
30	Charlotte Observer	299.45	238.554
31	Hartford Courant	323.084	231.177
32	Rochester Democratic and Chronicle	262.048	133.238
33	St. Paul Pioneer Press	267.781	201.86
34	Providence Journal-Bulletin	268.059	197.119
35	L.A. Daily News	202.613	185.735
36		.	.
37		.	190

Next, click **User>Core Statistics>Prediction, using most recent regression (confint)** (or type **db confint**) and enter **90** in the "Confidence Level in %" field:


Generate confidence intervals for prediction and estimated mean
Main
if/in
OK
Cancel
Submit

Column names for prediction, SE_est_mean, SE_prediction: predicted se_est_mean se_ind_pred

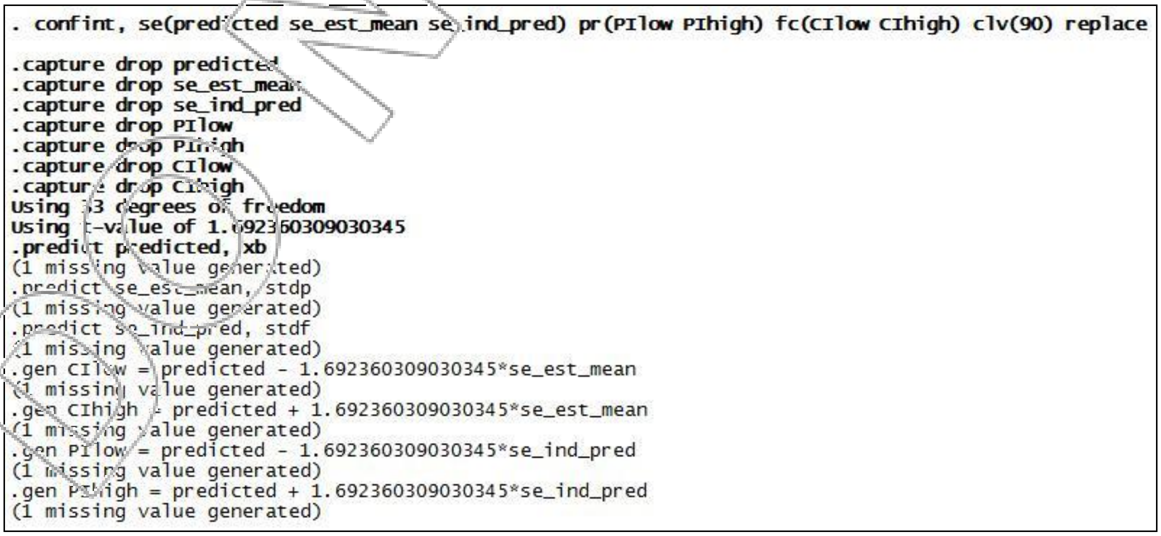
Low and high column names for PI: PIlow PIhigh

Low and high column names for CI for estimated mean: CIlow CIhigh

Confidence Level in %: 90

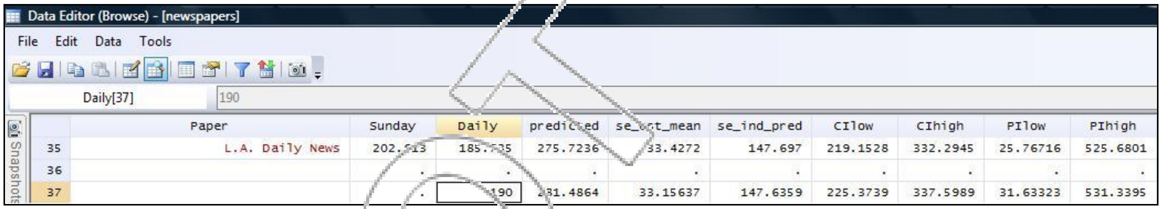
☑ Replace existing variables

Click **OK**, and Stata will generate the following:



```
. confint, se(predicted se_est_mean se_ind_pred) pr(PIlow PIhigh) fc(CIlow CIhigh) clv(90) replace

.capture drop predicted
.capture drop se_est_mean
.capture drop se_ind_pred
.capture drop PIlow
.capture drop PIhigh
.capture drop CIlow
.capture drop CIhigh
Using 33 degrees of freedom
Using t-value of 1.692360309030345
.predict predicted, xb
(1 missing value generated)
.predict se_est_mean, stdp
(1 missing value generated)
.predict se_ind_pred, stdf
(1 missing value generated)
.gen CIlow = predicted - 1.692360309030345*se_est_mean
(1 missing value generated)
.gen CIhigh = predicted + 1.692360309030345*se_est_mean
(1 missing value generated)
.gen PIlow = predicted - 1.692360309030345*se_ind_pred
(1 missing value generated)
.gen PIhigh = predicted + 1.692360309030345*se_ind_pred
(1 missing value generated)
```

Now, open the Data Browser and scroll down to the 37th observation (i.e. Daily=190). You will see that Stata has generated the information we need, as shown in Figure 4.6. The **predicted** column gives us the actual estimate (the fitted or predicted value) $\hat{y}_{190}$ = 281.4864, the **se_est_mean** column gives us the **standard error of the estimated mean**, $s_{\hat{y}_{190}}$ = 33.15637, which is the estimated standard deviation of $\hat{y}_{190}$ treated as an estimate of average circulation, and the **CIlow/CIhigh** columns give the requested 90% confidence interval. This data sheet also gives output relevant to prediction intervals, which are the subject of the next section. The **se_ind_pred** column gives us the **standard error of prediction**, the estimated standard deviation we use in calculating prediction intervals. The **PIlow/PIhigh** columns give the 90% prediction interval for the fitted value.



	Paper	Sunday	Daily	predicted	se_est_mean	se_ind_pred	CIlow	CIhigh	PIlow	PIhigh
35	L.A. Daily News	202.613	185.735	275.7236	33.4272	147.697	219.1528	332.2945	25.76716	525.6801
36		.	.	.	.	.	.	.	.	.
37		.	190	281.4864	33.15637	147.6359	225.3739	337.5989	31.63323	531.3395

Figure 4.6: Prediction of Sunday sales.

PREDICTION INTERVALS AND THE FITTED VALUE

Prediction intervals are particularly useful tools. A 90% confidence interval gives us a range of values in which we are 90% confident that the mean value falls, i.e., the mean Sunday circulation of all papers with daily circulation of 190,000. In contrast, a 90% prediction interval is a range of values which we are 90% confident would contain the circulation of a particular Sunday paper selected at random from all those with a daily circulation of 190,000.

Below is the formula for prediction intervals:

$$\boxed{\hat{y}_{190} \pm t_{\alpha/2,n-2}\sqrt{s^2 + s^2_{\hat{y}_{190}}}}$$

It is similar to the one for confidence intervals except that it uses a different (larger) standard deviation. This standard deviation is often referred to as the **standard error of prediction**. In the formula, *s* is the **standard error of regression**, $s_{\hat{y}_{190}}$ is the **standard error of the estimated mean**, and $\sqrt{s^2 + s^2_{\hat{y}_{190}}}$ is the **standard error of prediction**. The other symbols are familiar: $\hat{y}_{190}$ is the estimated value of the dependent variable, $t_{\alpha/2,n-2}$ is the α/2 t-statistic with n-2 degrees of freedom, and n is the sample size, i.e., $t_{\alpha/2,n-2}$ = invttail(n-2,α/2).

A word about terminology is called for at this point for power Stata users. Confusingly, Stata's built-in **predict** command – accessed via dialog box by typing **db predict** – refers to the standard error of prediction as the "standard error of the forecast", and uses the term "standard error of the prediction" to refer to what we call the standard error of the estimated mean.

HYPOTHESIS TESTS WITH THE FITTED VALUE

The fitted value, $\hat{y}_{190}$, is our estimator for the population average *y* when x = 190 as well as for an individual value, y_i, when *x* = 190. As we have seen in the previous two sections, we can determine the range around $\hat{y}_{190}$ where the population average should fall (when *x* = 190 with a given confidence) using the standard error of the estimated mean, and a similar range where an individual value, y_i, should be using the standard error of prediction. We can use these standard errors to develop hypothesis tests regarding the population average *y* at *x* = 190 and confidence statements about an individual y_i at *x* = 190.

First, we try to prove the average of the Sunday circulations of all newspapers with a daily circulation of 190,000 is greater than 260,000. The basic steps are the same as in all previous hypothesis tests. The hypotheses are the following:

H_0: at x = 190, population average y ≤ 260

H_a: at x = 190, population average y > 260.

We calculate the test statistic that shows by how many standard errors the estimator is greater than 260. The estimator is $\hat{y}_{190}$, and since the hypothesis is about the population average, the correct standard error to use is the standard error of the estimated mean (at x = 190), which we can find from Figure 4.6.

t = ($\hat{y}_{190}$ -260) / (se_est_mean at x = 190) = (281.49-260)/33.16 = 0.648.

Now we can proceed to calculate the p-value of the test with the following:

p-value = ttail(#degrees of freedom, t-value) = ttail(33, 0.648) = 0.26.

Since p = 26%, we cannot reject the null at a 5% significance level. In other words, we cannot prove at a 5% significance level that the average of the Sunday circulations of newspapers with a daily edition of 190,000 copies is greater than 260,000.

As a manager at your newspaper, you are not necessarily interested in the above result. You are more interested in determining whether your own Sunday edition will exceed 260,000 copies per day (you do not directly care about the industry average for papers with your daily circulation); that is, you might want to test the following:

$$H_0\text{: at } x = 190\text{, individual } y_i \leq 260$$

$$H_a\text{: at } x = 190\text{, individual } y_i > 260.$$

This is not a statement about population parameters, and is therefore not a valid hypothesis test. However, you can still use a similar procedure to better understand your newspaper's potential Sunday circulation. The estimator for your Sunday circulation is the same as before, $\hat{y}_{190}$, but the correct standard error to use in the calculation is now the standard error of prediction at x=190. Using the information from Figure 4.6, we calculate that the break-even level of 260 is 0.146 standard errors of prediction below the estimator:

$$\text{t-value} = (\hat{y}_{190}\text{ -260})/(\text{se\_ind\_pred at } x = 190) = (281.49\text{-}260)/147.64 = 0.146.$$

If this were indeed a hypothesis test, its p-value would be given by the following:

$$\text{p-value} = \text{ttail}(\#\text{degrees of freedom, t-value}) = \text{ttail}(33, 0.146) = 0.442,$$

or 44.2%. This is the area that lies below a prediction interval with lower endpoint 260. In this sense, we are only 1-0.442 = 0.558 or 55.8% confident that Sunday circulation exceeds break-even. In other words, we expect the Sunday circulation of an individual newspaper with a daily circulation of 190,000 to exceed 260,000 but there is still a reasonable chance it does not.

THE DECISION

Remember that the break-even point for this project was a circulation of 260,000. The regression gives a point prediction of 281,486, but if we look at the 90% prediction interval (31,633 to

531,340), we see this point prediction is of little use because the margin of error is enormous. The same conclusion arises from our latter hypothesis test, as we cannot prove at any reasonable significance level that our newspaper will have a Sunday circulation in excess of 260,000. In other words, knowing the daily circulation is not informative enough about Sunday circulation.

Was this regression useless? No; however, it suggests we need to collect more information to obtain a prediction accurate enough to make the decision. Newspaper circulation can be predicted much more accurately if we add in some extra variables (various demographics) and perform a multiple regression. We will examine multiple regression techniques in coming chapters.

Though daily circulation on its own is not informative enough to make the kind of prediction we need, it did explain a large fraction of the variation in Sunday circulation. We know this because of something called the R-squared statistic, which you can see in the regression output (R-squared = 0.8649).

THE R-SQUARED STATISTIC

If you have ever studied regression before, you will likely recognize the R-squared. It is a number that tells you how much variation in the *y* or dependent variable is explained by the regression equation. In this example, the dependent variable is Sunday circulation, and its total variation is defined this way:

$$\boxed{\sum\left(y_i - \bar{y}\right)^2}$$

$\bar{y}$ is the mean Sunday circulation of all the papers in the sample. This quantity is usually known as the total sum of squares (SST). You can find it by running a regression using Stata, and looking at the value in the **Total** row and the **SS** column. Next, we take the estimated regression equation $\hat{y}$ = b_0+b_1x and ask how much variation it predicts. Taking each paper in the sample in turn, look at its daily circulation x_i, and calculate the Sunday circulation that the regression equation predicts for that x_i. This number is called the ith fitted value $\hat{y}_i$. This value $\hat{y}_i$ is not equal to the true Sunday circulation of the ith paper because the regression is not totally accurate. But we can ask how much variance there would be if this regression were totally accurate, so that each $\hat{y}_i$ was the true value for its paper. This is given by applying the variation formula to the $\hat{y}_i$'s instead of to the y_i's:

$$\sum(\hat{y}_i - \bar{y})^2$$

This quantity is known as the sum of squares due to regression (SSR). You can find it on the regression output table in the **Model** row and the **SS** column. The SSR tells us how much variance there would be in our sample if Sunday circulation were exactly related to daily circulation by the estimated regression equation, i.e., if our best-fit line were a perfect fit. If the best-fit line is close to the data points, the SSR will be close to the SST since in that case the best-fit line is predicting accurately; if the best-fit line is a poor fit, the SSR will be much smaller than the SST.

This intuition leads to the mathematical definition of the R-squared:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

In this case, the high R-squared (0.8649 or 86.49%) tells us that daily circulation does an impressive job of explaining/predicting the variation in Sunday circulation; however, there is an enormous amount of variation overall. The remaining 13.51% variation that is unexplained represents a thin slice of a very large pie. In this example, the unexplained fraction is too much to make the prediction useful. One moral of this example is you should not overvalue the R-squared. One of the most common mistakes in using regression is thinking that a high R-squared means the regression is automatically useful for prediction. As we have seen, that is not the case. Similarly, a low R-squared does not mean that a regression is useless.

R-SQUARED AND ASSET BETAS

If you look back to our regression of McDonald's excess returns against the market, you will see that the R-squared in that regression is only about 53%. Should we have worried about this? The answer is no. All we were interested in was the accuracy of our beta estimate, and the R-squared is irrelevant to this. What it does tell us is how much of the variance in McDonald's share price is explained by the market's movements as a whole. This has a nice interpretation. Recall that a basic idea behind the beta and the CAPM model is that some risk is specific to each firm, and therefore diversifiable; the rest is due to the movements of the whole market and cannot be avoided by diversification. The R-squared is the ratio of variance (i.e., risk) due to the market and the total variance in McDonald's stock. In other words, it tells you what proportion of the risk in holding McDonald's shares cannot be diversified away. So, in this case, about 47% of McDonald's risk is firm-specific, related to things such as the success or failure of its new ad campaign or new food ideas; the other 53% is systematic risk, related to factors like people spending less at McDonald's in hard times.

SUMMARY

In this chapter, we learned two of the important things that regression can be used for: studying how changes in an independent variable relate to changes in the dependent variable through the coefficient and using a particular value of the independent variable to make predictions of the dependent variable. In both cases, we observed point estimates and interval estimates. In the finance case, we estimated a beta and gave confidence intervals reflecting our uncertainty about the estimate. With our newspaper circulation case, we estimated Sunday sales for a paper with a certain daily sales level and gave a prediction interval to demonstrate the limitations of our estimate.

Between the two cases, we used four different standard errors computed by Stata. Though each of these represents the same basic idea, a measure of the uncertainty of some estimate, you must keep track of which estimates are being assessed by which standard errors.

NEW TERMS

Fitted value The value of the dependent variable (y) predicted by the regression equation for a given value of the independent variable (x). It is a prediction for the average value of y given x and for an individual realization of y given x

Standard error of the coefficient An estimate of the standard deviation of a regression coefficient

Standard error of the estimated mean An estimate of the standard deviation of the fitted value. It is used in constructing confidence intervals for the average value of y given x and in conducting hypothesis tests concerning the average value of y given x

Standard error of prediction An estimate of the standard deviation of our estimate for an individual value of y given x. Calculated by combining the standard error of regression with the standard error of the estimated mean. Used in constructing prediction intervals for an individual value of y given x and in conducting hypothesis tests concerning an individual value of y given x

NEW FORMULAS

CAPM formula $r - r_f = \beta(r_m - r_f)$

Confidence Interval for the average value of y given x = p $\hat{y}_p \pm t_{\alpha/2,n-2}\, s_{\hat{y}_p}$

Prediction Interval for y given x = p $\hat{y}_p \pm t_{\alpha/2,n-2}\sqrt{s^2 + s_{\hat{y}_p}^2}$

Total Sum of Squares (SST) $\sum\left(y_i - \bar{y}\right)^2$

Sum of Squares due to Regression (SSR) $\sum\left(\hat{y}_i - \bar{y}\right)^2$

R-squared $R^2 = \frac{\text{SSR}}{\text{SST}}$

NEW STATA FUNCTIONS

User>Core Statistics>Prediction, using most recent regression (confint)

Equivalently, you may type **db confint**. This command generates fitted or predicted values, the standard error of the estimated mean, the standard error of prediction as well as prediction and confidence intervals. Because it uses the output of the most recent regression, you must run a regression before using this command. After running the regression, open the Data Editor and in some blank row(s) enter values in the respective column(s) of the independent variable(s) from which you want to generate the fitted value. Then, follow the menu path or type **db confint** to open the dialog box and set the desired level of confidence. The default is 95% confidence. When you click **OK**, results will be calculated for each set of values you have entered (as well as for all of the original observations on your datasheet).

If you want to generate only predicted values, only the standard error of the estimated mean, or only the standard error of prediction after running a regression, you can click **Statistics>Postestimation>Prediction, residuals, etc.** or type **db predict**. In the “New variable name” field, type in the name for which you want your predicted values or standard errors to be displayed as, and choose the appropriate variable from the “Produce” list:

a. To generate predicted values, choose “Linear prediction (xb).”

b. To generate the standard error of the estimated mean, choose “Standard error of the prediction.”

c. To generate the standard error of prediction, choose “Standard error of the forecast.”

The corresponding direct commands are:

a. **predict *newvar*, xb**

b. **predict *newvar*, stdp**

c. **predict *newvar*, stdf**

where ***newvar*** is the name of the newly generated variable.

CASE EXERCISES

1. Estimating betas

Access the **stocks** dataset and use it to estimate betas for the following stocks: Apple, IBM, and HP. Suppose the excess returns on the stock market (as measured by the S&P500, stored under **ESP** in the dataset) were to be negative 20% next month.

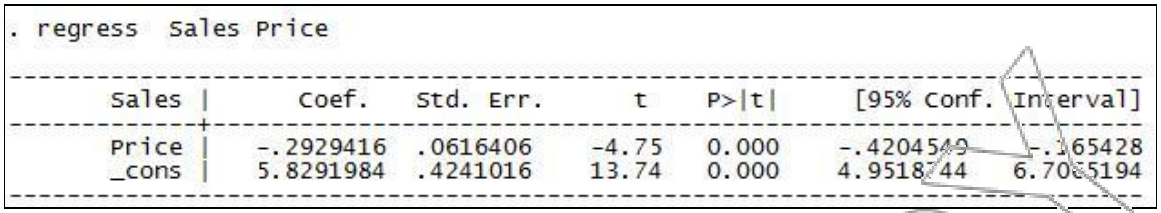
a. What would you expect to be the excess return on Apple shares next month? How about IBM and HP shares? Base your estimate on the estimated beta and the theoretical CAPM equation; that is, discard the estimated constant (alpha), as we did in the chapter.

b. How much money would you expect to lose next month if you had $10,000 invested in Apple shares at the beginning of the month? For the purposes of answering this part of the question only, assume that the risk-free rate next month is 0.25%.

In the example from the chapter, we used the variable **MARKET** to measure the market excess return. In this problem, we ask you to use an alternative method of measuring the market excess return using the variable ESP. So, for this exercise, use ESP. One problem with the CAPM is that it is not obvious how to measure the market return. Market is a combination of bonds and the S&P 500, and ESP includes only the S&P 500. Finally, all the variables in the dataset are **excess market returns** (i.e., market returns minus the risk-free interest rate).

2. Slippery soap sales

Greenfield, Inc., a manufacturer of a popular bathing soap, tried to find the relation between its product's price and its sales. It supplies over 2,000 retail outlets in the United States. It collected data from 25 of these stores during one week and ran a regression using these data. For each store in the sample, it observed the independent variable **Price** (measured in dollars), and the dependent variable **Sales** (measured in thousands of dollars). The results were as follows:



```
. regress  Sales Price

------------------------------------------------------------------------------
       Sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Price |  -.2929416   .0616406    -4.75   0.000    -.4204549    -.165428
       _cons |  5.8291984   .4241016    13.74   0.000    4.9518744   6.7065194
------------------------------------------------------------------------------
```

a. If the price of the bathing soap is reduced by $0.50, what is the expected increase in sales per store? Additionally, provide a 95% confidence interval for the expected increase.

b. The product manager claims that if the price is reduced by $0.50, average sales will increase by at least $160 per store. Do the data allow you to reject this claim at a level of significance of 5 percent?

c. The price in all stores next week is going to be $9.99. Predict the total expected sales including all of the 2,000 stores during next week.

3. Shore Realty revisited

Retrieve the **shore** dataset, which we used in Case Exercise 3 in Chapter 3, and run the regression again. Provide a 90% confidence interval for the coefficient on the sqfoot variable, and explain clearly and concisely what this interval means. Predict the selling price for a home with 2,600 square feet, provide the associated 95% confidence and prediction intervals and explain clearly and concisely what each means. Suppose that Shore Realty sells a large number of houses of this size: what proportion of them would you expect to sell for over $383,000?

PROBLEMS

Access the **Retailsales** data file to answer problems 1–3. This data file reports the percentage change in total domestic retail sales and the percentage change in the U.S. GDP over a recent ten-year period. (from A.C. Nielsen’s *Facts, Figures and the Future*. Feb. 2003).

1. Perform a regression of **percent_chginRetailSales** using **percent_chginGDP** as the independent variable.

   a. Write the estimated regression equation.

   b. Use the regression to estimate how much a one percentage point increase in GDP will affect retail sales.

   c. Provide a 95% confidence interval for your estimate in part b.

   d. Provide a 90% confidence interval for your estimate in part b.

   e. Using α = 0.05, can you reject the null hypothesis that the true coefficient multiplying **percent_chginGDP** is zero?

2. Use the regression from problem 1.

   a. Predict the **percent_chginRetailSales** in a year where the GDP increases by 3.0%.

   b. Provide a 95% prediction interval for your estimate.

   c. Provide a 98% prediction interval for your estimate.

   d. Using the same prediction, estimate the probability that the **percent_chginRetailSales** will be greater than 8.5.

3. Overall how would you rate the quality of this regression? Justify your answer.

Access the **Salaries** file to answer questions 4–6. These data represent the salaries of 41 workers at a major corporation based on the number of years employed with the company.

4. Perform a regression of **Salary** vs. **Years Experience**.

   a. Write out the estimated regression equation.

b. Use the regression to estimate the effect of one additional year of work experience at the company on a worker’s salary.

c. Provide a 95% confidence interval for your estimate in part b.

d. Provide a 99% confidence interval for your estimate in part b.

e. Using $\alpha = 0.05$, can you reject the null hypothesis that the true coefficient is zero?

5. Use the regression from problem 4.

a. Predict the salary of a worker with nine years of experience at the company.

b. Provide a 95% prediction interval for your estimate.

c. Provide a 75% prediction interval for your estimate.

d. Provide an interval that you are 90% confident contains the true mean salary of workers with nine years of experience.

e. How confident can we be that work experience is significantly related to salary?

6. What percentage of salary can be explained using an employee’s work experience with the company? Does this number sound reasonable to you?

For problems 7–9, you will need to access the **eurodata** file, which contains information from the *Statistical Annex of the European Economy, 2003.* The dataset consists of 42 years worth of wage rate growth and unemployment rates for 10 countries in Europe. Multinational corporations might be interested in studying how unemployment impacts the growth in wages for some or all of these 10 countries.

7. Perform a regression of wage growth vs. unemployment in Belgium (BE). Do the same for Denmark (DK).

a. Write both estimated regression equations.

b. How does a one percentage point increase in unemployment relate to the growth rate of wages in each country?

c. Provide a 95% confidence interval for the coefficient multiplying unemployment for each country.

d. Predict the growth rate in wages for each country in a year that has 3% unemployment

e. Provide a 90% confidence interval for each prediction from part d.

8. Perform a regression of wage growth vs. unemployment in Germany (DE). Do the same for Greece (EL).

a. Write both estimated regression equations.

b. How does a one percentage point increase in unemployment relate to the growth rate of wages in each country?

c. Provide a 95% confidence interval for the coefficient multiplying unemployment for each country.

d. Predict the growth rate in wages for each country in a year that has 3% unemployment.

e. Provide a 90% confidence interval for each prediction from part d.

9. Perform a regression of wage growth vs. unemployment in Spain (ES). Do the same for France (FR).

a. Write both estimated regression equations.

b. How does a one percentage point increase in unemployment relate to the growth rate of wages in each country?

c. Provide a 95% confidence interval for the coefficient multiplying unemployment for each country.

d. Predict the growth rate in wages for each country in a year that has 3% unemployment.

e. Provide a 90% confidence interval for each prediction from part d.

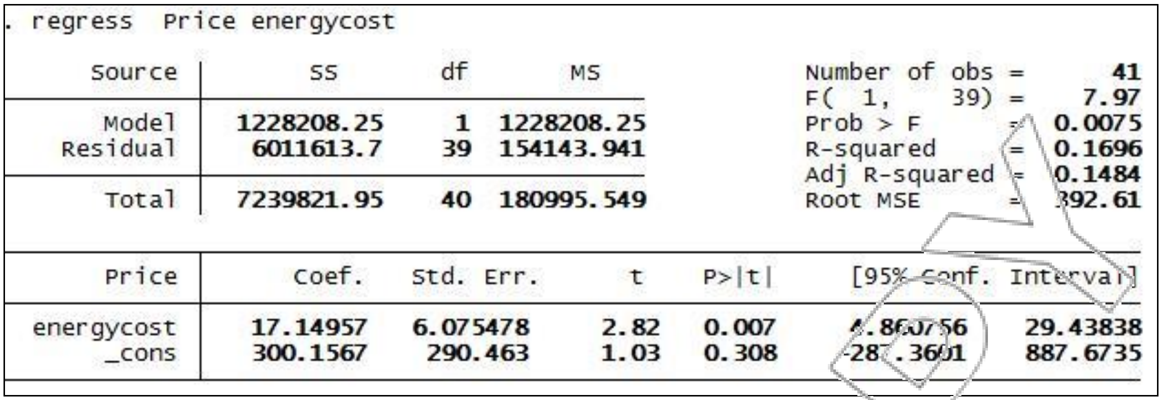
CASE INSERT 1

ENERGY COSTS AND REFRIGERATOR PRICING

As a manager in charge of a brand of refrigerators, you are confronted with the following scenario: A representative from your company’s research and development team sends you a report announcing a breakthrough in energy-efficient refrigeration technology. Specifically, the team believes that for an additional production cost of $80 per refrigerator, the consumer’s annual energy costs to run the refrigerator will drop by $20. Should you incorporate this new technology into your next refrigerator model?

One key issue is how much extra you could charge for a more energy-efficient fridge. To get an estimate of this, you order a study of the relationship between the annual energy costs and price of a refrigerator. The data gathered for this study provide information on 41 popular models of refrigerators.[1] Using these data, a regression of price on annual energy costs is performed. The variables are “Price,” which gives the refrigerator price (in $), and “Energy cost,” which gives the annual energy cost of running the refrigerator (in $/year).

[1] You can access this data in the **newfridge** file. Source: *Consumer Reports*, July 2003, Vol. 68, No. 7.



```
. regress  Price energycost

      Source |       SS       df       MS              Number of obs =      41
-------------+------------------------------           F(  1,    39) =    7.97
       Model |  1228208.25     1  1228208.25           Prob > F      =  0.0075
    Residual |   6011613.7    39  154143.941           R-squared     =  0.1696
-------------+------------------------------           Adj R-squared =  0.1484
       Total |  7239821.95    40  180995.549           Root MSE      =  392.61

------------------------------------------------------------------------------
       Price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  energycost |   17.14957   6.075478     2.82   0.007     4.860756    29.43838
       _cons |   300.1567    290.463     1.03   0.308    -287.3601    887.6735
------------------------------------------------------------------------------
```

Case Questions

1. Given this output, what is an estimate for the change in price of a refrigerator model when its annual energy costs decrease by $20?
2. Given this estimate, would you go ahead with the new technology?
3. Does this estimate make sense? Explain

CHAPTER 5

CALIFORNIA STRAWBERRIES:

DUMMY AND SLOPE DUMMY VARIABLES

In this chapter, we will learn about using two kinds of dummy variables to capture qualitative features in regression in the California Strawberries and the CEO Seek Cases.

5.1 Dummy Variables

DUMMY VARIABLES: REVISITING THE PACKAGING CASE

A “dummy” or “qualitative” variable is one that only takes on the values 0 and 1. The idea of a dummy variable is it measures not a quantity but a quality. For an example, go back to the consumer packaging example from Chapter 2. The dataset consisted of 72 sales figures, 36 from locations using packaging one and 36 from locations using packaging two. If we number these locations 1 through 72, we can define y_i to be sales at location i (so y_i is a regular, quantitative variable) and x_i to be a dummy variable defined by the following:

$$x_i = \begin{cases} 0 \text{ if location i uses packaging one} \\ 1 \text{ if location i uses packaging two} \end{cases}$$

You will see that dummy variables are one of the most useful techniques available in regression because they enable us to measure the effect of qualitative differences. This section introduces you to dummy variables and how to use them in regression by reproducing the two-sample results we obtained in Chapter 2.

INTERPRETING DUMMY VARIABLES IN THE REGRESSION MODEL

Suppose we regress sales on our packaging dummy. What is the meaning of this regression? Remember the regression model: The assumption is that we may write the following:

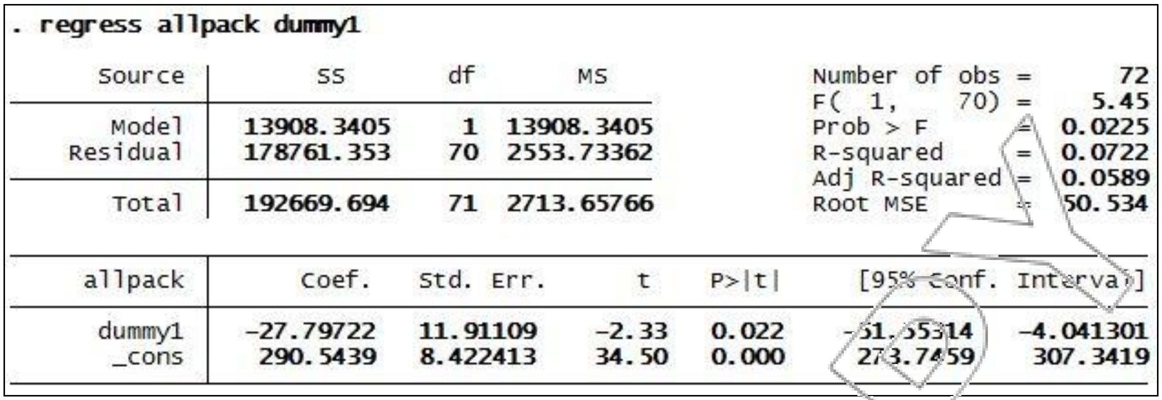
$$\mathrm{E}(y) = \beta_0 + \beta_1 x$$

That is, the average value of *y* for a given *x* is a linear function of *x*. That seemed to make sense when *x* was measuring income and *y* auto price. What does it mean when *x* is a dummy variable? Suppose *x* = 0; then, the equation says the expected value of *y* is β_0. So, β_0 is the expected value of y when *x* = 0, i.e., expected sales in districts using packaging one. For *x* = 1, the equation says the expected value of *y* is β_0+$\beta_1$1; so, the expected sales in districts using packaging two equals β_0+β_1. What is the difference in expected sales between districts using packaging two and districts using packaging one? It is β_0+β_1-β_0 = β_1. When we run the regression and estimate β_1, what we are estimating is the difference in expected sales between the two types of packaging, which is what we wanted to estimate in the first place in Chapter 2 because it tells us which packaging we should choose.

THE REGRESSION

Go ahead and run this regression using the **allpack** file. Our data should consist of two columns. The first (called allpack) is a list of sales figures, one for each district, and the second (called dummy1) is 0 for the first 36 entries since the first 36 sales figures come from districts that used packaging one (P1), and 1 for the next 36 since the next 36 sales figures come from districts that used packaging two (P2).[1]

[1] This dataset was generated from the original **package** file from Chapter 2. To create the **allpack** variable, we opened a blank datasheet in Stata and pasted the sales figures for Pack1 and Pack2 into one column (i.e., the first 36 entries were from Pack1, and the next 36 were from Pack2). To create the **dummy1** variable, we typed the following commands: 1) **generate dummy1=0 in 1/36**, and 2) **replace dummy1=1 in 37/72**.



```
. regress allpack dummy1

      Source |       SS       df       MS              Number of obs =      72
-------------+------------------------------           F(  1,    70) =    5.45
       Model |  13908.3405     1  13908.3405           Prob > F      =  0.0225
    Residual |  178761.353    70  2553.73362           R-squared     =  0.0722
-------------+------------------------------           Adj R-squared =  0.0589
       Total |  192669.694    71  2713.65766           Root MSE      =  50.534

------------------------------------------------------------------------------
     allpack |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      dummy1 |  -27.79722   11.91109    -2.33   0.022    -51.55314   -4.041301
       _cons |   290.5439   8.422413    34.50   0.000     273.7459    307.3419
------------------------------------------------------------------------------
```

Figure 5.1: Allpack regression.

If you look back at the consumer packaging section, you will see that we estimated the difference in average sales with P1 versus P2 to be 27.79 in favor of P1. Here in the regression output we have b_1 = -27.80 which says that we estimate that when x goes from 0 to 1, i.e., when we change from P1 to P2, sales go down on average by 27.80. So, the regression has given us the same estimate we had before (the 0.01 difference is due to rounding when we estimated the difference in average sales).

One convenient thing about using the regression is Stata has automatically tested this coefficient for significance: The t-statistic is -2.33, giving a p-value of .022. Recall that this is the p-value for the following hypothesis test:

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

So, the p-value of .022 tells us we are quite confident (over 97% confident) that $\beta_1 \neq 0$. What does this mean in the context of our example? Since we worked out that $\beta_1 = \mu_2 - \mu_1$, it means that we are quite confident that there is a difference in true average sales between the two types of packaging. Is this what we wanted to know? Not exactly. We wanted to see if the data provided strong evidence that average sales with P1 were above those with P2 using the hypothesis test:

$$H_0: \mu_1 - \mu_2 \leq 0$$

$$H_a: \mu_1 - \mu_2 > 0$$

Since $\beta_1 = \mu_2 - \mu_1$, this may be rewritten as

$$H_0: \beta_1 \geq 0$$

$$H_a: \beta_1 < 0$$

Using the regression output, we calculate the p-value for this test to be p = 1-ttail(70,-2.33) = .01135. Thus this data provides very strong evidence that average sales with P1 are higher, supporting a decision to go with packaging 1 rather than continue the marketing experiment.

When we used Stata to conduct the same hypothesis test using the two-sample t-test in Chapter 2.4, it reported a p-value of .0113 and we reached the same conclusion.

A NOTE ON OUR ASSUMPTIONS

Even though the p-values in the example are similar for the test based on the regression as for the two-sample t-test in Chapter 2.4, the two methods of comparing two population means rely on different assumptions. As you know, regression assumes the *y* values have the same variances for different *x* values, which, in this example, is equivalent to assuming the y values have the same variance for each of the two populations. The two sample t-test used in Chapter 2.4 did not use this assumption. Formally, using regression with a single dummy variable yields the same results as using a two-sample t-test assuming equal variances, and these results may differ from those obtained by using a two-sample t-test without assuming equal variances.

SUMMARY

Dummy variables capture qualitative differences rather than quantitative ones. When we have data from two populations, we can define a dummy variable to represent which population each data point comes from, run a regression to estimate differences in the two population means, and test the difference for statistical significance, etc. This is an alternative technique to the two-sample methods we learned earlier and provides a first application of dummy variables.

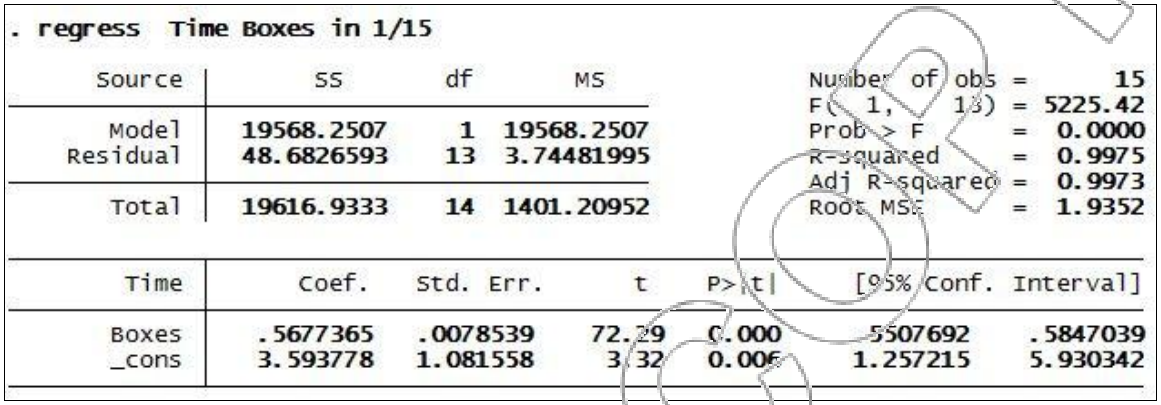
5.2 California Strawberries

Susan Lee is the chief manager of California Strawberries, Inc. Her firm transports strawberries from local farmers to a chain of grocery stores. The strawberries are packed into the retail boxes in two locations, using two different packaging systems. One is used at the plant in Bakersfield and the other in Monterey. Susan wants to compare the efficiency of the two systems and decide if one of the systems should be abandoned. The personnel and equipment needed for the two systems are basically identical. However, the time taken to pack a box of strawberries in Bakersfield and Monterey differs. Susan wants to adopt the quickest system. She asked her assistant to observe the time (measured in minutes) taken to pack different amounts of strawberries (measured in number of boxes) at Bakersfield and Monterey. The data he obtained is in the **california** file and is shown in Figure 5.2:

	Monterey			Bakersfield	
Row	Time	Boxes		Time	Boxes
1	102	175		95	140
2	69	110		104	153
3	133	225		48	70
4	37	57		108	161
5	28	47		89	128
6	124	217		85	125
7	71	120		90	133
8	36	60		81	122
9	41	65		68	95
10	104	180		98	143
11	126	210		109	161
12	63	106		54	80
13	34	50		85	128
14	38	60		137	205
15	88	150		85	125

Figure 5.2: California Strawberries, Inc. data.

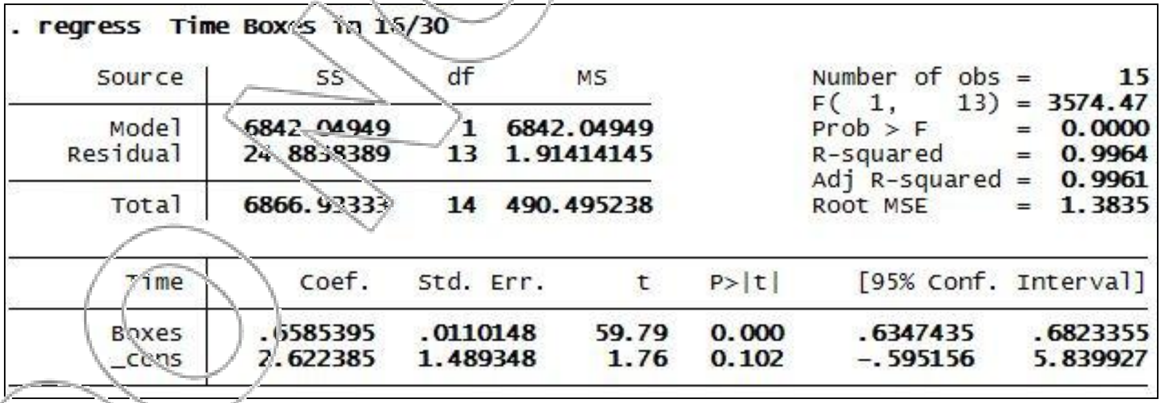
We can use a regression analysis to study the relationship between the two variables. Time is the dependent variable and Boxes is the independent variable. In this first regression (see Figure 5.3), we use only the data obtained at the Monterey plant.[2]



```
. regress  Time Boxes in 1/15

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) = 5225.42
       Model |  19568.2507     1  19568.2507           Prob > F      =  0.0000
    Residual |  48.6826593    13  3.74481995           R-squared     =  0.9975
-------------+------------------------------           Adj R-squared =  0.9973
       Total |  19616.9333    14  1401.20952           Root MSE      =  1.9352

------------------------------------------------------------------------------
        Time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Boxes |   .5677365   .0078539    72.29   0.000     .5507692    .5847039
       _cons |   3.593778   1.081558     3.32   0.006     1.257215    5.930342
------------------------------------------------------------------------------
```

Figure 5.3: Simple regression for the Monterey system.

Now, consider a similar regression for the Bakersfield system (see Figure 5.4). In this regression, we use only the data obtained at the Bakersfield plant.



```
. regress  Time Boxes in 16/30

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) = 3574.47
       Model |  6842.04949     1  6842.04949           Prob > F      =  0.0000
    Residual |  24.8838389    13  1.91414145           R-squared     =  0.9964
-------------+------------------------------           Adj R-squared =  0.9961
       Total |  6866.93333    14  490.495238           Root MSE      =  1.3835

------------------------------------------------------------------------------
        Time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Boxes |   .6585395   .0110148    59.79   0.000     .6347435    .6823355
       _cons |   2.622385   1.489348     1.76   0.102    -.595156     5.839927
------------------------------------------------------------------------------
```

Figure 5.4: Simple regression for the Bakersfield system.

[2] As shown in Figure 5.3, we need to add the command **in 1/15** to specify that we want to run the regression of Time on Boxes using only observations from the Monterey plant (observations 1 to 15). Similarly, we need to add the command **in 16/30** when running a similar regression for the Bakersfield plant (observations 16 to 30) as shown in Figure 5.4. If using the **regress** menu option or dialog box, these restrictions on the observations to use can be entered by selecting the **by/if/in** tab in the dialog box, checking the box next to "Obs. in range," and specifying the appropriate range.

What is the interpretation of these two regressions? The constant term indicates the time needed to start the system (literally, the time to pack 0 boxes). The coefficient on Boxes indicates the time it takes to pack each additional box. The regression analysis suggests it takes a longer time to set up the Monterey system (3.59 min) than the Bakersfield system (2.62 min). However, once the system is ready, the Monterey system (0.57 min per box) is faster than the Bakersfield system (0.66 min per box).

Susan believes the time to set up both systems should be similar, and she decides to maintain this hypothesis unless she discovers strong evidence against it.

Before she examines the regressions, Susan does not have any reason to believe that the time to pack each additional box in Monterey is smaller than in Bakersfield, nor does she have any reason to believe that the time per additional box in Bakersfield is smaller than in Monterey. By looking at the regressions, she feels tempted to abandon the Bakersfield system. However, she decides not to do so unless significant statistical evidence shows the Bakersfield system is slower.

Susan has good reasons to be cautious. Suppose the Bakersfield system is actually faster than the system in Monterey. In this case, if Susan switches to the Monterey system on the basis of the sample data, California Strawberries, Inc. will incur the costs of forcing the workers to adapt themselves to a new (and slower) system. Moreover, she will not be led to correct her mistake in the future because, once the Bakersfield system is abandoned, no more data will be available from it.

If the current sample evidence is not strong enough to prove that one system is faster than the other, it may be wise to obtain more data before making a decision. On the other hand, if the statistical evidence strongly convinces her it takes less time to pack an additional box in the Monterey system than to pack it in the Bakersfield system but there is no strong statistical evidence that shows the time to set up the Bakersfield system is shorter than the time to set up the Monterey system, then Susan can safely decide to abandon the Bakersfield system. How can Susan perform these statistical tests?

A simple and effective solution to this problem is to use dummy and slope dummy variables. A **slope dummy variable** is a variable that takes the value zero in some rows and the value of another independent (i.e., x) variable elsewhere.

Such a slope dummy variable may be constructed by multiplying a dummy variable times another x variable.

In simple regressions, we fit the data to a single straight line. However, in this case, the data come from two different sources and may not be well modeled by a single straight line, but may fit two different straight lines. A simple illustration of this possibility is given in Figure 5.5.

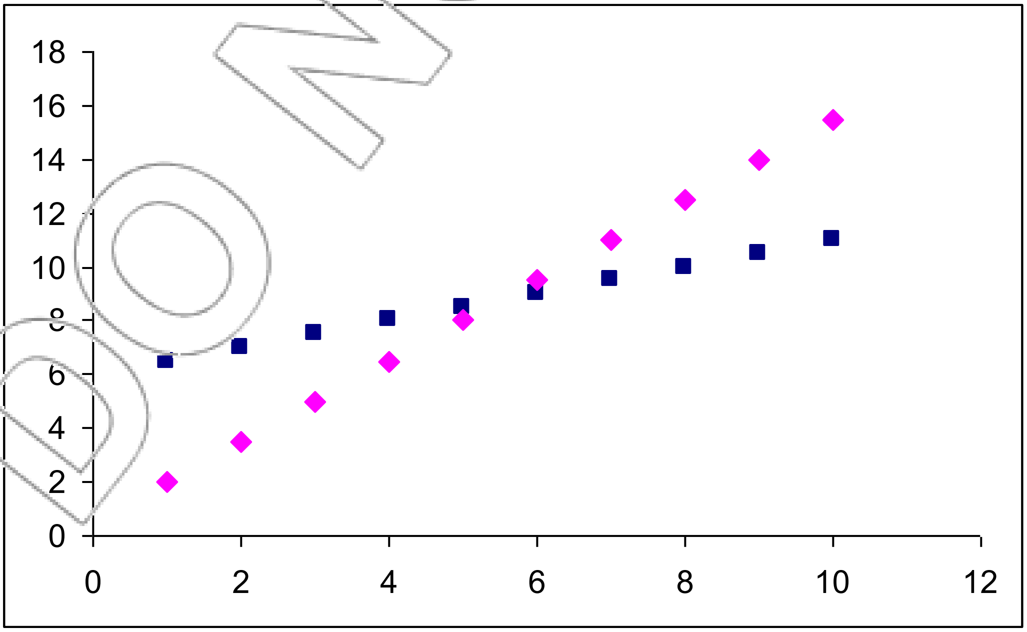


Figure 5.5: Example of data well-modeled by two straight lines.

If the Bakersfield and Monterey systems are different, then the data may fit naturally in two straight lines. One line is associated with the Monterey system, and another line is associated with the Bakersfield system. A dummy variable allows for differences in the intercepts of these two lines. A slope dummy variable allows for differences in the slopes of these two lines. Next we apply these important dummy variable techniques to Susan's problem.

Consider the dummy and slope dummy variables Plant and Boxplant. Plant equals 1 if the data come from the Bakersfield plant and 0 if the data come from the Monterey plant. Boxplant is equal to the variable Boxes if the data come from the Bakersfield plant and 0 if the data come from the Monterey plant (i.e., Boxplant = Plant*Boxes).

If we put all the data together, we obtain Figure 5.6.

Row	Time	Boxes	Plant	Boxplant
1	102	175	0	0
2	69	110	0	0
3	133	225	0	0
4	37	57	0	0
5	28	47	0	0
6	124	217	0	0
7	71	120	0	0
8	36	60	0	0
9	41	65	0	0

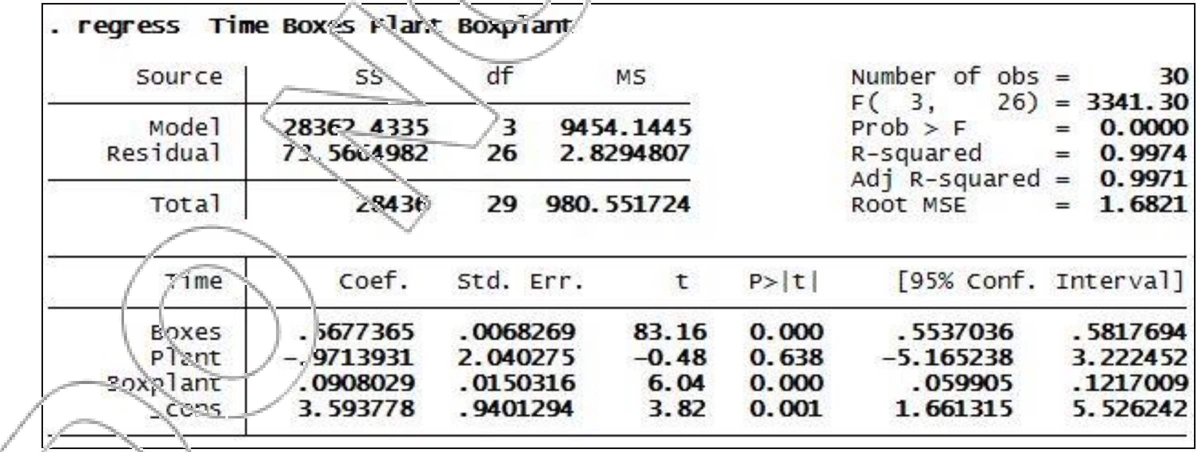
0	104	180	0	0
11	126	210	0	0
12	63	106	0	0
13	34	50	0	0
14	38	60	0	0
15	88	150	0	0
16	95	140	1	140
17	104	153	1	153
18	48	70	1	70
19	108	161	1	161
20	89	128	1	128
21	85	125	1	125
22	90	133	1	133
23	81	122	1	122
24	68	95	1	95
25	98	143	1	143
26	109	161	1	161
27	54	80	1	80
28	85	128	1	128
29	137	205	1	205
30	85	125	1	125

Figure 5.6: Complete dataset for California Strawberries, Inc.

Consider a new regression (see Figure 5.7) making use of all the data. Time is the dependent variable. The independent variables are Boxes and the dummy and slope dummy variables (Plant and Boxplant).

MULTIPLE REGRESSION ANALYSIS INCLUDING A DUMMY AND A SLOPE DUMMY VARIABLE

Examine the results in Figure 5.7. The constant term indicates the time needed to set up the Monterey system. The coefficient on Boxes indicates the additional packing time for each additional box under the Monterey system. The constant plus the coefficient on Plant indicates the time needed to set up the Bakersfield system. The coefficient on Boxes plus the coefficient on Boxplant indicates the additional time to pack each additional box under the Bakersfield system. (This is not obvious. A good exercise to understand dummy and slope dummy variables is to think about the interpretation of these coefficients.)



```
. regress  Time Boxes Plant Boxplant

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  3,    26) = 3341.30
       Model |  28362.4335     3  9454.1445            Prob > F      =  0.0000
    Residual |  73.5664982    26  2.8294807            R-squared     =  0.9974
-------------+------------------------------           Adj R-squared =  0.9971
       Total |       28436    29  980.551724           Root MSE      =  1.6821

------------------------------------------------------------------------------
        Time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Boxes |   .5677365   .0068269    83.16   0.000     .5537036    .5817694
       Plant |  -.9713931   2.040275    -0.48   0.638    -5.165238    3.222452
    Boxplant |   .0908029   .0150316     6.04   0.000      .059905    .1217009
       _cons |   3.593778   .9401294     3.82   0.001     1.661315    5.526242
------------------------------------------------------------------------------
```

Figure 5.7: Multiple regression for California Strawberries, Inc.

For the Monterey system, the regression equation simplifies to the following:

$$\begin{aligned}\text{Time} &= 3.593 + 0.568\ \text{Boxes} - 0.971\ \text{Plant} + 0.091\ \text{Boxplant}\\ &= 3.593 + 0.568\ \text{Boxes} - 0.971\ (0) + 0.091\ (0)\\ &= 3.593 + 0.568\ \text{Boxes}\end{aligned}$$

For the Bakersfield system, the regression equation simplifies to the following:

$$\begin{aligned}\text{Time} &= 3.593 + 0.568\ \text{Boxes} - 0.971\ \text{Plant} + 0.091\ \text{Boxplant}\\ &= 3.593 + 0.568\ \text{Boxes} - 0.971\ (1) + 0.091\ (\text{Boxes})\\ &= 2.622 + 0.659\ \text{Boxes}\end{aligned}$$

These are exactly the same equations as we obtained before using two simple regressions. What is the difference? Our regression equation using dummy and slope dummy variables allows Susan to perform the desired statistical tests, which she could not easily do using two separate regressions.

The key coefficients are the coefficients on the dummy and slope dummy variables. The coefficient on Plant measures difference in the time needed to set up (i.e., the constant term for) the Bakersfield and Monterey systems. The coefficient on Boxplant measures the difference in the time needed to pack each additional box (i.e., the slope term) in the Bakersfield and Monterey systems.

The coefficient on Plant (-0.971) is not significant. The reported p-value is 0.638. Thus, we cannot conclude that the time to set up the Bakersfield system is different than to set up the Monterey system. On the other hand, the coefficient on Boxplant (0.0908) is significant. The reported p-value is 0.000. The p-value for the one-tailed test with alternative hypothesis that the true coefficient on Boxplant is greater than 0 is therefore 0.000 as well. So, we can conclude that the time to pack each additional box under the Bakersfield system is significantly longer than the time to pack each additional box under the Monterey system.

Our conclusions are as follows:

1. The time to pack each additional box under the Monterey system is significantly shorter than the time to pack each additional box under the Bakersfield system.
2. The time to set up the Monterey system is not significantly different than the time to set up the Bakersfield system.
3. Susan decides to abandon the Bakersfield system.

5.3 Head-Hunting Agency

Having finally completed your MBA, you have landed work at a prestigious consulting firm. Your first project is with CEO Seek, a head-hunting agency. CEO Seek looks for CEOs as well as lower-level managers.

To stay ahead of competition, CEO Seek recently came up with a “Within 15 days. Guaranteed!” marketing scheme. The agency wants to guarantee finding a well-suited candidate within 15 days, or the service is free of charge. You are asked to evaluate the scheme and propose possible improvements. Naturally, you have inquired where the number 15 came from. However, the answer you got was, “It’s a nice round number and will catch the eye.’’ This did not satisfy you. You decide to investigate further.

You suspect it is harder to find a CEO to manage a bigger company than one to head a small firm. It is, after all, a more responsible job, involving more skills and experience. So, fewer candidates may be suitable for it.

However, the staff at CEO Seek does not agree with your hypothesis. They had the same idea in the past, and they intensified all searches on behalf of larger clients. This method brought no improvement. Thus, they concluded, no relation exists between the size of the firm to manage and the time needed to find a candidate.

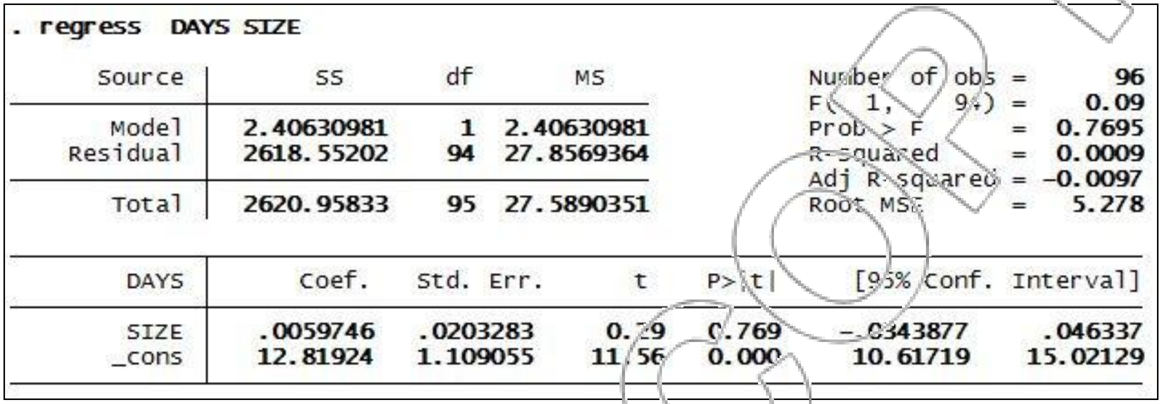
But is it true? You decide to check this hypothesis using regression analysis. From the past performance of the agency, you take a random sample of 48 observations from each of the two categories of searches that CEO Seek conducts: CEO searches and lower-level searches. Each observation includes the size of the firm to be managed and the time it took to produce a well-suited candidate.

The dataset is in the **headhunting** file. In the variable **SIZE**, the size of the client firm is measured in hundreds of employees. **DAYS** denotes the number of days it took CEO Seek to find a suitable candidate. The first 48 observations are from lower-level searches and the remaining 48 observations are from CEO searches.

You would like to use the data to answer the following questions:

1. Is the size of the firm related to the number of days needed to find a suitable candidate? If it is, describe the relationship.
2. What would you recommend concerning the 15-day guarantee?
3. Is it efficient to treat searches for large firms the same as for small ones? If not, do you have any recommendations for improving the system?

Start with a simple regression. DAYS is the dependent variable, and SIZE is the independent variable (see Figure 5.8).



```
. regress  DAYS SIZE

      Source |       SS       df       MS              Number of obs =      96
-------------+------------------------------           F(  1,    94) =    0.09
       Model |  2.40630981     1  2.40630981           Prob > F      =  0.7695
    Residual |  2618.55202    94  27.8569364           R-squared     =  0.0009
-------------+------------------------------           Adj R-squared = -0.0097
       Total |  2620.95833    95  27.5890351           Root MSE      =   5.278

------------------------------------------------------------------------------
        DAYS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SIZE |   .0059746   .0203283     0.29   0.769    -.0343877     .046337
       _cons |   12.81924   1.109055    11.56   0.000     10.61719    15.02129
------------------------------------------------------------------------------
```

Figure 5.8: Simple regression of DAYS on SIZE.

The estimated slope coefficient is 0.006 with a p-value of 0.769. At first glance, there does not appear to be any relationship between the size of the client firm and the number of days CEO Seek took to find a well-suited candidate. This explains why focusing search effort more on searches for larger clients did not improve the system.

Nevertheless, the plot of DAYS and SIZE (see Figure 5.9) indicates the size of the firm and the search time are related. However, there appear to be two relationships; a positive one for CEO searches and a negative one for lower-level management searches.

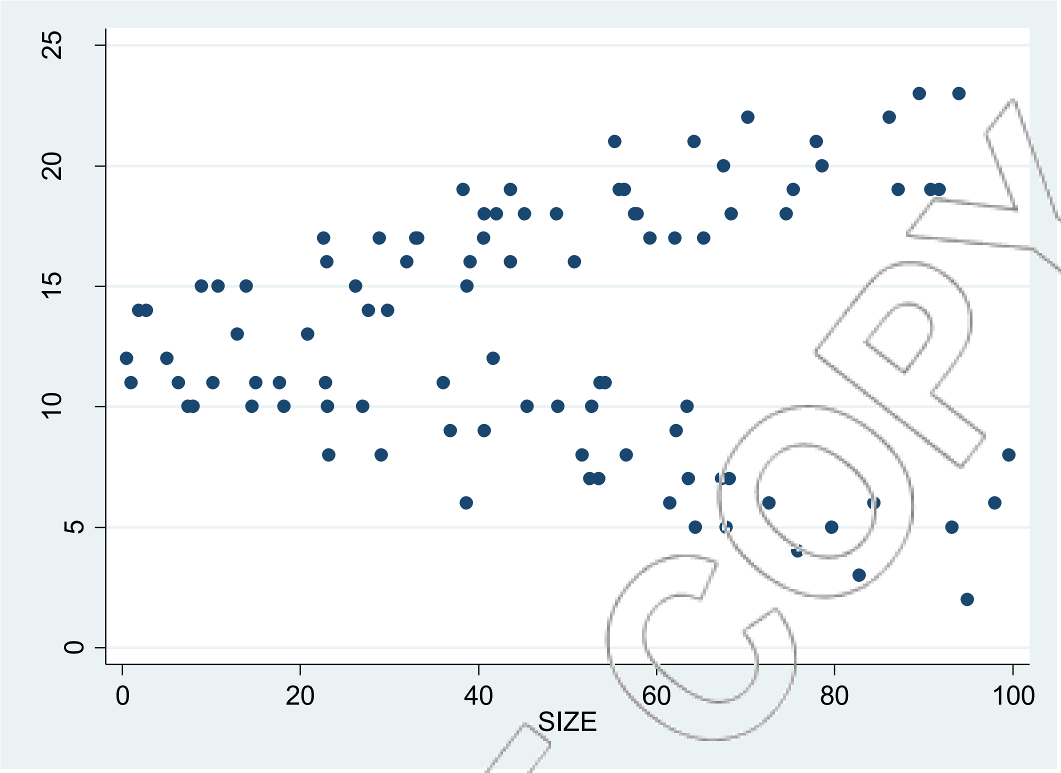
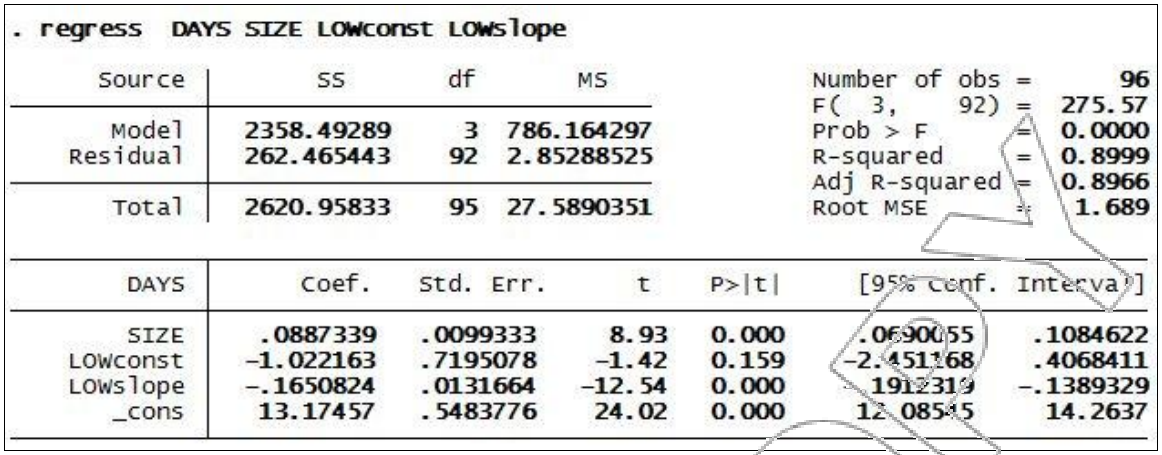


Figure 5.9: Scatterplot of DAYS vs. SIZE.

We could proceed in two ways. One is to run separate simple regressions for CEO and lower managerial positions. The other is to run a multiple regression with a dummy and a slope dummy variable. We choose the latter here because it is more convenient and facilitates comparisons. It would have been fine to do this analysis with separate regressions.

First, we create two new variables. We will call the first one LOWconst . It is equal to 1 if the position is lower-level management and 0 if a CEO is demanded. The second new variable we call LOWslope. It is a slope dummy variable and is the product of LOWconst and SIZE. It is equal to SIZE if the position is lower-level managerial, and it is equal to zero if the position is CEO. Figure 5.10 shows the output from a regression of DAYS on SIZE, LOWconst and LOWslope.



```
. regress  DAYS SIZE LOWconst LOWslope

      Source |       SS       df       MS              Number of obs =      96
-------------+------------------------------           F(  3,    92) =  275.57
       Model |  2358.49289     3  786.164297           Prob > F      =  0.0000
    Residual |  262.465443    92  2.85288525           R-squared     =  0.8999
-------------+------------------------------           Adj R-squared =  0.8966
       Total |  2620.95833    95  27.5890351           Root MSE      =   1.689

------------------------------------------------------------------------------
        DAYS |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SIZE |   .0887339   .0099333     8.93   0.000     .0690055    .1084622
    LOWconst |  -1.022163   .7195078    -1.42   0.159    -2.451168    .4068411
    LOWslope |  -.1650824   .0131664   -12.54   0.000    -.1912319   -.1389329
       _cons |   13.17457   .5483776    24.02   0.000     12.08545     14.2637
------------------------------------------------------------------------------
```

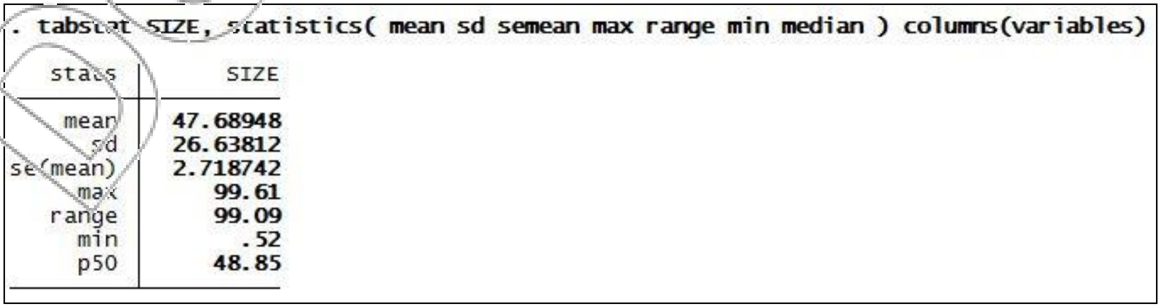
Figure 5.10: Regression of DAYS using dummy and slope dummy variables.

The estimated coefficient on SIZE, 0.0887, is the effect on DAYS of increasing the size of the client firm by 100 employees when looking for a CEO. Testing H_a: $\beta_1 > 0$, we see we are convinced the time to find a suitable CEO candidate increases as the client firm's size grows.

For a lower-level managerial position, the estimated effect on DAYS of increasing the size of a client firm by 100 employees is given by the sum of the coefficients on SIZE and LOWslope or 0.089+ -0.165 = -0.076.

The basic descriptive statistics for SIZE (for CEO and lower-level management searches) can be seen in Figure 5.11.

User>Core Statistics>Univariate Statistics>Custom (tabstat) (or **db tabstat**)



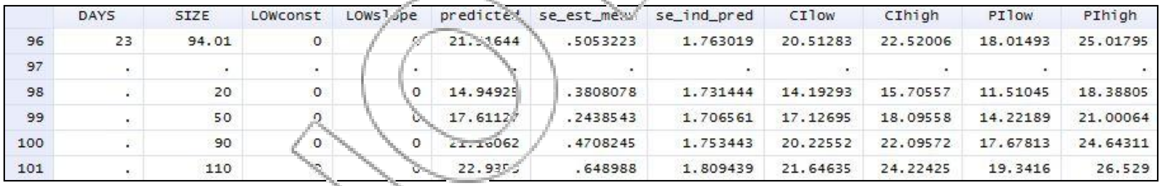
```
. tabstat SIZE, statistics( mean sd semean max range min median ) columns(variables)

    stats |      SIZE
----------+----------
     mean |  47.68948
       sd |  26.63812
 se(mean) |  2.718742
      max |     99.61
    range |     99.09
      min |       .52
      p50 |     48.85
---------------------
```

Figure 5.11: Univariate Statistics for SIZE.

The descriptive statistics tell us client firms have between 0.52 and 99.61 hundred employees. The mean is 47.69 and the median is 48.85. Thus, we can consider a firm where SIZE equals 90 as a large firm and where SIZE = 110 as an exceptionally large firm. A client firm with 2,000 employees is relatively small, while 5,000 is typical.

We will use our new regression with the dummy and slope dummy variable to make predictions about the time needed to find suitable candidates of both categories for different sized clients. The 95% confidence and prediction intervals for time to find a well-suited CEO candidate for firms with SIZE = 20, 50, 90, and 110 respectively can be obtained using Stata (see Figure 5.12):[3]

User>Core Statistics>Univariate Statistics>Prediction, using most recent regression (confint) (or **db confint)**



	DAYS	SIZE	LOWconst	LOWslope	predicted	se_est_mean	se_ind_pred	CIlow	CIhigh	PIlow	PIhigh
96	23	94.01	0	0	21.51644	.5053223	1.763019	20.51283	22.52006	18.01493	25.01795
97	.	.	.	.	.	.	.	.	.	.	.
98	.	20	0	0	14.94925	.3808078	1.731444	14.19293	15.70557	11.51045	18.38805
99	.	50	0	0	17.61127	.2438543	1.706561	17.12695	18.09558	14.22189	21.00064
100	.	90	0	0	21.16062	.4708245	1.753443	20.22552	22.09572	17.67813	24.64311
101	.	110	0	0	22.9353	.648988	1.809439	21.64635	24.22425	19.3416	26.529

Figure 5.12: Predictions for CEO position search times.

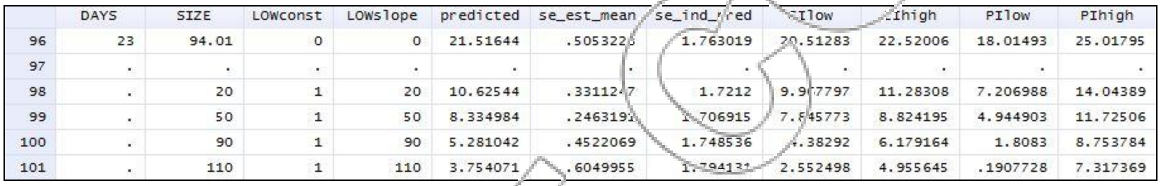
For CEO positions, the lower and upper levels of the confidence and prediction intervals increase as the size of the firm increases.

[3] Before using the **confint** dialog box, you need to enter the values for prediction of 20, 50, 90, and 110 in the **SIZE** column as well as 0's in the **LOWconst** and **LOWslope** columns (since we are interested in CEO positions) in some blank rows (we chose rows 98 through 101).

For firms of all sizes, the upper limits of confidence and prediction intervals are greater than fifteen. Thus, it appears it would be quite costly to attach a 15-day guarantee to CEO-level searches. You would not recommend applying the new guarantee for these searches.

The 95% confidence and prediction intervals for lower-level management searches for firms with SIZE = 20, 50, 90, and 110, respectively, are also easily obtained (see Figure 5.13).[4]

User>Core Statistics>Univariate Statistics>Prediction, using most recent regression (confint) (or **db confint**)



	DAYS	SIZE	LOWconst	LOWslope	predicted	se_est_mean	se_ind_pred	CIlow	CIhigh	PIlow	PIhigh
96	23	94.01	0	0	21.51644	.5053225	1.763019	20.51283	22.52006	18.01493	25.01795
97	.	.	.	.	.	.	.	.	.	.	.
98	.	20	1	20	10.62544	.3311247	1.7212	9.967797	11.28308	7.206988	14.04389
99	.	50	1	50	8.334984	.2463191	1.706915	7.845773	8.824195	4.944903	11.72506
100	.	90	1	90	5.281042	.4522069	1.748536	4.38292	6.179164	1.8083	8.753784
101	.	110	1	110	3.754071	.6049955	1.794131	2.552498	4.955645	.1907728	7.317369

Figure 5.13: Predictions for lower-level management search times.

For the case of lower-level management, the upper and lower levels of the prediction and confidence intervals for time to find a well-suited candidate decrease as the size of the firm increases.

For all sizes of the client firm, the confidence and prediction intervals are below 14.05. Thus, the 15-day guarantee could be offered at little cost for lower-level managerial positions. Therefore, it would be advisable to apply the new policy only for lower-level managerial searches but not for CEO searches.

Our conclusions can be summarized as follows:

[4] To generate the predicted values for lower-level management, change the values for prediction in the **LOWconst** and **LOWslope** columns to those shown in rows 98 through 101 of Figure 5.13. Then, use the **confint** dialog box again.

1. The size of the firm and the search time are related but the relationship depends on the category of employee desired. When a CEO is needed, it takes more time to find a suitable candidate for large firms than for small firms. On the other hand, it takes less time to find a suitable lower-level candidate for large firms than for small firms.

2. The 15-day guarantee policy is quite feasible for the case of lower-level positions. This policy would work poorly for the CEO searches. A longer time horizon for the guarantee should be considered for candidates in this category.

3. We might improve the current system (in terms of reducing the lengthiest searches) by allocating more effort to finding CEO candidates for large firms. Alternatively, CEO Seek might want to solicit more business from small firms looking for CEOs and large firms looking for lower-level management since it seems to handle these searches more efficiently. Since it takes more time to find a CEO candidate than a candidate for a lower managerial position, a policy recognizing the increased difficulty of finding CEOs would be sensible.

SUMMARY

Dummy and slope dummy variables can be used to test statistical differences between the constant and slope coefficients (respectively) of two regressions.

When we have to decide between adopting different systems, these statistical tests are useful. It may not be easy to tell which system is best and these statistical tests help quantify the strength of our evidence for this question.

A single simple regression may be unsuccessful when the relationship between the independent and dependent variables is changed by a third factor. You need dummy and slope dummy variables to deal with this.

Situations in which slope dummy variables can prove useful can often be detected through graphical analysis. The regression output on its own can be inadequate or misleading as in the simple regression in the head-hunting agency case.

NEW TERMS

Dummy variable An artificially constructed variable which takes on the values of zero and one only. Used to quantify non-numerical qualities or categories. When included in a regression, effectively allows the constant to change depending on the value of the dummy variable

Slope dummy variable A variable that takes the value zero in some rows and the value of an independent variable elsewhere. The product of a dummy variable and another variable. When included in a regression, effectively allows the slope on the independent variable used in its construction to change depending on the value of the dummy variable used in its construction

CASE EXERCISES

1. Valuing an MBA for yourself

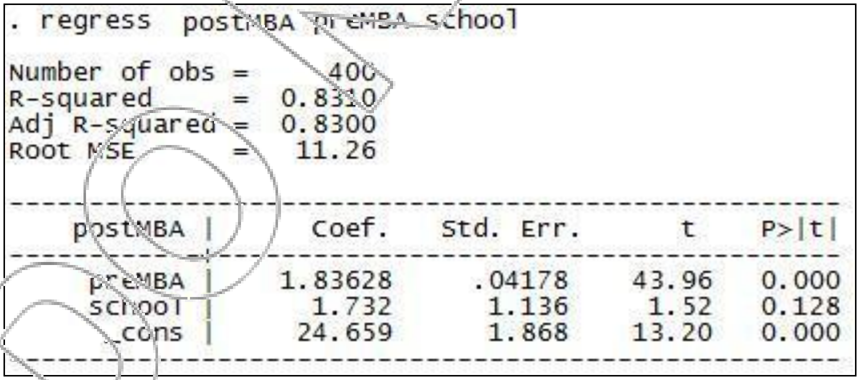
The purpose of this example is to compare the “value-added” of two different business schools by looking at the incomes of the student body prior to beginning the MBA program, and comparing it to the incomes after completing the program. The data consist of information on 400 students, half from school A and the other half from school B.

‘preMBA’ = income in year before beginning the program, in thousands of dollars

‘postMBA’ = income in year after completing the program, in thousands of dollars

‘school’ = a dummy variable equal to 0 for students attending school A, and 1 for students attending school B

The following regression output was obtained:



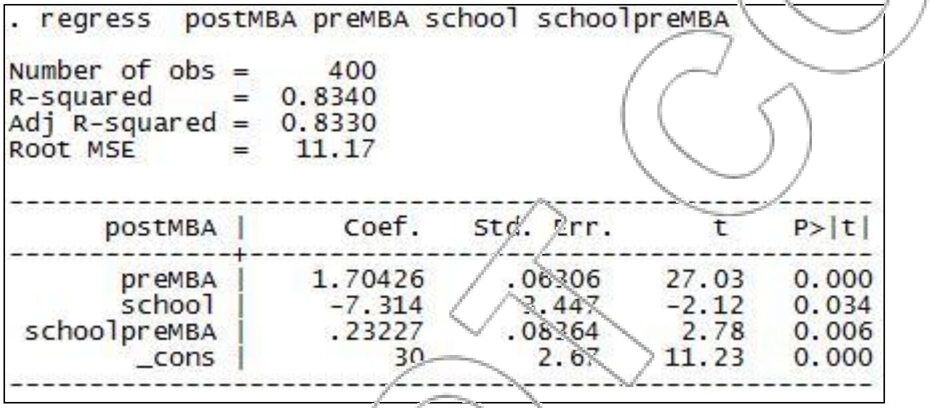
```
. regress  postMBA preMBA school

Number of obs =      400
R-squared     =   0.8310
Adj R-squared =   0.8300
Root MSE      =    11.26

------------------------------------------------------------
    postMBA |       Coef.   Std. Err.       t    P>|t|
------------+-----------------------------------------------
     preMBA |     1.83628     .04178    43.96    0.000
     school |       1.732      1.136     1.52    0.128
      _cons |      24.659      1.868    13.20    0.000
------------------------------------------------------------
```

a. Explain clearly, and as concisely as possible, the interpretation of the coefficient on the school variable.

Suppose we define a new variable as follows:

'schoolpreMBA' = 'school' multiplied by 'preMBA'.

We redo the regression with this extra variable added as another predictor and obtain the following regression output:



```
. regress  postMBA preMBA school schoolpreMBA

Number of obs =      400
R-squared     =   0.8340
Adj R-squared =   0.8330
Root MSE      =    11.17

------------------------------------------------------------------
      postMBA |      Coef.   Std. Err.       t    P>|t|
-------------+----------------------------------------------------
       preMBA |    1.70426     .06306    27.03    0.000
       school |     -7.314      3.447    -2.12    0.034
 schoolpreMBA |     .23227     .08364     2.78    0.006
        _cons |         30       2.67    11.23    0.000
------------------------------------------------------------------
```

Answer the remaining questions, basing your answers on this second regression:

b. Suppose your income this year is $15,000 and you are choosing between the two schools' programs. Assume the two schools have the same fees, similar locations, etc. Which one should you choose? What if your current income is $65,000?

We ask Stata to predict the post-MBA income of someone entering school A with a pre-MBA income of $40,000 and to give 90% confidence and prediction intervals for post-MBA income. This gives the following additional output:

predicted	se_est_mean	CIlow	CIhigh	PIlow	PIhigh
98.171	0.79	96.868	99.474	79.71	116.632

c. What is the predicted post-MBA income of graduates of school A having pre-MBA income of $40,000? If 60 students entering school A this year have pre-MBA incomes of $40,000, about how many of those students do you estimate will make less than $80,000 the year they leave?

d. Explain briefly the meaning of the R-squared statistic in this context (i.e., do not simply say what it means in the abstract, but say what it means for this regression and application).

e. In a few, non-technical words, summarize what the difference seems to be between the two schools.

2. Valuing an MBA for your employer

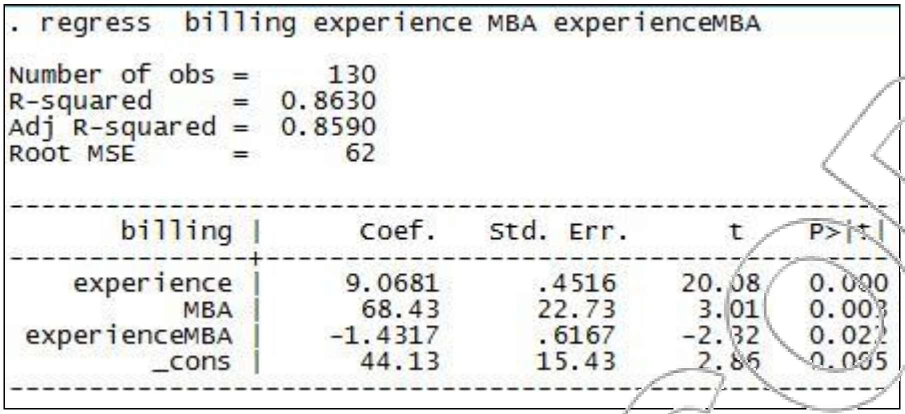
A well-known consulting company is interested in comparing the performance of the consultants it recruits from MBA programs with that of consultants it recruits from non-traditional backgrounds (such as Ph.D. programs). The accounting department has developed a method of allocating all billing to individuals, so it is possible to say how much revenue any given consultant has produced in the last year. You collect data on 130 consultants. For each person, you get three pieces of information, stored as follows:

experience = the length of time they have been with the company (measured in months)
billing = the revenue they brought in in the last year (in thousands of dollars)
MBA = 1 if they came from an MBA program; 0 for those from non-MBA programs

You define a slope dummy variable as follows:

experienceMBA = experience multiplied by MBA

Then, you run the following regression:



```
. regress  billing experience MBA experienceMBA

Number of obs =      130
R-squared     =   0.8630
Adj R-squared =   0.8590
Root MSE      =       62

------------------------------------------------------------
       billing |      Coef.    Std. Err.      t     P>|t|
---------------+--------------------------------------------
    experience |     9.0681       .4516    20.08    0.000
           MBA |     68.43       22.73      3.01    0.003
 experienceMBA |    -1.4317       .6167    -2.32    0.022
         _cons |     44.13       15.43      2.86    0.005
------------------------------------------------------------
```

Answer the following questions.

(a) What do you predict to be the average billing of consultants with two years of experience if they came in with an MBA? What if they came in with a PhD?

(b) Does the extra value to the company of an MBA as compared to a non-MBA change over the time the MBA is with the company? Test at the 1% level of significance.

(c) The sample consists of consultants who have been at the company for up to five years. Suppose you are asked to use your results to predict what the difference in billing (between MBAs and non-MBAs) will be after 10 years. What does the estimated regression equation predict?

(d) Use your judgment: What do you think of this last prediction and why?

PROBLEMS

For problems 1–4, you will need to access the **pizzasales** file.

The Waialua Pizza Company is a medium-sized chain of pizzerias located at beaches all over the South Pacific. The chain is known for its delicious pizzas served at all the nice beaches, and it is known for its use of statistical techniques to improve operations.

The company has obtained data reflecting its sales in its 50 beachfront stores. The Waialua Pizza Company feels the income levels of the nearby community and the presence or absence of competition might be major factors in determining sales.

The following variables were tallied:

Sales = $ per day

Income = Average per-capita income in $ per week in the surrounding neighborhood

Competitor = 1 when one or more competing pizzerias are located within ½ mile; 0 when no other pizzerias are located nearby

1. Conduct a regression of Sales vs. Competitor (only use this one independent variable for now) and use the results to answer the following questions:

a. Estimate the daily sales for a store that has no competition.
b. Estimate the daily sales for a store that faces competition.
c. Calculate the difference between your two estimates and comment on the practical and statistical significance of this gap.
d. Provide a 95% confidence interval for the effect of competition on sales.
e. What percentage of the variance in sales can be explained using only the Competitor variable?

2. Conduct a regression of Sales vs. Income (only use this one independent variable for now) and use the results to answer the following questions:

a. Estimate the daily sales for a store whose neighborhood income is $200 per week.

b. Estimate the daily sales for a store whose neighborhood income is $300 per week.

c. Estimate the impact of a $100 increase in neighborhood income per week on sales.

d. Provide a 95% confidence interval for your estimate in part c.

e. What percentage of the variance in sales can be explained using only the Income variable?

3. Create a scatterplot of Sales vs. Income and plot the regression line as well. Does the picture reveal any likely opportunities to improve your model?

4. Construct a new variable, CompInc, by multiplying the Competitor and Income variables together. Run a regression to predict sales using all three variables: Competitor, Income, and CompInc.

a. Is the Competitor variable in this model statistically significant?

b. Estimate the daily sales for a store without competition whose neighborhood income is $300 per week.

c. Estimate the daily sales for a store with a competitor whose neighborhood income is $300 per week.

d. Compare your answers to part b and part c. Reconcile the results of this comparison with your answer to part a.

5. Access the **eurodata2a** dataset, which is a restructured version of the file **eurodata** used in problems 7–9 in Chapter 4. This file contains information about unemployment and wage growth in Belgium and Denmark. The dummy variable Belgium is set to 1 in Belgium and 0 in Denmark.

Perform a regression of Wage Growth vs. Unemployment, Belgium, and BEUnemployment.[5]

a. Write out the full estimated regression equation.

b. Write out the estimated regression equation for Belgium.

c. Write out the estimated regression equation for Denmark.

d. Compare the equations from part b and c to your answers from Problem 7, Chapter 4.

e. How does a one percentage point increase in unemployment relate to the growth rate of wages in Belgium?

f. How does a one percentage point increase in unemployment relate to the growth rate of wages in Denmark?

g. Estimate the difference in how unemployment relates to wage growth between the two countries.

h. Provide a 95% confidence interval for the difference in how unemployment relates to wage growth between the two countries.

i. Predict the growth rate in wages for each country in a year that has 3% unemployment.

j. Provide a 90% confidence interval for each prediction from part i.

6. Access the **eurodata2b** dataset, which is a restructured version of the file eurodata used in problems 7–9 in Chapter 4. This file contains information about unemployment and wage growth in Germany and Greece. The dummy variable Germany is set to 1 in Germany and 0 in Greece.

[5] BEUnemployment = Belgium*Unemployment

Perform a regression of Wage Growth vs. Unemployment, Germany, and DEUnemployment.[6]

a. Write out the full estimated regression equation.

b. Write out the estimated regression equation for Germany.

c. Write out the estimated regression equation for Greece.

d. Compare the equations from part b and c to your answers from Problem 8, Chapter 4.

e. How does a one percentage point increase in unemployment relate to the growth rate of wages in Germany?

f. How does a one percentage point increase in unemployment relate to the growth rate of wages in Greece?

g. Estimate the difference in how unemployment relates to wage growth between the two countries.

h. Provide a 95% confidence interval for the difference in how unemployment relates to wage growth between the two countries.

i. Predict the growth rate in wages for each country in a year that has 3% unemployment.

j. Provide a 90% confidence interval for each prediction from part i.

7. Access the **eurodata2c** dataset, which is a restructured version of the file eurodata used in problems 7–9 in Chapter 4. This file contains information about unemployment and wage growth in Spain and France. The dummy variable Spain is set to 1 in Spain and 0 in France.

Perform a regression of Wage Growth vs. Unemployment, Spain, and ESUnemployment.[7]

a. Write out the full estimated regression equation.

b. Write out the estimated regression equation for Spain.

c. Write out the estimated regression equation for France.

[6] DEUnemployment = Germany*Unemployment
[7] ESUnemployment=Spain*Unemployment

d. Compare the equations from part b and c to your answers from Problem 9, Chapter 4.

e. How does a one percentage point increase in unemployment relate to the growth rate of wages in Spain?

f. How does a one percentage point increase in unemployment relate to the growth rate of wages in France?

g. Estimate the difference in how unemployment relates to wage Growth between the two countries.

h. Provide a 95% confidence interval for the difference in how unemployment relates to wage growth between the two countries.

i. Predict the growth rate in wages for each country in a year that has 3% unemployment.

j. Provide a 90% confidence interval for each prediction from part i.

CHAPTER 6

FORESTIER WINE: GRAPHICAL ANALYSIS, NON-LINEAR REGRESSION AND SPURIOUS CORRELATION

In this chapter, we will learn how to use graphical analysis to supplement regression. We will study residuals and how to use residual plots to supplement our regression analysis. Additionally, we will expand our regression model's domain of applicability by learning how to conduct one type of non-linear regression. Finally, we will explore the notions of outliers, influential observations, and spurious correlation.

6.1 Snowfall, Unemployment, And Spurious Correlation

The following data (see the **unemploy** file[1]) provides the annual inches of snowfall in Amherst, Massachusetts, and the annual U.S. national unemployment (in %) for the years 1973 to 1982 (see Figure 6.1).

In principle, should we expect any relationship between snowfall in Amherst and U.S. unemployment? Look at the plot of these two variables in Figure 6.2.

Row	Snowfall	Unemployment	Year
1	45	4.9	1973
2	59	5.6	1974
3	82	8.5	1975
4	80	7.7	1976
5	71	7.1	1977
6	60	6.1	1978
7	55	5.8	1979
8	69	7.1	1980
9	79	7.6	1981
10	95	9.7	1982

Figure 6.1: Snowfall data.

[1] From *Statistics for Business and Economics*, by Heinz Kohler, Thomson Learning, 2002.

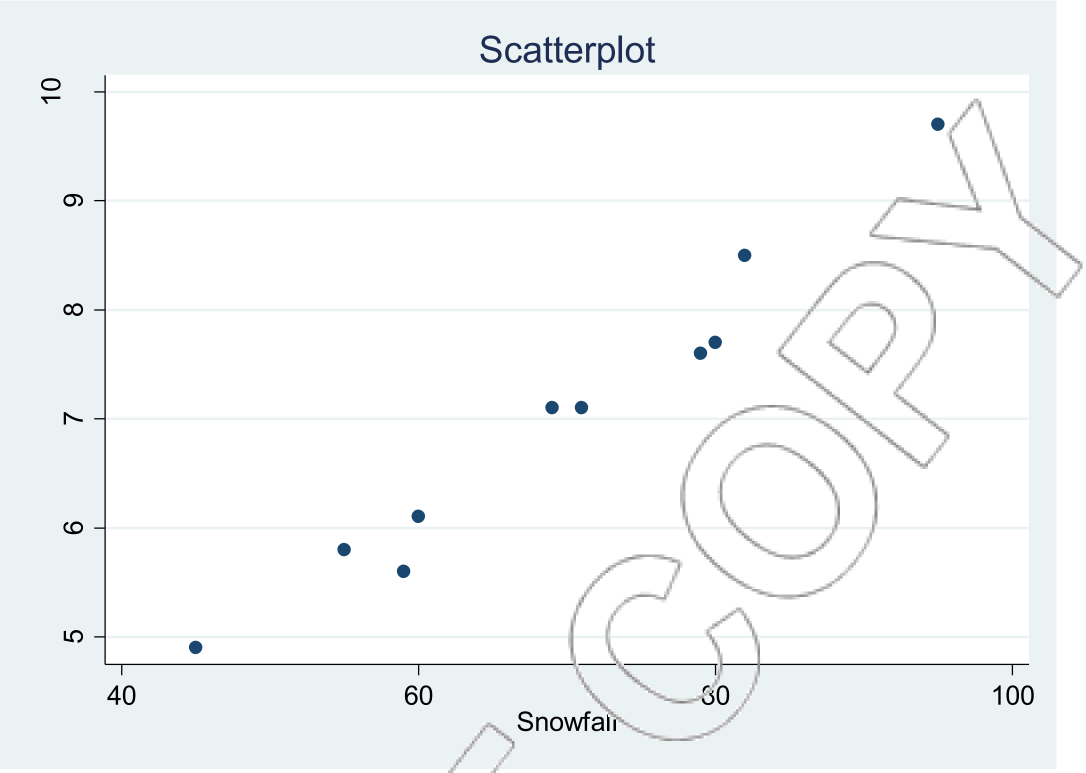
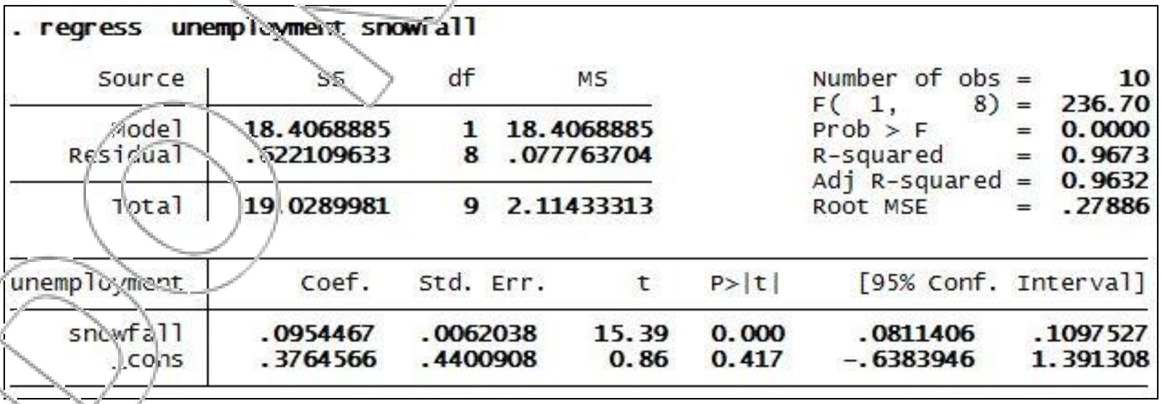


Figure 6.2: Snowfall vs. unemployment in Amherst.

There is clearly a linear relationship between the two variables in the sample, and a regression will do well here (see Figure 6.3).



```
. regress  unemployment snowfall

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =  236.70
       Model |  18.4068885     1  18.4068885           Prob > F      =  0.0000
    Residual |  .622109633     8  .077763704           R-squared     =  0.9673
-------------+------------------------------           Adj R-squared =  0.9632
       Total |  19.0289981     9  2.11433313           Root MSE      =  .27886

------------------------------------------------------------------------------
unemployment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    snowfall |   .0954467   .0062038    15.39   0.000     .0811406    .1097527
       _cons |   .3764566   .4400908     0.86   0.417    -.6383946    1.391308
------------------------------------------------------------------------------
```

Figure 6.3: Regression of unemployment on snowfall.

The R-squared of 0.9673 (96.73%) is exceptionally high, which indicates we are explaining most of the variation in U.S. unemployment. Based on our data, should we conclude that there exists a significant relationship between snowfall in Amherst and U.S. unemployment?

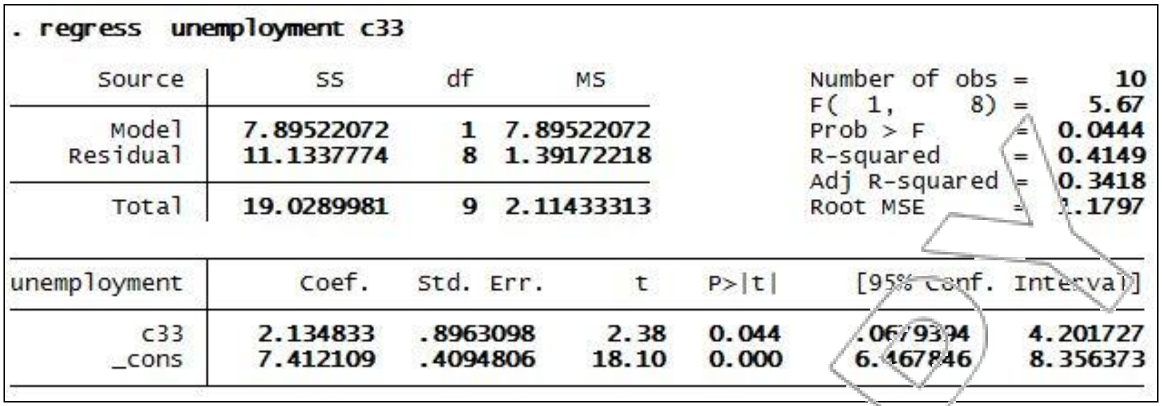
To answer this question we can do a hypothesis test on the slope coefficient to find out if it is significant. The t-statistic is 15.39 and the associated p-value is 0; thus, we reject the null hypothesis that the slope is zero and conclude there is a significant relationship.

This example shows that on occasion, clear patterns pop up at random. Since our inferences are based on data, we will make errors. The relationship between unemployment and snowfall is spurious.

Spurious correlation occurs when the data coming from two unrelated variables are apparently linearly related.

The example suggests that if people want to reach a certain conclusion, and they search for data with this in mind, they can often find a dataset which supports the conclusion.

For example, we generated 40 columns of random data with 10 numbers in each column. We know that none of them are related to unemployment or to any other real dataset because the data was randomly generated in Stata. However, some of the regressions turned out to fit the unemployment data pretty well with the slope coefficient statistically significant at a standard 5% level of significance. For example, a regression relating unemployment and the 33rd randomly generated column turned out this way (see Figure 6.4).



```
. regress  unemployment c33

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =    5.67
       Model |  7.89522072     1  7.89522072           Prob > F      =  0.0444
    Residual |  11.1337774     8  1.39172218           R-squared     =  0.4149
-------------+------------------------------           Adj R-squared =  0.3418
       Total |  19.0289981     9  2.11433313           Root MSE      =  1.1797

------------------------------------------------------------------------------
unemployment |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         c33 |   2.134833   .8963098     2.38   0.044     .0679394    4.201727
       _cons |   7.412109   .4094806    18.10   0.000     6.467846    8.356373
------------------------------------------------------------------------------
```

Figure 6.4: Regression of unemployment on random data.

Our conclusions are as follows:

1. Unemployment and snowfall in Amherst have a statistically significant linear relationship over this period. This relationship is spurious.
2. It is always possible to find a spurious relation between an independent variable and a dependent variable if you try many different independent variables. This occurs because each relationship you examine has some chance of appearing significant due to luck or sampling error even if there is no underlying relationship. Using a level of significance α when testing a single relationship ensures the probability of finding this type of spurious result is at most α. However, if you examine 100 different possible relationships, the probability that at least one of them appears significant even if none of the relationships are real may be as high as $1-(1-\alpha)^{100}$. So, when $\alpha = 0.05$, this probability is $1-(0.95)^{100} = 0.994$.
3. For this reason, always think hard about what variables are sensible to use in a regression analysis before running the regressions. This helps to limit your risk of obtaining spurious results. Similarly, when presented with others' analyses, make sure to find out the process that led to the reported results. If they were the result of searching through a large number of relationships and reporting only significant results, you should be skeptical.

6.2 Wine and Wealth

In this section, we present some simple (yet deceptive) regression examples. The purpose is to motivate techniques that move beyond an examination of the basic regression output.

Robert Owen is the new chief manager of Forestier, a company that produces, markets, and distributes wine. Forestier produces four brands of wine: Almaden, Bianco, Casarosa, and Delacroix. Almaden and Casarosa are high-quality wines. Bianco is a regular wine. Delacroix is a specialty dessert wine sold only in specific locations.

Robert believes that wine sales are directly related to the average household income of the neighborhoods in which the wine shops are located. Robert is considering expanding the business to rich neighborhoods with $15,000 monthly average income. To learn how the various wines are likely to sell in these neighborhoods, he wants to estimate how average income affects sales of the four Forestier brands.

Robert obtained some data on average monthly household income (measured in units of $1,000) and average monthly wine sales (measured in units of $1,000). He has figures from 11 neighborhoods for each brand. The data are in Figure 6.5 and in the **wineandwealth** file.[2]

Almaden		Bianco		Casarosa		Delacroix	
Income A	Sales A	Income B	Sales B	Income C	Sales C	Income D	Sales D
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76

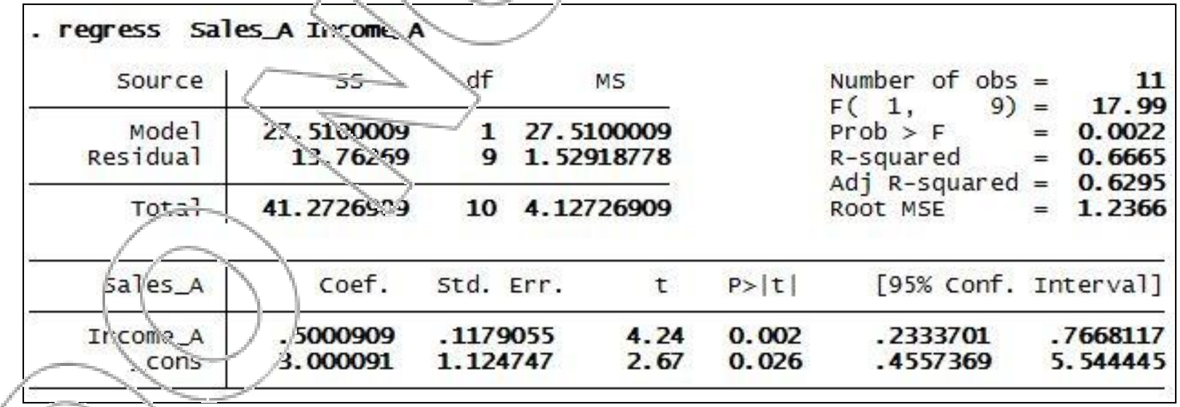
[2] Data adapted from Anscombe, F.J., *Graphs in Statistical Analysis*, American Statistician, (27) February 1973, pp17-21.

13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Figure 6.5 Forestier data.

Robert decides to use regressions to get a feel for the effect of average income on wine sales. He intends to use the regressions to predict wine sales in neighborhoods of $15,000 monthly income.

Consider the Almaden data. Sales A is the dependent variable. Income A is the independent variable (see Figure 6.6).



```
. regress  Sales_A Income_A

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.99
       Model |  27.5100009     1  27.5100009           Prob > F      =  0.0022
    Residual |    13.76269     9  1.52918778           R-squared     =  0.6665
-------------+------------------------------           Adj R-squared =  0.6295
       Total |  41.2726909    10  4.12726909           Root MSE      =  1.2366

------------------------------------------------------------------------------
     Sales_A |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Income_A |   .5000909   .1179055     4.24   0.002     .2333701    .7668117
       _cons |   3.000091   1.124747     2.67   0.026     .4557369    5.544445
------------------------------------------------------------------------------
```

Figure 6.6: Simple regression analysis using the Almaden data.

The regression indicates that monthly sales of Almaden increase, on average, by 50 cents for each extra dollar (equivalently, by $500 for each extra $1,000) in average monthly household income of the neighborhood where the wine shop is located.

The coefficient on Income A (0.50) is statistically significant at our standard 5% level of significance. The t-ratio is 4.24 with a p-value of 0.002.

The regression estimate and 95% confidence and prediction intervals for Almaden sales when Income A is 15 are 10.501, (8.692, 12.310) and (7.170, 13.833), respectively (as you may calculate by entering 15 for Income A in an empty row and clicking the **User>Core Statistics>Prediction, using most recent regression (confint)** menu option[3]). Thus, in any single neighborhood with $15,000 monthly average income, our estimated monthly average sales of Almaden are $10,501, and, with 95% confidence, monthly average sales of Almaden will be between $7,170 and $13,833. Similarly, the average, over the whole population of neighborhoods with $15,000 monthly income, of the monthly average sales of Almaden is between $8,692 and $12,310 with 95% confidence.

Plot Almaden sales and average income (see Figure 6.7). That is, plot Sales A versus Income A. There does not seem to be anything unusual or troubling about this plot. The data seem to fit a generally linear pattern with some variance about the line.

Next, Robert analyzes the effects of average income on Bianco sales. In the next regression (see Figure 6.8), Sales B is the dependent variable and Income B is the independent variable.

[3] You may also type **db confint** instead.

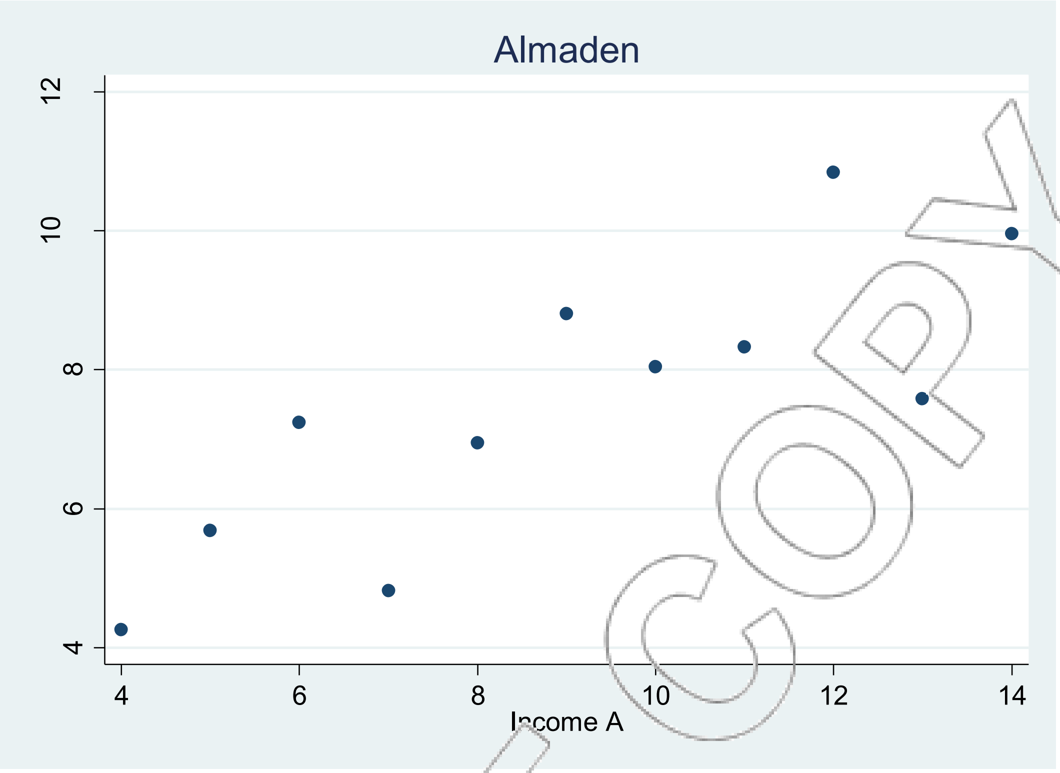
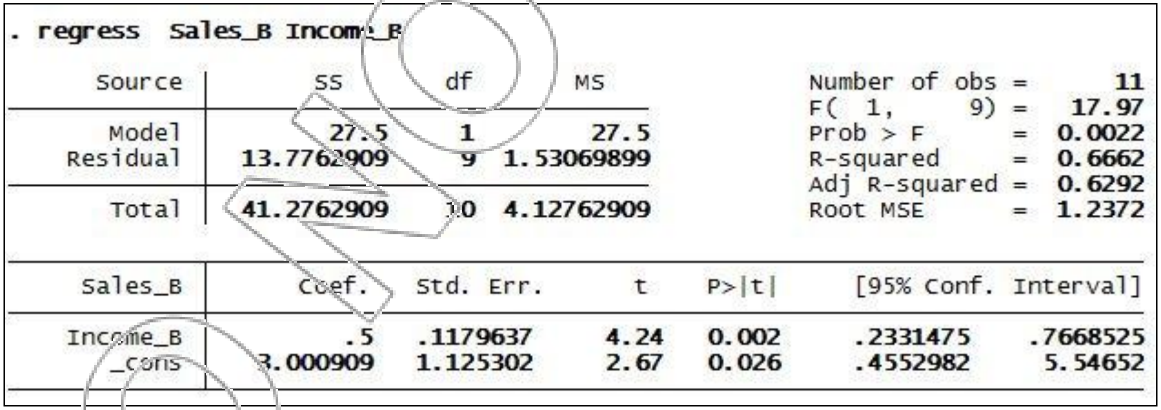


Figure 6.7: Plot of Almaden Sales vs. Income.



```
. regress  Sales_B Income_B

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.97
       Model |        27.5     1        27.5           Prob > F      =  0.0022
    Residual |  13.7762909     9  1.53069899           R-squared     =  0.6662
-------------+------------------------------           Adj R-squared =  0.6292
       Total |  41.2762909    10  4.12762909           Root MSE      =  1.2372

------------------------------------------------------------------------------
     Sales_B |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Income_B |         .5   .1179637     4.24   0.002     .2331475    .7668525
       _cons |   3.000909   1.125302     2.67   0.026     .4552982     5.54652
------------------------------------------------------------------------------
```

Figure 6.8: Simple regression analysis using the Bianco data.

The regression output when using the Bianco data is almost exactly the same as the regression output when using the Almaden data. Thus, the conclusions we would obtain from this regression are the same as the conclusions we obtained from the regression using the Almaden data. In particular, this regression indicates that Bianco monthly average sales increase, on average, by 50 cents for each extra dollar of average monthly household income. The confidence and prediction intervals for Bianco sales are virtually identical to the ones for Almaden.

The data on Bianco sales are different from the data on Almaden, but the regressions using the Bianco and the Almaden data are essentially the same. This seems odd. Robert is puzzled. After all, Almaden is a high-quality wine and Bianco is merely ordinary. Many times, a background graphical analysis can help us understand a regression analysis better. Plot Bianco sales and average income (see Figure 6.9). That is, plot Sales B versus Income B.

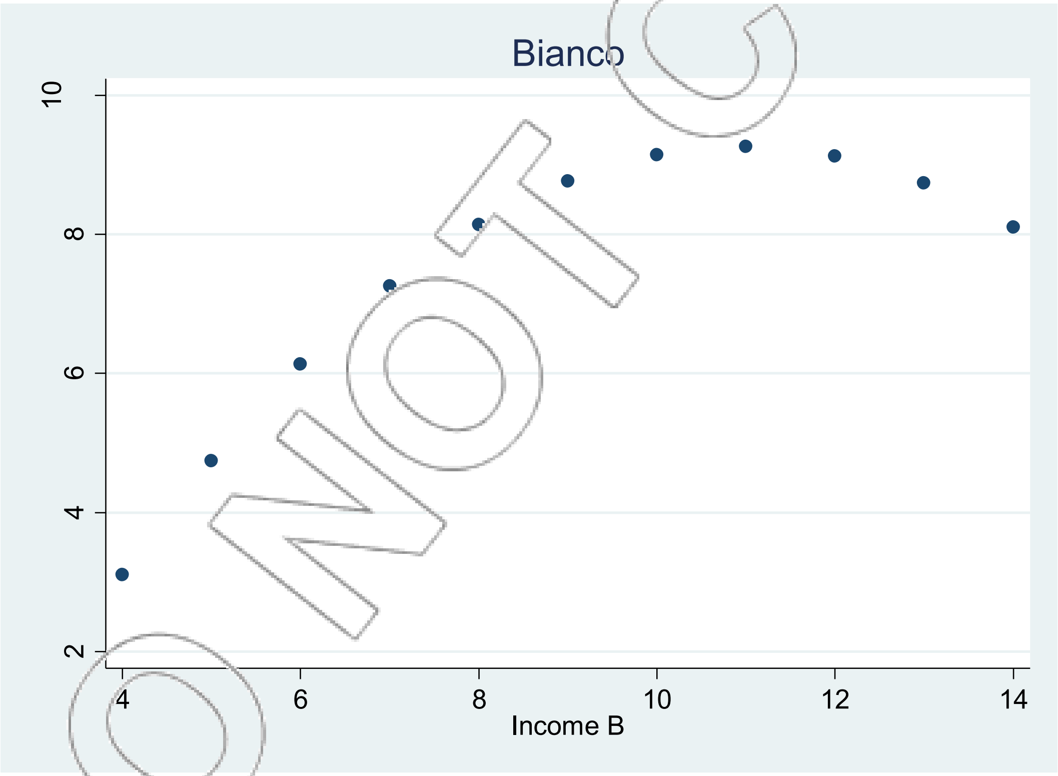


Figure 6.9: Plot of Bianco Sales vs. Income.

The plot clearly indicates that the relationship between Bianco sales and average income is not linear. Thus, one of the most fundamental assumptions of regression (linearity) has been violated. The conclusions we obtained concerning Bianco must be revisited.

The regression using the Bianco sales seems, at first glance, to confirm the conclusion obtained from the regression analysis using the Almaden sales. However, this is incorrect. The effects of average income on Almaden sales are not the same as on Bianco sales. The plots indicate that the Almaden sales are higher if the shops are located in richer neighborhoods. The Bianco sales increase if the wine shops are located in richer neighborhoods but only up to a certain point. After this point, the Bianco sales decrease if the wine shops are located in richer neighborhoods. This probably happens because the quality of the Bianco wine is worse than the quality of the Almaden wine. The crucial point, however, is that the relationship between Bianco sales and average income is non-linear, i.e., not a straight-line relationship.

How can we estimate the effects of average income on Bianco sales when this relationship is non-linear?

It may seem that everything we have learned so far only applies to the linear case, and therefore, these techniques are useless if the relationship between the independent and dependent variable is non-linear. Fortunately, this is untrue: We can apply the techniques we have learned to the case of a non-linear relationship between the independent and dependent variable. One useful and important kind of non-linear relationship is a quadratic relationship. Below, we will learn to use regression to estimate such a relationship.

A **quadratic function** is a function of the form $f(x) = a + bx + cx^2$.

If the coefficient on the squared term is negative, i.e., if $c < 0$, then the plot of the function f looks like an inverted U. For example, Figure 6.10 shows the plot of the function $f(x) = 5+10x-x^2$ for values of x between 0 and 8.

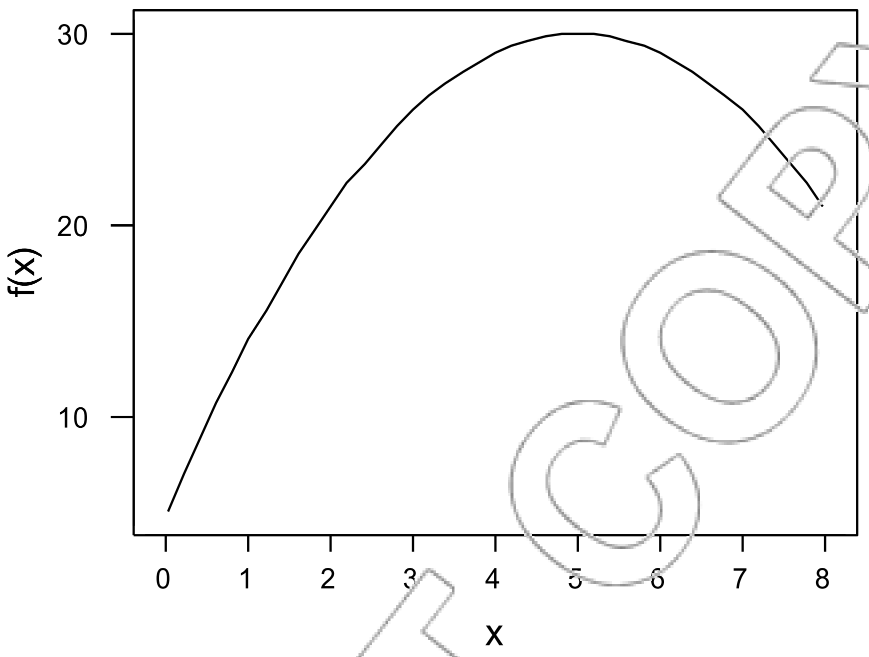


Figure 6.10: Quadratic equation with negative coefficient on the squared term.

On the other hand, if the coefficient on the squared term is positive, i.e., if c > 0, then the plot of the function f looks like a U. For example, Figure 6.11 shows the plot of the function f(x) = 5-10x+x^2 for values of x between 0 and 8.

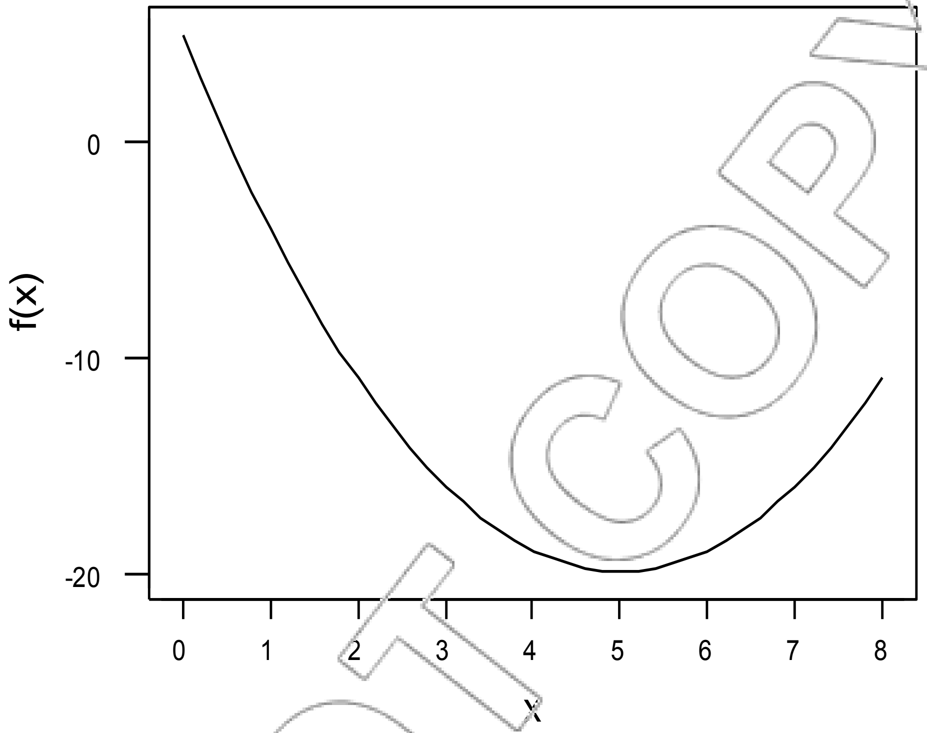


Figure 6.11: Quadratic equation with positive coefficient on the squared term.

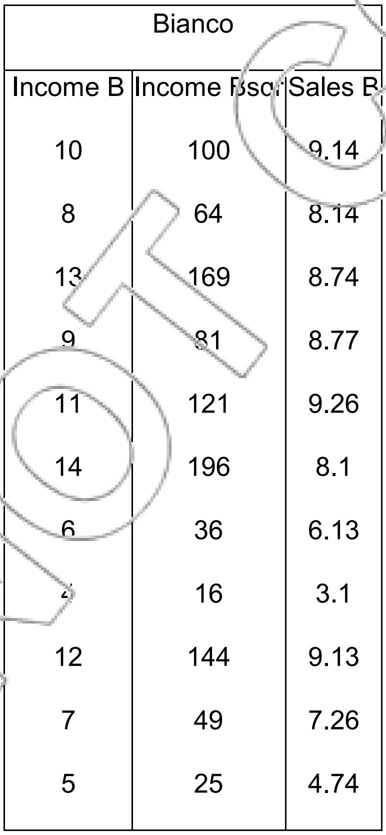
Looking at these plots, we can reasonably conjecture that Bianco sales are a quadratic function (with negative coefficient on the squared term) of the average household income of the neighborhoods in which the wine shops are located. That is, we can reasonably conjecture that Bianco sales and average income are related in the following way:

Bianco sales = a+b(Average Income)+c(Average Income)^2+error

We can estimate the coefficients a, b, and c by running a multiple regression. The dependent variable is Sales B. The independent variables are Income B and Income Bsqr, where Income Bsqr is the square of Income B:[4]

$$\boxed{\text{Income Bsqr} = (\text{Income B})^2}$$

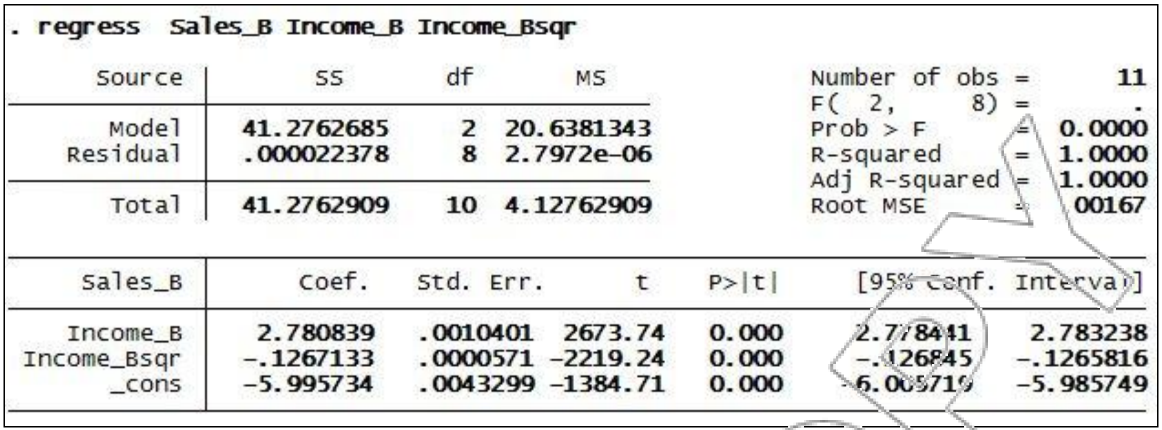
The relevant data for this regression are in Figure 6.12:



Bianco		
Income B	Income Bsqr	Sales B
10	100	9.14
8	64	8.14
13	169	8.74
9	81	8.77
11	121	9.26
14	196	8.1
6	36	6.13
4	16	3.1
12	144	9.13
7	49	7.26
5	25	4.74

Figure 6.12 Bianco data with squared term.

[4] To generate Income_Bsqr in Stata, you can click **User>Manipulate Variables and Obs>Generate New Variable (generate)** or type **db generate**. Type **Income_Bsqr** in the “New variable name” field, and type **Income_B^2** in the “Contents of new variable” field. Alternatively, you can directly type the command **generate Income_Bsqr = Income_B^2**. See the Appendix for detailed explanation on generating new variables in Stata.



```
. regress  Sales_B Income_B Income_Bsqr

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  2,     8) =       .
       Model |  41.2762685     2  20.6381343           Prob > F      =  0.0000
    Residual |  .000022378     8  2.7972e-06           R-squared     =  1.0000
-------------+------------------------------           Adj R-squared =  1.0000
       Total |  41.2762909    10  4.12762909           Root MSE      =  .00167

------------------------------------------------------------------------------
     Sales_B |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Income_B |   2.780839   .0010401  2673.74   0.000     2.778441    2.783238
 Income_Bsqr |  -.1267133   .0000571 -2219.24   0.000     -.126845   -.1265816
       _cons |  -5.995734   .0043299 -1384.71   0.000    -6.005719   -5.985749
------------------------------------------------------------------------------
```

Figure 6.13: Regression analysis of the Bianco data with a quadratic term.

The regression (see Figure 6.13) appears extremely successful in capturing the relationship. In fact, the R-squared is 1 (100%), indicating a perfect fit. The coefficient on the linear term is positive (2.7808) and is significantly greater than zero, and the coefficient on the squared term is negative (-0.1267) and is significantly below zero. This makes sense. The estimated coefficient on the linear term in a quadratic regression is the estimated slope of the relationship when x = 0. Here, this tells us that if average monthly income is close to zero, increasing it by a dollar yields an average of $2.78 in extra sales. Thus, for low levels of income the slope relating income to sales is positive and steep.

The estimated coefficient on the squared term in a quadratic regression tells us how quickly the slope of the relationship changes as x increases. The fact that this coefficient is negative in the example tells us that increases in income provide less of a boost in Bianco sales for higher income neighborhoods than for lower income neighborhoods. We expected these signs for the coefficients because we observed (in Figure 6.9) at low levels of income Bianco sales increase as the average income of the wine shops' neighborhoods increases, but gradually this effect lessens, until, eventually, Bianco sales start decreasing as the average income of the wine shops' neighborhoods increases.

What is the meaning of the constant term? It is our estimate of average sales of Bianco when average monthly household income is zero. The estimated constant (-6) is significantly negative. This does not make sense as a prediction. After all, we should not expect sales to be negative for the wine shops located in extremely poor neighborhoods. However, an examination of the data indicates no such neighborhoods were in our sample for Bianco. Thus, although the quadratic regression appears be an excellent model for incomes closer to the range of our data, we should exercise caution in using our regression equation to forecast Bianco sales in poor neighborhoods.

Robert wants to predict Bianco sales in wine shops located in neighborhoods with $15,000 monthly average income. Using the quadratic regression, the estimated sales when Income B is 15 (and therefore Income Bsqr is $15^2 = 225$) are $7,206 per month. The corresponding 95% confidence and prediction intervals for Bianco Sales are shown in Figure 6.14.

CIlow	CIhigh	PIlow	PIhigh
7.202128	7.210599	7.200635	7.212092

Figure 6.14: 95% confidence and prediction intervals for Bianco sales.

The confidence and prediction intervals are narrow, indicating little error in our sales estimate. The non-linear regression predicts that the average sales will be $7,206 per month. The linear regression predicted average monthly sales of $10,501. The difference is large (almost 50%). It would have been a big mistake to ignore the non-linearity present in the data.

How do we know if a non-linear model should be used? One way is to plot the dependent against the independent variable and look for distinct curvature. We used this method in the Bianco example. Another method (explained below) involves plotting residuals versus predicted or fitted values and examining this plot for distinct curvature. This method is extremely useful, especially if there is more than one independent variable. The reason is simple. Since a plot can have no more than three dimensions, plotting the dependent versus the independent variables is impossible if more than two independent variables are used. Plotting residuals versus predicted values is always possible because the plot remains two-dimensional no matter how many independent variables are used. Examining such plots to detect non-linearities should become a regular supplement to your basic regression analysis.

According to the simple regression model, every observation, y_i, consists of a part that is linear in x, plus an error term:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

In the case of m independent variables, every observation, y_i, consists of a part which is linear in x_1, x_2, …. x_m, plus an error term:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + ... + \beta_n x_{m,i} + \varepsilon_i$$

We use regression to estimate the linear part via the fitted (or predicted) value $\hat{y}$:

$$\hat{y}_i = b_0 + b_1 x_i$$

In the case of multiple regression, the fitted value $\hat{y}$ is given by the following:

$$\hat{y}_i = b_0 + b_1 x_{1,i} + b_2 x_{2,i} + ... + b_n x_{m,i}$$

The **fitted value** (or **predicted value**), $\hat{y}$, is the value of the dependent variable predicted by the regression model.

The **residual** is the difference between the observed value and the fitted value. That is, the residual for the ith observation in our sample, e_i, is given by the following equation:

$$\boxed{e_i = y_i - \hat{y}_i}$$

Since the residuals depend on our estimates (via the fitted values), it makes sense to talk about their sampling distribution. If the standard assumptions of the regression model are correct, the residuals will be normally distributed with a mean equal to zero, a constant variance, and independent of each other.

For the Almaden and Bianco wines, we can use a plot of the residuals to check our linearity assumption. Consider the Almaden data. To plot residuals against the fitted values, we first have to run the regression for Sales A against Income A again since Stata uses only the most recent regression in calculating the residuals and fitted values. Then, click **User>Core Statistics>Model Analysis, using most recent regression>Plot residuals vs predicted values (rvfplot)** or type **db rvfplot**.[5] Click **OK**, and Stata will plot residuals against the fitted values (see Figure 6.15).

[5] Alternatively, you can type **rvfplot** into the Command box and generate the graph without using the dialog box.

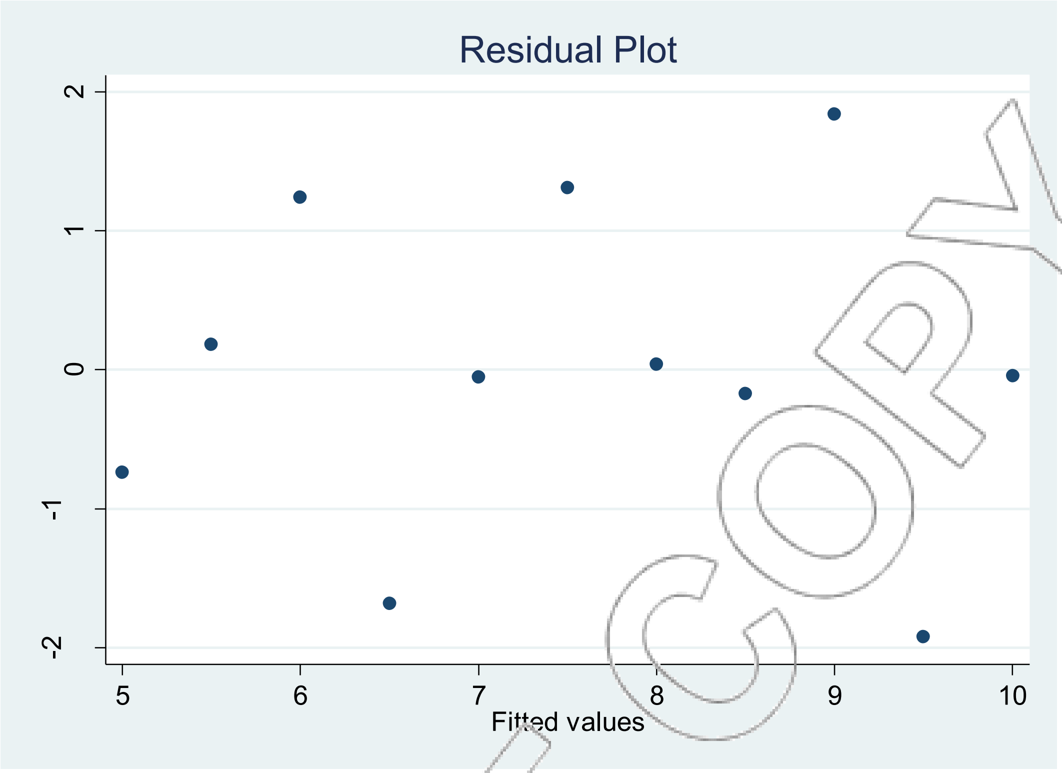


Figure 6.15: Residual plot for Almaden sales.

In this plot, the residuals seem to be displayed at random. No distinct curved pattern can be detected as we move from left to right across the plot. This is a good sign, because it indicates that our linearity assumption appears satisfied.

Consider the Bianco data. A plot of the residuals against the fitted values for the regression without the squared income term reveals distinct curvature (see Figure 6.16).

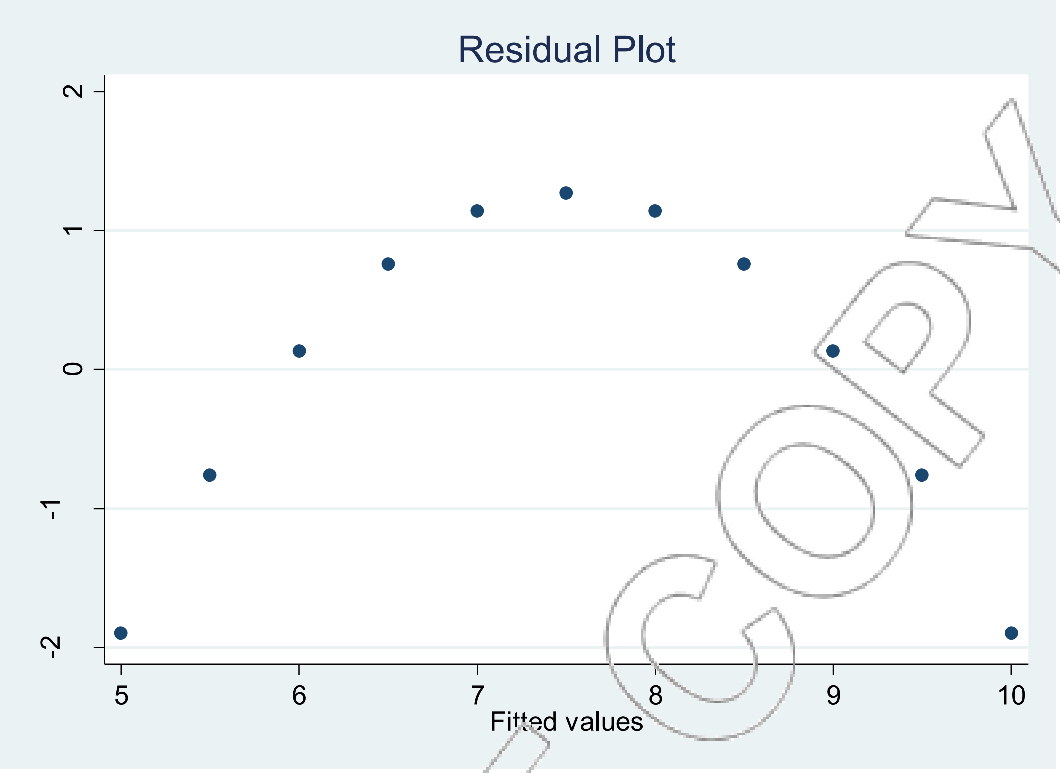


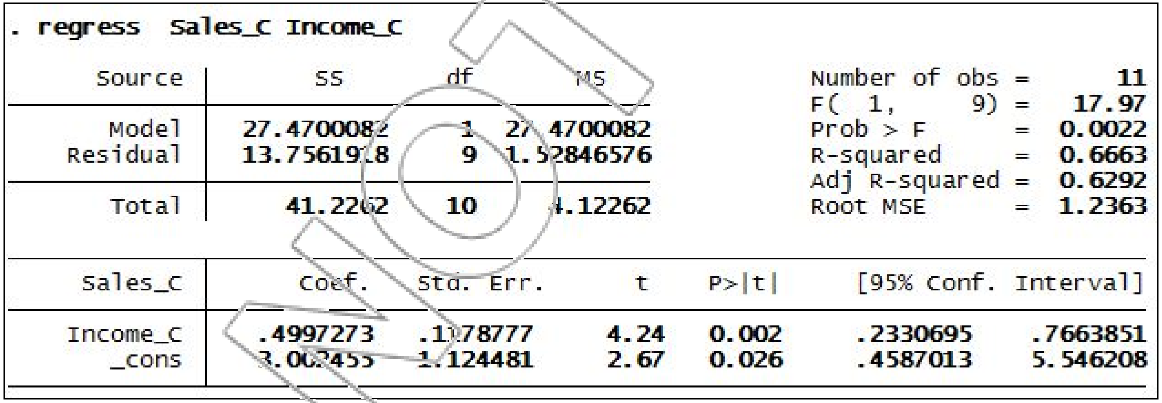
Figure 6.16: Residual plot of Bianco sales with linear model.

All the residuals are negative when the fitted values are low or high. On the other hand, all the residuals are positive for middle fitted values. This inverted-U pattern indicates a non-linear relationship (in fact, a quadratic relationship in this case) between the dependent and independent variables. In general, distinct curvature in the plot of residuals against fitted values suggests a non-linear relationship between the dependent (y) and independent (x) variables.

Try running the quadratic regression using the Bianco data and plotting the residuals versus predicted values from that regression. If the quadratic form is successful in capturing the curvature in the relationship, there should no longer be a distinct curved pattern across the residual plot. You will see that is the case. If distinct curvature had remained, that would have suggested that a model other than the quadratic was needed.

It is important to check the linearity assumption whenever you try a regression model. If distinct curvature is ignored, the regression estimates and standard errors will be biased and may be quite misleading. In addition to checking the linearity assumption, residual plots have another use that we will see in Chapter 7 when we learn how to check the assumption of constant variance.

Now, we will move on and analyze the effects of average income on Casarosa sales. As you can see in Figure 6.17, the regression using the Casarosa data is almost identical to the regressions using the Almaden and Bianco (the linear case) data. Thus, a direct interpretation of the regression would indicate that average monthly sales of Casarosa increase, on average, by 50 cents for each extra dollar of average monthly household income for the neighborhood in which the wine shop is located.



```
. regress  Sales_C Income_C

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   17.97
       Model |  27.4700082     1  27.4700082           Prob > F      =  0.0022
    Residual |  13.7561918     9  1.52846576           R-squared     =  0.6663
-------------+------------------------------           Adj R-squared =  0.6292
       Total |     41.2262    10     4.12262           Root MSE      =  1.2363

------------------------------------------------------------------------------
     Sales_C |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Income_C |   .4997273   .1178777     4.24   0.002     .2330695    .7663851
       _cons |   3.002455   1.124481     2.67   0.026     .4587013    5.546208
------------------------------------------------------------------------------
```

Figure 6.17: Simple regression analysis using the Casarosa data.

The coefficient on income (0.4997) is statistically significant as the t-ratio is 4.24, with a p-value of 0.002. The 95% confidence and prediction intervals evaluated at income of 15 are (8.6898022, 12.30692) and (7.167809, 13.82892), respectively, which are almost identical to the intervals we first obtained with the other two wines.

Plot Casarosa sales against average income (see Figure 6.18). That is, plot Sales C and Income C.

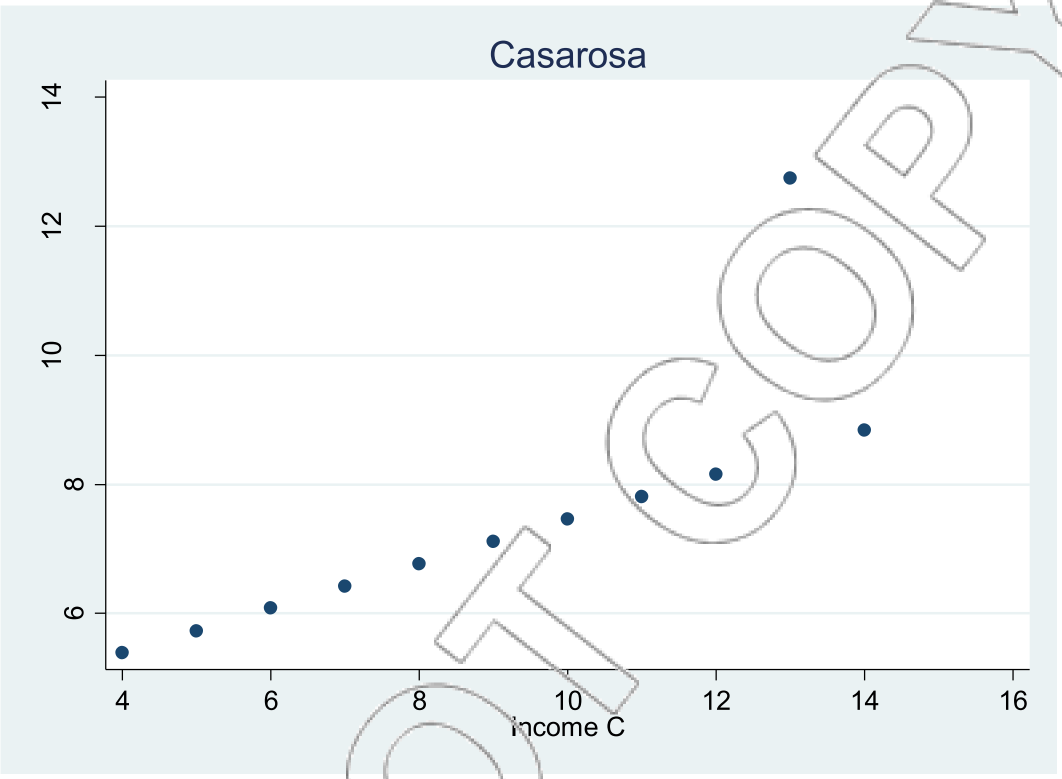


Figure 6.18: Plot of Casarosa Sales vs. Income.

The plot indicates a linear relationship between Casarosa sales and average income, except for one point. In this case, this unusual observation is called an outlier.

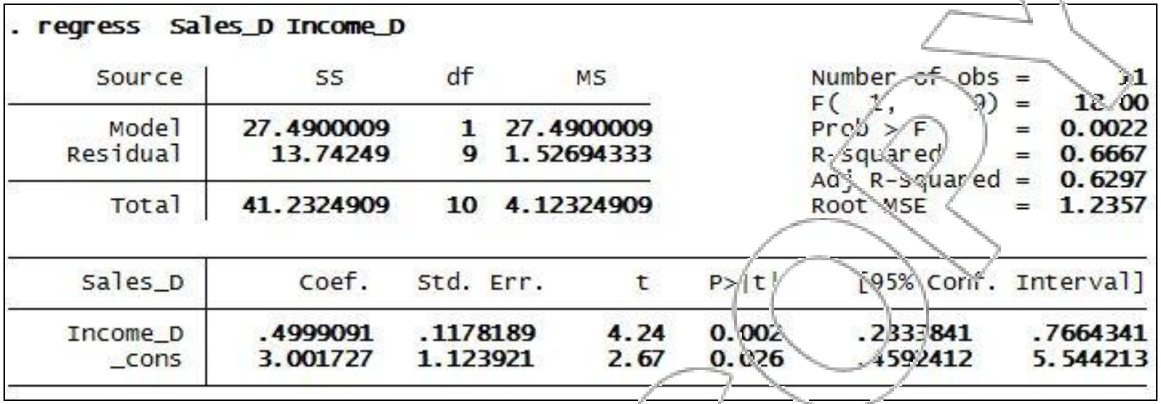
An **outlier** is an observation with an unusually large residual. Stata can identify outliers for you. This is especially useful in multiple regressions or large datasets where they may not be visualized as readily. To have Stata do this, run the regression (here Sales C vs. Income C) and click **User>Core Statistics>Model Analysis, using most recent regression>Residuals, outliers and influential observations (inflobs)**. (You can also type **db inflobs**.) Click **OK** and examine the data browser. The **stdized** column contains the studentized residuals. The studentized residual tells you the number of standard deviations that this residual is from zero, which is the expected value of residuals. The studentized residuals for any outliers will have a value of 1 in the **Ystdized** column. The cutoff for determining if an observation is an outlier can be seen in Stata's Results window, where it is listed under **Flag values** next to **Studentized residual**. In this example, the cutoff is 2.2621572, so any studentized residual with an absolute value above this value generates a 1 in the **Ystdized** column. The formula used to determine the cutoff value is invttail(df, .025), where df is the residual degrees of freedom of your regression. In other words, this cutoff is determined so that if the residuals are normally distributed, approximately 5% of the observations would typically be classified as outliers.

When you encounter outliers (especially if they are large, as in Figure 6.18), you should initially check whether they are due to a mistake such as a data entry error or a measurement error. If that is not the case, it may be worthwhile to try to find out what led to the unusually high or low value: for example, if these are financial data, an outlier might be linked to a stock market crash. In this example, the outlier could be related to a single buyer who is particularly fond of Casarosa wine.

You should not remove outliers from your dataset unless they are due to a mistake: Weird things happen, and it is foolish to pretend otherwise.

On the other hand, if you have a data entry error or a measurement error, then the data should be corrected or removed. In the case of an error, we would have to run a new regression with the corrected data. The results would probably indicate that average Casarosa sales increase by less than 50 cents for each extra dollar on the average income of the wine shops' neighborhood. We can see this in the slope of the line formed by the remaining points being smaller than 0.5.

Finally, we will analyze the effects of average income on Delacroix sales (see Figure 6.19). In this regression, Sales D is the dependent variable and Income D is the independent variable.



```
. regress  Sales_D Income_D

      Source |       SS       df       MS              Number of obs =      11
-------------+------------------------------           F(  1,     9) =   18.00
       Model |  27.4900009     1  27.4900009           Prob > F      =  0.0022
    Residual |    13.74249     9  1.52694333           R-squared     =  0.6667
-------------+------------------------------           Adj R-squared =  0.6297
       Total |  41.2324909    10  4.12324909           Root MSE      =  1.2357

------------------------------------------------------------------------------
     Sales_D |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Income_D |   .4999091   .1178189     4.24   0.002     .2333841    .7664341
       _cons |   3.001727   1.123921     2.67   0.026     .4592412    5.544213
------------------------------------------------------------------------------
```

Figure 6.19: Simple regression analysis on the Delacroix data.

The regression using the Delacroix data is essentially identical to the regression using the Almaden data, the Bianco data (the linear case), and the Casarosa data. A direct interpretation of this regression would lead to the same conclusions as before. However, we have seen that before deriving conclusions from the regression analysis, it is useful to look further.

Again, click **User>Core Statistics>Model Analysis, using most recent regression>Residuals, outliers and influential observations (inflobs)** or type **db inflobs** and examine the Data Browser. One of the values in the **Yleverage** column has a value of 1. This indicates an observation has high **leverage**. The corresponding entry in the **YCook** column is also 1. This indicates an observation has a disproportionately large influence on the regression results. Cook's distance (or **Cook's D**) is a measure of this influence.

Plot Delacroix sales against average income (see Figure 6.20). That is, plot Sales D and Income D.

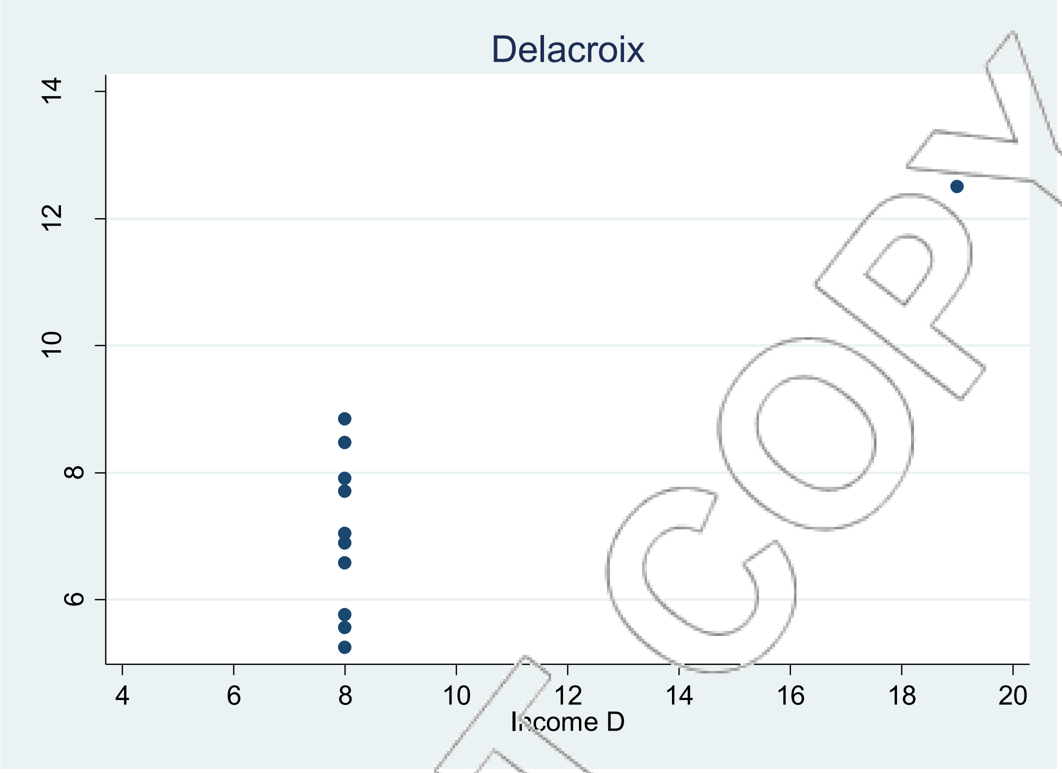


Figure 6.20: Plot of Delacroix Sales vs. Income.

The plot clearly indicates that the regression is entirely driven by a single observation. The estimated regression coefficients would be drastically different if the sales number for just the one influential observation were changed.

An **influential observation** is a data point that has a disproportionately large effect on the regression results.

An influential observation can be an outlier. In this example, however, the influential observation is not an outlier. In fact, the residual associated with the influential observation is zero, i.e., the estimated regression line goes through this point. An influential observation can happen because the point has an unusual *x* value, i.e., one far above or below the average of the *x* values (these are called **high leverage points)**. This is the case here.

As with outliers, you should check that the influential observation is not due to some data error. If it is not due to error, then you should keep it.

It is often a good idea to run the regression with and without an influential observation, and report both. This is a way to explicitly see the influence on the regression estimates. In this example, however, it makes no sense to run a regression without the influential observation. (Can you explain why not?)

Robert should be hesitant to rely on the results from the Delacroix regression. The results are all driven by a single observation. More data are necessary for a reliable analysis. In particular, data from more income levels are needed.

We have shown four different datasets generating the same regression output. These examples demonstrate we have to be careful when analyzing data to guarantee we do not mistakenly miss any of these problems. In addition, since these problems do occur with some regularity in real applications, we must have a "toolbox" of fixes at our disposal.

Our conclusions are as follows:

1. The initial regression output for the Almaden, Bianco, Casarosa, and Delacroix data is the same.

2. The regression using the Almaden data seems to work fine. The analysis predicts average Almaden sales of $10,501 in a neighborhood with average household income of $15,000 a month.
3. The simple regression using the Bianco data is unreliable because the relationship between Bianco sales and average income is curved. Curvature may be detected by examining the plot of residuals versus predicted values. Once a quadratic term is introduced, the regression analysis predicts that Bianco sales should be on average $7,200 in a neighborhood with the average income of $15,000 a month. A further residual plot confirms that the quadratic regression has captured the curvature in the relationship.
4. The regression using the Casarosa data contains an outlier. If there is no error associated with this observation, the regression analysis is identical to the analysis of the regression on the Almaden data.
5. The regression using the Delacroix data is driven entirely by one influential observation. More data on Delacroix sales are necessary for reliable conclusions.

SUMMARY

Spurious correlation occurs when the data indicate a linear relationship that is a statistical artifact (i.e., is due to luck of the draw.) Examples of spurious correlation can be constructed deliberately by generating data at random or (sometimes accidentally) by looking at many different independent variables. This highlights the importance of judgment in constructing and interpreting regressions.

A regression must not be interpreted mechanically. Checking if the underlying assumptions are satisfied is important. If the relationship between dependent and independent variables is non-linear, then we must introduce non-linear terms in our regression. We should also check if outliers and influential observations are associated with some error. These observations should not be modified or deleted unless we find a measurement error or data entry error. Results driven primarily by a few influential observations should be used with care.

NEW TERMS

Spurious correlation The appearance of a significant relationship between unrelated variables

Quadratic function A function of the form f(x) = a + bx + cx^2

Fitted value The value of the dependent variable predicted by the regression model

Residual The difference between the observed value and the fitted value

Outlier A data point that is atypically distant from the regression line. Identified by an unusually large residual

Leverage A measure of how different from the norm the values of the independent variables are for a particular observation

High leverage point An observation whose leverage is more than twice the average for the dataset

Influential observation A data point which has a disproportionately large effect on the regression results

Cook’s D A measure of the influence a data point has on the regression results

NEW STATA FUNCTIONS

[3H]User>Core Statistics>Model Analysis, using most recent regression>Plot residuals vs predicted values (rvfplot)

Equivalently, you may type **db rvfplot**. This command generates a dialog box allowing you to plot the residuals against fitted values following the most recent regression.

Alternatively, you can bypass the dialog box and directly type the command **rvfplot**.

User>Core Statistics>Model Analysis, using most recent regression>Residuals, outliers and influential observations (inflobs)

Equivalently, you may type **db inflobs**. This command creates new variables (default variable names are in parentheses – these can be changed in the dialog box) containing residuals (**residuals**), Studentized residuals (**stdized**), leverage (**leverage**) and Cook's distance (**Cook_D**). It also creates flag (dummy) variables **Ystdized**, **Yleverage**, and **YCook** (again, these are the default names and may be changed). The **Ystdized** column alerts you to outliers by assigning them the value of 1. Observations that are not outliers have the value 0. The **Yleverage** column alerts you to high leverage points by assigning them the value of 1. The **YCook** column alerts you to influential observations by assigning them the value of 1.

An alternate way to generate the residuals, studentized residuals, leverage or Cook’s distance individually following a regression, is to click **Statistics>Postestimation>Predictions, residuals, etc.** and select the quantities of interest. The analogous commands are:

a. **predict *newvar1*, residuals**

b. **predict *newvar2*, rstudent**

c. **predict *newvar3*, leverage**

d. **predict *newvar4*, cooksd**

where ***newvar1-newvar4*** are the names that you want to apply to your respective variables.

CASE EXERCISES

1. The Denny Motors Case

A group of consultants has suggested to Denny Motors that it can predict sales using a forecasting model based on the S&P500. Specifically, as many people view a "Denny" as a luxury good, surges in the stock market may result in subsequent purchases from Denny Motors. After evaluating numerous potential lag times (how long before someone cashes their windfalls into luxury goods is unknown), the consultants have determined that a 30-month lag yields an accurate forecasting model. Specifically, they tried every possible lag time from 0 to 40 months and the highest R-Squared value was found when using a 30-month delay.

Access the data in the **dennymotors** file and run the regression of Denny Motors Quarterly Sales vs. S&P 500 Lagged 30 Months. Knowing that the average value of the S&P during the quarter ending 30 months ago was 1337, construct a 95% prediction interval for next quarter's sales and evaluate its precision. Is it a wide interval or does it seem pretty tight?

Do you agree with the consultants' conclusions?

2. Baseball

A professional baseball team wants to estimate attendance at their ballpark to help make decisions regarding concessions and turnstile revenues. One factor they suspect has an impact on the attendance is weather. The **baseball** data file has attendance data for the first half of the season including both temperature and attendance figures.

Estimate the effect of temperature on attendance. Explore the residuals using the model analysis feature. Are there any obvious explanations for the influential observations? Would removing any outliers improve your model? Can you suggest a way to improve the model without removing any outliers?

3. Television for life

The World Almanac and Book of Facts, 1993, reports the following data on televisions and life expectancy in 38 countries. Access the **tvforlife** file, and conduct a regression predicting life expectancy using TVs per person. Are you surprised by the output? Suggest a possible explanation for these results.

4. Show me the money

Running an agency that represents many professional athletes, you are often forced into serious contract negotiations. One of the baseball players that you represent has had a decent career but has been known to strike out a lot. The team is not offering him a significant contract based on his propensity to strike out more than the other players. To improve your negotiating leverage and to add force to your arguments, you have gathered data to conduct a preliminary analysis of ballplayers' salaries and the number of times they strike out. Your assistant, who has analyzed the data, tells you that every strikeout adds about $14,800 to a player's salary; thus, the assistant suggests encouraging your top players to strike out as often as possible.

The **strikeout** file[6] contains the data on 337 professional baseball players. Use these data to conduct a regression of salary vs. number of strikeouts to replicate the assistant's results. Should you go along with the assistant's suggestion?

[6] "Pay for Play: Are Baseball Salaries Based on Performance?" by Mitchell R. Watnik, *The Journal of Statistics Education*, Volume 6, Number 2 (July 1998).

PROBLEMS

1. Take the dataset from Case Exercise 4 called **strikeout** and run the regression of salary vs. number of strikeouts. Construct a listing of the studentized residuals.

   a. What do the 1’s in the **Ystdized** column tell you about the corresponding observations?

   b. How many studentized residuals large enough to be flagged as 1’s should you expect for a dataset of this size?

2. Access the data in the **burglary** file[7], which contains information about burglary arrests and employment levels in 90 counties in the United States. Conduct a regression of Burglary Arrests vs. Employed (which contains the number of employed people in the civilian workforce in that county.)

   a. What do these results suggest?

   b. Are these results surprising to you?

   c. Identify any counties that are outliers or highly leveraged or influential observations

   d. What is the probability that a normal random variable will be over 6.953 standard deviations from the mean (as the LA County residual is)?

3. Access the **beerdata** dataset[8], which contains data on beer consumption and income levels per capita for 19 European countries. Conduct a regression of beer consumption vs. income levels per capita.

[7] US Department of Justice, Bureau of Justice Satistics at http://www.ojp.usdoj.gov/bjs/dtdata.htm#crime.

[8] See http://www.brewersofeurope.org.

a. On average, as income increases by $1,000 per capita, how much does beer consumption increase?

b. Does this relationship make sense?

c. Identify any outliers in this dataset.

d. How would your answer to part a change if the outliers were removed from the data? (This is generally not a good idea, but we are using the removal of outliers to see how strongly they impact some of our results.)

4. A Midwestern hotel chain has noticed much variation in its electricity costs and would like to be able to explain these changes for planning and budgeting reasons. It has collected samples from random hotels during random months during the past year. The variables include the hotels' electricity costs per room and the average temperature that month. These data are available in the **electricitycosts** file. Conduct a regression of electricity costs per room vs. average temperature.

a. Does the relationship seem significant?

b. Plot residuals versus predicted values for this regression. Does this graph give you any thoughts on improving the model?

c. Use the tools discussed in this chapter to build an improved model.

CHAPTER 7

THE HOT DOG CASE: MULTIPLE REGRESSION, MULTICOLLINEARITY AND THE GENERALIZED F-TEST

In this chapter, we will further our understanding of multiple regression analysis. One new topic is multicollinearity, i.e., strong linear relationships between independent variables in a regression. Specifically, we will learn to use variance inflation factors to detect multicollinearity and use F-tests to test joint significance of regression coefficients. Other topics emphasized include omitted variable bias, hidden extrapolation, and conducting hypothesis tests concerning linear combinations of regression coefficients. Most of this is done in the context of a case involving the analysis of supermarket price data for several varieties of hot dogs.

7.1 The Hot Dog Case

You have just been hired by Dubuque[1], a hot dog manufacturer that produces Dubuque brand hot dogs for the retail market. On your first day at work, you receive a disturbing memo indicating that Ball Park[2], a competing brand, may substantially reduce the price of its hot dog. Dubuque is concerned about the negative impact this might have on its market share.

At the last staff meeting, some of your colleagues argued that Oscar Mayer[3] is Dubuque's leading competitor and that Ball Park's new campaign will not substantially reduce Dubuque's market share. Others, however, disagreed and no consensus was obtained on the strategy that Dubuque should take to protect its market share.

Ball Park produces two kinds of hot dogs. One is a regular hot dog, and the other is a special, all-beef hot dog. The current prices are $1.79 and $1.89 per package, respectively. Dubuque's current price is $1.49 and Oscar Mayer's current price is $1.69.

According to the memo, Ball Park intends to reduce the price of the regular hot dog to $1.45. Two rumors concern the price of Ball Park's special hot dog. One is that Ball Park will slightly increase the price of the special hot dog to $1.95, and the other is that Ball Park will set the price of the special hot dog to $1.55.

You want to predict Dubuque's market share under these different scenarios. Some data are available from a scanner study conducted at grocery stores located in the western suburbs of

[1] Dubuque is a trademark of Hormel Foods Corporation.
[2] Ball Park is a brand of Sara Lee Corporation.
[3] Oscar Mayer is a trademark of Kraft Foods Corporation.

Chicago (see the **hotdog** file). The data were compiled at a weekly level and consist of information on Dubuque's market share (MKTDUB) along with its price (pdub), as well as Oscar Mayer's prices (poscar) and Ball Park's prices (pbpreg and pbpbeef) where pbpreg stands for the price of Ball Park's regular hot dog, and pbpbeef stands for Ball Park's special hot dog. Prices are given in cents (i.e., 135 = $1.35) and market share is given in decimal form (i.e., 0.04 = 4%). There are 113 weeks of data.

Questions:

1. How does Dubuque's price affect its market share?
2. Does Oscar Mayer's price affect Dubuque's market share? If so, how?
3. Does Ball Park's price affect Dubuque's market share? If so, how?
4. Is Ball Park or Oscar Mayer Dubuque's leading competitor? Why?
5. Assume that Dubuque does not respond to Ball Park's new campaign. How much market share is Dubuque expected to lose? In what range is Dubuque's market share expected to be?
6. How much should Dubuque charge for its hot dog to maintain its current market share?

7.2 Hot Dog Case: Solutions, Multicollinearity, Hidden Extrapolation and Tests of Joint Significance

We begin by pointing out an interesting issue present in this data. Examine the correlation between Dubuque's market share and the various prices (see Figure 7.1). Calculating the correlation between two variables is a quick-and-dirty way of estimating the extent of the linear relationship between them. The correlation between Y and X may be found by regressing Y on X, taking the square root of the R-squared (expressed as a decimal), and making it positive or negative depending on the sign of the estimated coefficient multiplying X. Thus, correlations lie between -1 and 1 with correlations further from 0 corresponding to higher R-squared of the regression relating the two variables. The Stata menu option **User>Core Statistics>Bivariate Statistics>Correlations (correlate)** (also accessible by typing **db correlate**) calculates the correlations between each pair of variables in your data and reports them in a table.[4]

Correlations (correlate)

	MKTDUB	pdub	poscar	pbpreg	pbpbeef
MKTDUB	1.0000				
pdub	-0.4329	1.0000			
poscar	0.1695	0.4844	1.0000		
pbpreg	0.3517	0.3593	0.5488	1.0000	
pbpbeef	0.3695	0.3226	0.5337	0.9794	1.0000

Figure 7.1: Correlations.

What signs would we expect the correlations between MKTDUB and the various prices to have? Do we see what we expect?

Note the high correlation between pbpreg and pbpbeef (0.979). In this situation, estimating the separate effects from these two variables is likely to be difficult. When one goes up or down, so does the other: hence, it is difficult to tell if the resulting change in market share is due to pbpbeef or pbpreg. This will play a role in our analysis below.

[4] Alternatively, you can directly type in the command **correlate**. See the list of new Stata functions at the end of the chapter for more details.

Multicollinearity is the term used to describe the presence of linear relationships among the independent variables. A multicollinearity problem occurs when these relationships are strong. We describe it as a problem because it can make it difficult to accurately assess the separate contributions of the strongly related variables to a regression analysis. Specifically, multicollinearity increases the size of the standard errors of the estimated coefficients multiplying the related independent variables. However, we want to emphasize that multicollinearity does not cause any of the basic regression assumptions to be violated. In this sense, it is less serious a problem than the curvature issue discussed in Chapter 6. Multicollinearity simply decreases the precision with which we can estimate some of the regression coefficients.

In this example, we do have a problem of multicollinearity because pbpreg and pbpbeef are highly correlated. In the case of these two variables, the correlation is so strong that it can be seen by looking at the plot between them (see Figure 7.2).

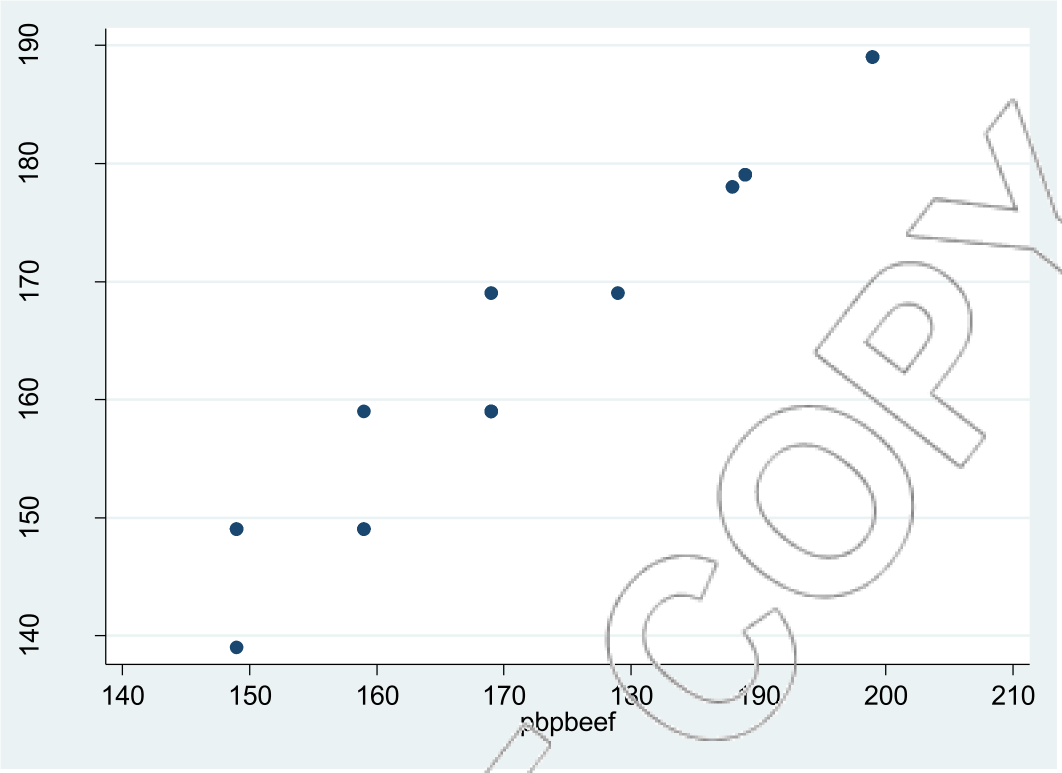
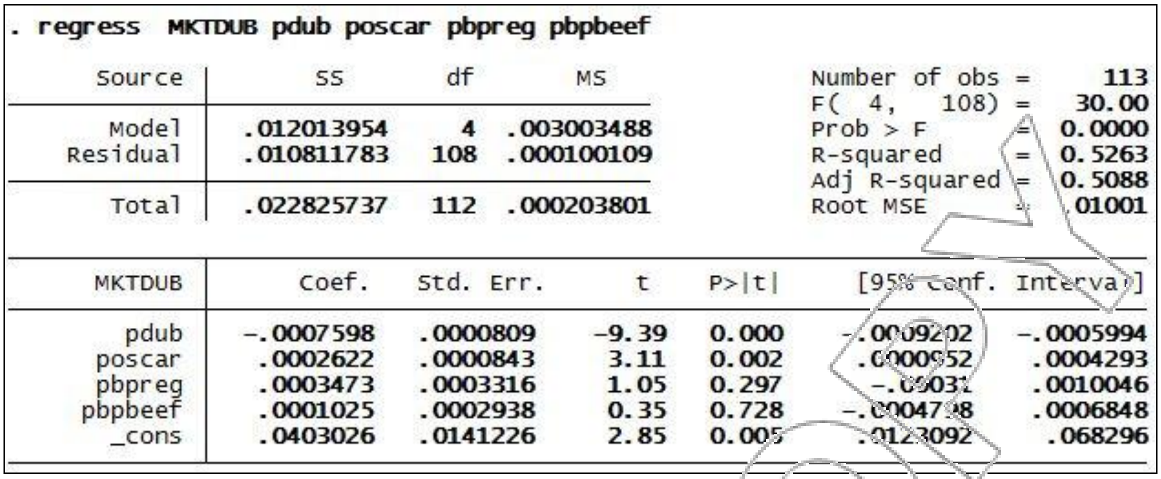


Figure 7.2: Scatterplot of Ball Park's prices.

These two prices move in almost a perfect one-to-one fashion, and so it will be essentially impossible to separate the impact of pbpreg from that of pbpbeef on Dubuque's market share. This is a graphical depiction of the multicollinearity problem we noted above.

Now begin the main analysis by running a regression of MKTDUB on the price variables (see Figure 7.3).



```
. regress  MKTDUB pdub poscar pbpreg pbpbeef

      Source |       SS       df       MS              Number of obs =     113
-------------+------------------------------           F(  4,   108) =   30.00
       Model |  .012013954     4  .003003488           Prob > F      =  0.0000
    Residual |  .010811783   108  .000100109           R-squared     =  0.5263
-------------+------------------------------           Adj R-squared =  0.5088
       Total |  .022825737   112  .000203801           Root MSE      =  .01001

------------------------------------------------------------------------------
      MKTDUB |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pdub |  -.0007598   .0000809    -9.39   0.000    -.0009202   -.0005994
      poscar |   .0002622   .0000843     3.11   0.002     .0000952    .0004293
      pbpreg |   .0003473   .0003316     1.05   0.297      -.00031    .0010046
     pbpbeef |   .0001025   .0002938     0.35   0.728    -.0004798    .0006848
       _cons |   .0403026   .0141226     2.85   0.005     .0123092     .068296
------------------------------------------------------------------------------
```

Figure 7.3: Multiple regression analysis of Dubuque's market share.

The 95% confidence and prediction intervals for market share evaluated at Dubuque's price of $1.49, Oscar Mayer's price of $1.69, Ball Park's (regular) price of $1.45, and Ball Park's (special) price of $1.95 are (0.01636, 0.067146) and (0.009533, 0.073973), respectively.

The 95% confidence and prediction intervals for market share evaluated at Dubuque's price of $1.49, Oscar Mayer's price of $1.69, Ball Park's (regular) price of $1.45, and Ball Park's (special) price of $1.55 are (0.032809, 0.042497) and (0.017238, 0.058069), respectively.

Consider the 95% confidence and prediction intervals for market share evaluated at Dubuque's prices of $1.49, Oscar Mayer's price of $1.69, Ball Park's (regular) price of $1.45, and Ball Park's (special) price of $1.95. The prediction we tried to do is far from typical. This is true, though the values we picked are within the range of the values we have in the data. (You can check this by examining the univariate statistics for the data.) In particular, while pbpreg has been near 145 and pbpbeef has been near 195, they have never been near these values simultaneously. This is an example of a problem called hidden extrapolation.

Extrapolation occurs when the values of the independent variables used for a prediction are far from those in the sample data. **Hidden extrapolation** occurs when these values, as a group, are far from the values in the sample data, even though for each independent variable individually the data seem reasonable enough.

The effect of extrapolation, hidden or not, is to increase $s_{\hat{y}}$, the standard error of the estimated mean, when we predict for such values. This will make our prediction and confidence intervals larger. In this example, quite large. The lower bound of the confidence interval (0.016) is four times smaller than the upper bound of the confidence interval (0.067). Predicting that Dubuque's average market share is expected to be between 1.6% and 6.7% seems not to be helpful. After all, with few exceptions, Dubuque's market share is in this range throughout the data.

Consider the 95% confidence and prediction intervals for market share evaluated at Dubuque's prices of $1.49, Oscar Mayer's price of $1.69, Ball Park's (regular) price of $1.45, and Ball Park's (special) price of $1.55. In this scenario, the prediction and confidence intervals are much narrower. The reason is we do not have a hidden extrapolation problem in this case. The values we are using for prediction are more typical of those in our data.

The lesson to take from this discussion of hidden extrapolation is that predictions using values of the independent variables far from those in the data will be less accurate than those for values more typical of the data. The "hidden" part of hidden extrapolation emphasizes that values for a group of independent variables may be far from those in the data even if the value for each variable individually is close to those in the data.

Now turn to the estimated effects of each price on Dubuque's market share, controlling for, or holding fixed, the other prices. The coefficients of the independent variables have the expected signs. They are positive for the competitors' prices and negative for Dubuque's price. In particular, the coefficient on Dubuque's price is -0.00076. The coefficient on Oscar Mayer's price is 0.000262. The coefficients of Ball Park's prices (regular and special) are 0.000347 and 0.000103, respectively.

Examining the p-values for the coefficient estimates, we see that the coefficient on the constant, Dubuques's price, and Oscar Mayers' price are significantly different from zero. However, the coefficients on the Ball Park prices do not seem to be significant. This is rather curious. The estimated coefficient on Ball Park's regular hot dog price is higher than the estimated coefficient on Oscar Mayer's price. This may indicate Ball Park is Dubuque's main competitor. On the other hand, the coefficient estimates on Ball Park's prices are not significant. This may indicate the opposite. That is, this may indicate our data do not show that Ball Park's prices have any effect on Dubuque's market share.

By looking at the t-ratios and associated p-values for Ball Park's prices, you might think from this first regression that we have little evidence that Ball Park's prices are related to Dubuque's market share. This conclusion seems to support the idea of not reacting to the Ball Park campaign though the estimated coefficient on Ball Park regular hot dog price is higher than the estimated coefficient on Oscar Mayer's price.

However, to decide this issue, we must test if both Ball Park's price coefficients taken together, or jointly, are statistically different from zero. This is particularly important in light of the strong multicollinearity between the Ball Park prices. As observed above, the effect of this multicollinearity is to make it hard to separate the effects of the two Ball Park prices. This appears as an increase in the standard errors of our Ball Park coefficient estimates. The larger standard errors, in turn, result in larger p-values for those coefficients, making them less statistically significant. By giving up on separating the effects of the two Ball Park prices and examining their joint effect on market share, we can sidestep the multicollinearity in the data and, hopefully, arrive at a more precise estimate of the joint effect.

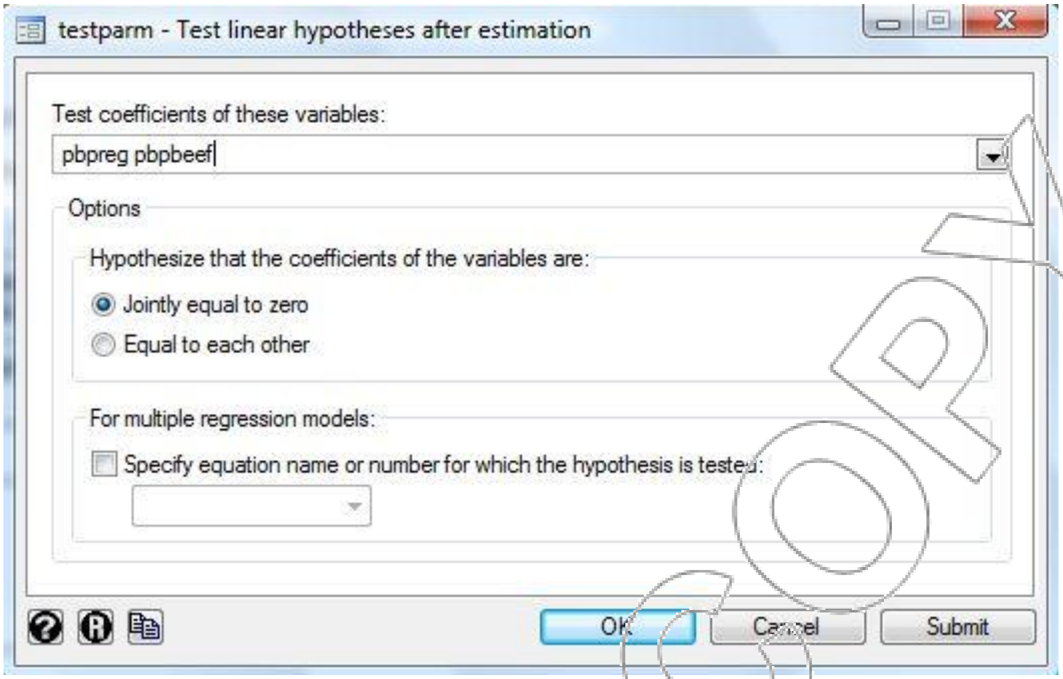
When we want to test whether at least one of a group of coefficients is different from zero, we must consider a hypothesis test called an F-test on the group of coefficients rather than the individual t-tests on each coefficient. As we will see, when x variables are strongly related, the F-test (so-called because the test statistic for this test follows an F distribution if the null hypothesis is true) can give a different answer from the t-tests.

Let's see how we can conduct such a test of joint significance using Stata. Specifically, we will test whether Ball Park's price coefficients taken together, are statistically different from zero. The null and alternative hypotheses are as follows:

H_o: $\beta_{pbpreg} = \beta_{pbpbeef} = 0$

H_a: At least one of β_{pbpreg} or $\beta_{pbpbeef}$ is not equal to zero.

To perform this test, after running the regression, click **User>Core Statistics>Test Hypothesis, using most recent regression>Joint significance (testparm)** (or type **db testparm**). You will obtain the dialog box in Figure 7.4.


testparm - Test linear hypotheses after estimation

Test coefficients of these variables:
pbpreg pbpbeef

Options

Hypothesize that the coefficients of the variables are:
● Jointly equal to zero
○ Equal to each other

For multiple regression models:
□ Specify equation name or number for which the hypothesis is tested:

Figure 7.4: Testparm dialog box.

Select **pbpreg** and **pbpbeef** in the "Test coefficients of these variables" field (in this test, these two variables are also called **added variables**; **pdub** and **poscar** are your **base variables**). Choose **Jointly equal to zero** under the "Hypothesize…variables are" option.[5] When you click **OK**, Stata will run an F-test where the null hypothesis is that coefficients of the added variables (**pbpreg** and **pbpbeef**) are equal to zero, and the alternative hypothesis is that at least one of the coefficients of the added variables is not equal to zero. Stata output for this test is shown in Figure 7.5.

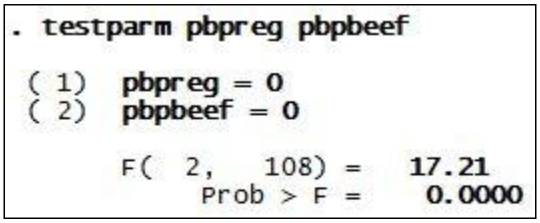

```
. testparm pbpreg pbpbeef

 ( 1)  pbpreg = 0
 ( 2)  pbpbeef = 0

       F(  2,   108) =   17.21
            Prob > F =    0.0000
```

Figure 7.5: Testparm results.

[5] Alternatively, you can directly type the command **testparm *varlist***, where ***varlist*** contains the name(s) of the added variable(s).

This output tells us the p-value (0.0000) associated with this test in the **Prob > F** row. Since the p-value is zero, we reject the null hypothesis:

$$H_o: \beta_{pbpreg} = \beta_{pbpbeef} = 0$$

Therefore, we can conclude that, holding Oscar Mayer's and Dubuque's prices fixed, at least one of the Ball Park prices has an effect on Dubuque's market share.

To understand the example above, we need to have a technical discussion on the use of F-tests. Consider a regression with p independent variables. The data consist of n observations of all the variables.

The regression equation is the following:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q + \beta_{q+1} x_{q+1} + \beta_{q+2} x_{q+2} + \ldots + \beta_p x_p + \varepsilon$$

We want to test if the coefficients $\beta_{q+1}, \ldots, \beta_p$ are jointly significant. The null and alternative hypothesis can be stated as follows:

$$H_0: \beta_{q+1} = 0, \beta_{q+2} = 0, \ldots, \beta_p = 0$$

H_a: One or more of the coefficients (betas) in the null hypothesis is not equal to zero.

Let SSE($x_1, \ldots, x_q, x_{q+1}, \ldots, x_p$) be the error (or residual) sum of squares of the regression equation using all independent variables (the "extended" model).

Let SSE($x_1, \ldots, x_q$) be the error (or residual) sum of squares of the regression equation using only the first q independent variables (the "base" model).

The following F statistic provides the basis for testing whether the additional p-q variables are jointly statistically significant.

$$F = ((SSE(x_1, \ldots, x_q)/SSE(x_1, \ldots, x_q, x_{q+1}, \ldots, x_p))-1)*((n-p-1)/(p-q))$$

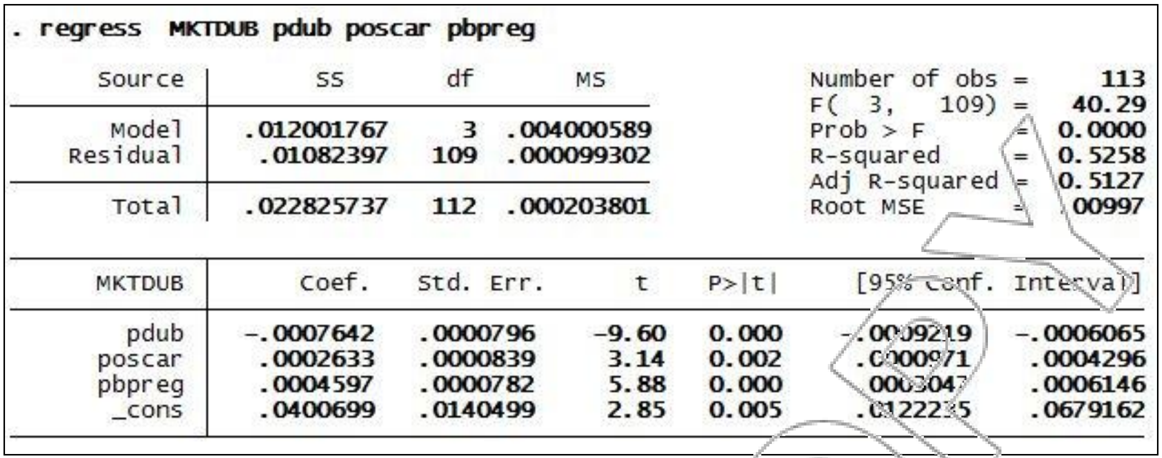
In general, p is the number of variables in the extended model, and q is the number of variables in the base model; thus, p-q is the number of variables being tested.

We have seen that when we run an F-test, Stata gives us the associated p-value for the test. Sometimes, you may only have access to someone else's output where only the F statistic is reported. In this case, you can use Stata's **Ftail** function to find the p-value corresponding to the F statistic. In the hotdog example, the F statistic was 17.21 (see the **F(2, 108)** row in Figure 7.5). To find the corresponding p-value, you can directly type the command **display Ftail(2, 108, 17.21)** (the numbers in the parentheses correspond to p-q (the number of variables being tested), n-p-1 (the degrees of freedom for the extended model with all the variables included), and the F statistic, respectively).

Alternatively, you can use Excel's **FDIST** function to find the p-value corresponding to the F statistic. Click **Insert>Function…**, and choose **Statistical** as the **Function category** and **FDIST** as the **Function name**. Enter the F statistic next to **X**, enter p-q (i.e., the number of variables being tested (= 2 in this example)) next to **Deg_freedom1**, and enter n-p-1 (i.e., the degrees of freedom for the extended model regression with all the variables included (= 108 in this example)) next to **Deg_freedom2**. With the **Formula result**, Excel will give you the p-value. You can also directly type **=FDIST(X, p-q, n-p-1)** into an empty cell and press **Enter**.

This analysis provides an excellent example of the danger of relying too heavily on the significance test of individual coefficients in a multiple regression context. Here, individual t-tests from the original regression would have led us to the incorrect conclusion that neither Ball Park price was significant. The test of joint significance showed that at least one of the Ball Park price coefficients is significant. The joint test does not try to distinguish the effects of the two prices while the individual tests do. The multicollinearity between the two prices explains why the joint test was able to succeed even though the individual tests failed: multicollinearity makes it harder to separate the effects of the two prices.

To carry this discussion a little further, watch what would happen if we run a new regression with only one of the Ball Park prices included, as in Figure 7.6. This is for illustration purposes only. Do not take this to mean that the proper response to multicollinearity is to drop one of the variables. This is not generally correct and, as in this case, may lead to regressions that will be interpreted incorrectly if the multicollinearity present in the original set of variables is not explicitly acknowledged.



```
. regress  MKTDUB pdub poscar pbpreg

      Source |       SS       df       MS              Number of obs =     113
-------------+------------------------------           F(  3,   109) =   40.29
       Model |  .012001767     3  .004000589           Prob > F      =  0.0000
    Residual |   .01082397   109  .000099302           R-squared     =  0.5258
-------------+------------------------------           Adj R-squared =  0.5127
       Total |  .022825737   112  .000203801           Root MSE      =  .00997

------------------------------------------------------------------------------
      MKTDUB |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        pdub |  -.0007642   .0000796    -9.60   0.000    -.0009219   -.0006065
      poscar |   .0002633   .0000839     3.14   0.002     .0000971    .0004296
      pbpreg |   .0004597   .0000782     5.88   0.000     .0003047    .0006146
       _cons |   .0400699   .0140499     2.85   0.005     .0122235    .0679162
------------------------------------------------------------------------------
```

Figure 7.6 Multiple regression analysis without pbpbeef

As you can see from this output, there is almost no qualitative difference in the overall fit of this regression equation. Once we have removed pbpbeef from the regression equation, pbpreg becomes highly significant (p-value = 0). As noted above, it would have been a mistake to have concluded from the results of the first regression that neither variable matters. It follows from the results of the earlier F-test that at least one of the two Ball Park prices does matter, but because of the multicollinearity problem described above, we cannot tell which does matter in the first regression. The coefficient on pbpreg in the regression in Figure 7.6 is approximately the sum of the two Ball Park coefficients in the first regression. You should not conclude from the regression in Figure 7.6 that the effect of Ball Park's regular price on Dubuque's market share is significant. Rather, its coefficient is an estimate of the combined effect of pbpreg and pbpbeef, and we cannot determine which part belongs where.

You should not conclude from this exercise that there was something special about the choice of pbpreg. We could have as easily chosen pbpbeef to leave in the regression. If you do this, the results will be quite similar. This exercise supports the results of our F-test: That the Ball Park prices do matter in determining Dubuque's market share. In the regression with both Ball Park prices, we must remember that the t-ratios should be interpreted recognizing a high degree of multicollinearity.

We can see, from adding together the two Ball Park coefficients in the original regression, that the estimated effect of changing both Ball Park prices by one cent (0.00045) is larger than the estimated effect of changing Oscar Mayer's price by one cent (0.00026). This suggests that Ball Park seems to be Dubuque’s main competitor. Of course, to know if we should be confident in this conclusion, we need to know if the difference between the two estimates is statistically significant. Section 7.3, entitled “Analyzing sums and differences of regression coefficients,” explains how this can be done.

Our responses to the case questions are as follows:

1. Dubuque’s market share falls by an estimated 0.076% for each cent of increase in its hot dog price, holding fixed the Ball Park and Oscar Mayer prices.
2. Dubuque’s market share falls by an estimated 0.026% for each cent of decrease in Oscar Mayer’s price, holding fixed the Dubuque and Ball Park prices.
3. Dubuque’s market share falls by an estimated 0.045% for each cent of decrease in both of Ball Park’s prices, holding fixed the Dubuque and Oscar Mayer prices.
4. Ball Park seems to be Dubuque’s main competitor.
5. Assume that Dubuque does not react to Ball Park’s campaign. Also, assume that Ball Park’s regular hot dog price goes to $1.45, and Ball Park’s special hot dog price goes to $1.55. Dubuque’s average market share is expected to fall by 1.529%. In this case, we are 95% confident that Dubuque’s average weekly market share lies between 3.28% and 4.25%. We are 95% confident that its market share for any given week at these prices will lie between 1.724% and 5.81%.

6. If Dubuque wants to reduce its price to keep its market share, then the correct price reduction will depend upon Oscar Mayer's reaction to Ball Park's campaign. For example, suppose that Oscar Mayer does not change its price. Then, if Ball Park prices are at \$1.45 and \$1.55, Dubuque must reduce its price by approximately 20 cents (≈ market share to make up/market share gained per cent decrease = 1.529%/0.076%).

We can take away two additional lessons from this case:

Ball Park's prices are highly correlated. This creates a multicollinearity problem. As a result, we cannot accurately estimate separate effects for the two Ball Park prices using these data.

Predicting Dubuque's market share is difficult where Ball Park's regular hot dog price is \$1.45 and Ball Park's special hot dog is \$1.95 because of the hidden extrapolation problem. In our sample, these two prices are almost always only 10 cents apart.

7.3 Analyzing Sums and Differences of Regression Coefficients

In the case, we asked: "Who is Dubuque's leading competitor, Ball Park or Oscar Mayer? Why?" Since the sum of the estimated coefficients on Ball Park's two prices was larger than the estimated coefficient on Oscar Mayer's price, it appeared that Ball Park was Dubuque's leading competitor. Because these coefficients are estimates, being able to use statistics to say how confident we are in our conclusion that the effect of a Ball Park price change is larger is important. As usual, we will use a hypothesis test (and the resulting p-value) to evaluate the strength of our evidence. The only twist will be that we will have to use a new test command in Stata to calculate the standard deviation we will need for our test statistic.

Since we would like to know if we have strong evidence that a change in Ball Park's prices has a larger effect on Dubuque's market share than an identical change in Oscar Mayer's price, we should make that the alternative hypothesis. Therefore, using the regression with the four prices as in Figure 7.3, our null and alternative hypotheses are the following:

$$H_0\text{: } \beta_3+\beta_4-\beta_2 \leq 0$$

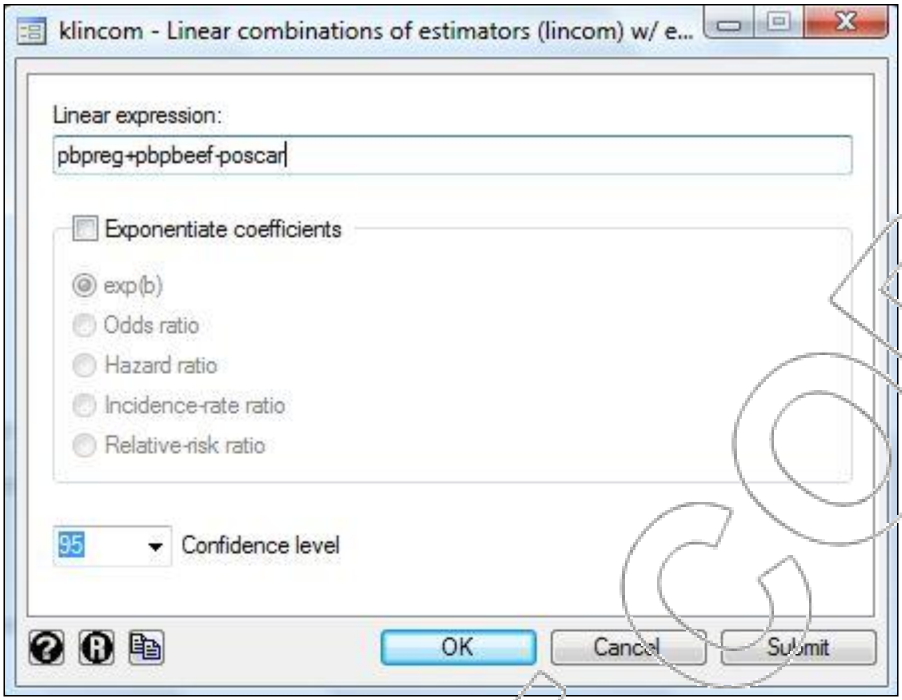
$$H_a\text{: } \beta_3+\beta_4-\beta_2 > 0.$$

Unfortunately, the p-value for such a test is not part of the standard regression output on Stata or any other regression program. However, Stata does have a separate command for us to find the p-value, which we will cover later. As usual, the next step after writing the hypotheses is to calculate the test statistic. The test statistic is similar to those for the hypothesis tests concerning individual coefficients:

$$t = \frac{\text{estimator} - \text{value in the null hypothesis}}{\text{standard deviation of the estimator}} = \frac{b_3 + b_4 - b_2 - 0}{s_{b_3+b_4-b_2}}.$$

If the null hypothesis is true, this will have a t-distribution with degrees of freedom equal to the residual degrees of freedom reported by Stata (= n-# of regression coefficients). So, the only problem is, where can we get the value of $s_{b_3+b_4-b_2}$?

To do this, run the regression of MKTDUB on pdub, poscar, pbpreg, and pbpbeef. Click **User>Core Statistics>Test Hypotheses, using most recent regression>Linear combinations of coefficients (klincom)** or type **db klincom**. This will open the **klincom** dialog box:



klincom - Linear combinations of estimators (lincom) w/ e...

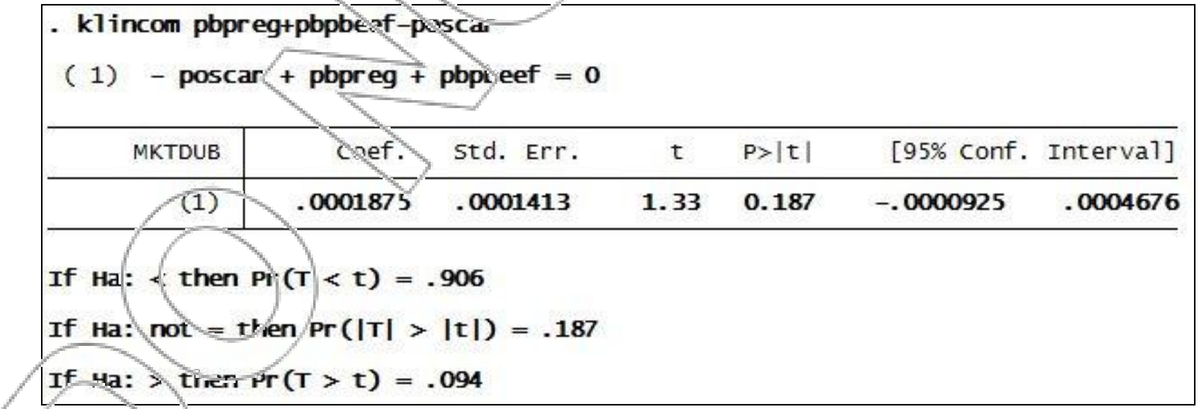
Linear expression:
pbpreg+pbpbeef-poscar

□ Exponentiate coefficients

● exp(b)
○ Odds ratio
○ Hazard ratio
○ Incidence-rate ratio
○ Relative-risk ratio

95 Confidence level

Type **pbpreg+pbpbeef-poscar** into the "Linear expression" field and click **OK**.[6] The Stata output should look like Figure 7.7.



```
. klincom pbpreg+pbpbeef-poscar

 ( 1)  - poscar + pbpreg + pbpbeef = 0

      MKTDUB |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
         (1) |   .0001875   .0001413    1.33   0.187    -.0000925    .0004676

If Ha: < then Pr(T < t) = .906
If Ha: not = then Pr(|T| > |t|) = .187
If Ha: > then Pr(T > t) = .094
```

Figure 7.7: Stata's klincom test output.

[6] Alternatively, you can directly type the command **klincom pbpreg+pbpbeef-poscar**.

First, the value under **Coef.** (0.0001875) is exactly b_3+b_4-b_2 (our estimator). Second, the value under **Std. Err.** (0.0001413) is exactly $s_{b_3+b_4-b_2}$, the standard error (or estimated standard deviation) of our estimator. Therefore, the test statistic for our hypothesis test is 0.0001875/0.0001413 = 1.327, which is precisely the test statistic that Stata reports after rounding (t=1.33). We can calculate the p-value = ttail(108, 1.327) = 0.0936537 or 9.4%. The **klincom** command actually calculates this value automatically and displays it in the last row of Figure 7.7 (**If H_a: > then Pr(T > t) = 0.094**). It looks as if we have fairly strong (though maybe not as strong as we hoped) evidence that Ball Park is our leading competitor.

The method presented here is general and will work for any hypotheses comparing a linear combination of regression coefficients to a number. For example, suppose you wanted to estimate if the effect on our market share would be bigger from a 10-cent drop in the Oscar Mayer price or a reduction in the Ball Park prices of 15 cents on the regular brand and 9 cents on the special hot dog. You would want to compare $-10*\beta_2$ with $-15*\beta_3$-$9*\beta_4$. Therefore, if you were doing a two-tailed test, the alternative hypothesis would be the following:

H_a: $-10*\beta_2+15*\beta_3+9*\beta_4$ 🕒 0.

If you wanted to see if the effect of the Ball Park changes was at least 0.001 larger than the effect of the Oscar Mayer changes, the alternative would be the following:

H_a: $-10*\beta_2+15*\beta_3+9*\beta_4 < -0.001$.

In the first case, you would type **15*pbpreg+9*pbpbeef-10*poscar** in the "Linear expression" field of the **klincom** dialog box.[7] In the second case, you would type **15*pbpreg+9*pbpbeef-10*poscar+0.001**.[8] The Stata output would give you the needed estimated standard deviation (as well as the estimator), test statistic, and the appropriate p-values.

7.4 Detecting Multicollinearity

In the hot dog example, the presence of a multicollinearity problem was clear from looking at the correlation between pbpreg and pbpbeef. However, in general, it may be not so clear if a multicollinearity problem is present. For example, suppose you found the correlation between two independent variables is 0.65 or 0.75. Is there a multicollinearity problem? How can we quantify this? More importantly, looking at the correlation between pairs of variables often may miss important interactions among three or more variables. These can cause multicollinearity problems as well.

Is there an indicator of a multicollinearity problem that may overcome these shortcomings of simple correlations? The answer to this question is the variance inflation factor.

Variance inflation factors measure how much the variance of the estimated regression coefficients are enlarged compared to when the independent variables are not linearly related. For example, suppose the variance of a coefficient is 6, and the variance inflation factor is 2. In this case, the variance of this coefficient should be 3 (6 divided by 2) in the absence of

[7] The direct command would be **klincom -10*poscar+15*pbpreg+9*pbpbeef**.
[8] The direct command would be **klincom -10*poscar+15*pbpreg+9*pbpbeef+0.001**.

multicollinearity. Clearly, the larger the variance inflation factors, the more severe are the multicollinearity problems (i.e., the more that multicollinearity is contributing to the lack of precision in our estimates).

For example, assume the t-ratio of a coefficient estimate is 0.5. In this case, the coefficient might appear to be insignificant. On the other hand, assume the variance inflation factor is 36. This means that the standard deviation of this coefficient is six times (because the square root of 36 is 6) larger than the standard deviation of this coefficient would be in the absence of multicollinearity. The t-ratio is the estimated coefficient divided by its standard deviation. Thus, the t-ratio (0.5) is six times smaller than it would be in the absence of a multicollinearity problem. In conclusion, in the absence of a multicollinearity problem, the t-ratio of this coefficient would be 3 (= 0.5*6) and the coefficient estimate would have been significant. Of course, since multicollinearity is present in our data, we cannot conclude we have significant evidence of an effect. We can say, however, that multicollinearity was severe enough to have led to the insignificance in the t-test.

Consider the same example as before, but now assume the variance inflation factor is 4. In this case, the t-ratio of the coefficient would be only 1 in the absence of multicollinearity.

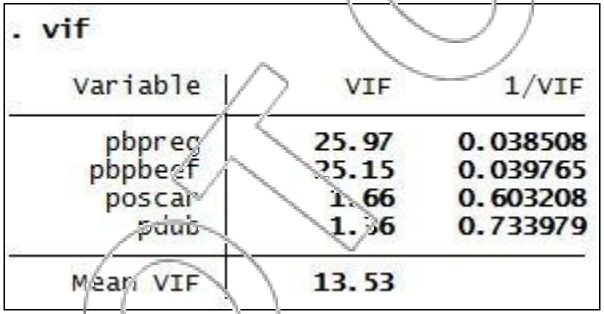
A threshold often used for the variance inflation factor is 10. That is, if the variance inflation factor is above 10, then a serious multicollinearity problem exists in the data.

To obtain the variance inflation factors using Stata, after running a regression click **User>Core Statistics>Model Analysis, using most recent regression>Variance Inflation Factors (vif)**.[9]

[9] Alternatively, you can directly type the command **vif** or **db vif**.

Click **OK**, and Stata will report the variance inflation factors for all independent variables. To illustrate how to check the variance inflation factors, we will reexamine the hot dog regression.

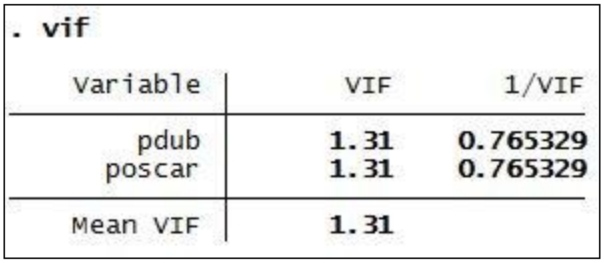
Consider the regression with all the prices (see Figure 7.3). MKTDUB is the dependent variable. The independent variables are all four of the price variables. The variance inflation factors may be found in Figure 7.8 in the **VIF** column. The variance inflation factors of the two Ball Park prices are 25.97 and 25.15. These are well above 10. Therefore, as we determined before, a multicollinearity problem exists in this regression and the two Ball Park prices are the multicollinear variables.



```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      pbpreg |     25.97    0.038508
     pbpbeef |     25.15    0.039765
      poscar |      1.66    0.603208
        pdub |      1.36    0.733979
-------------+----------------------
    Mean VIF |     13.53
```

Figure 7.8: Variance inflation factors for the Hot Dog case.

Consider another regression. MKTDUB is once more the dependent variable. The independent variables are all the price variables except the Ball Park prices. The variance inflation factors are the following:



```
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
        pdub |      1.31    0.765329
      poscar |      1.31    0.765329
-------------+----------------------
    Mean VIF |      1.31
```

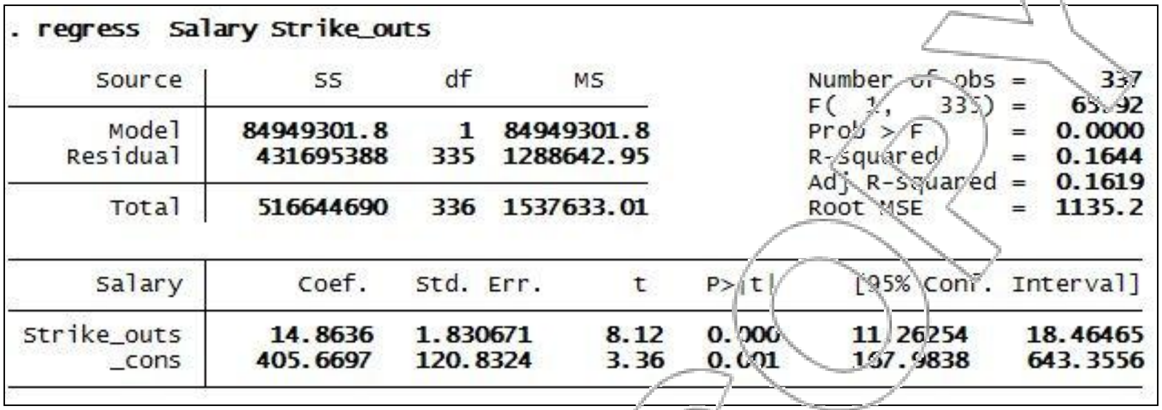
The variance inflation factors of Dubuque’s price and Oscar Mayer’s price are 1.31; therefore, both the variance inflation factors are below 10. This indicates we do not have a serious multicollinearity problem in this regression.

7.5 Omitted Variable Bias

Multicollinearity can make it difficult to obtain precise estimates of the coefficients of strongly related variables in the regression equation. A different and often more serious problem can occur if we leave out one or more related independent variables from a regression. This is called an **omitted variable bias** and we’ve seen it at work in the refrigerator case and some of the case exercises in Chapter 6.

Examine Case Exercise 4 from Chapter 6 called **Show me the money**. In that case, we were surprised to see that the more often a baseball player strikes out, the higher his salary tends to be. This outcome is neither spurious nor phony but is the result of an omitted variable bias. That is, players who strike out a lot actually do make more money then those who do not, but they also hit a lot of home runs. (For instance, Sammy Sosa is, as of this writing, third in the all-time career strike-out list behind Reggie Jackson and Andres Galarraga, and all three are in the top-40 career home run list.) The **strikeouts2** dataset extends the dataset used in the case exercise.

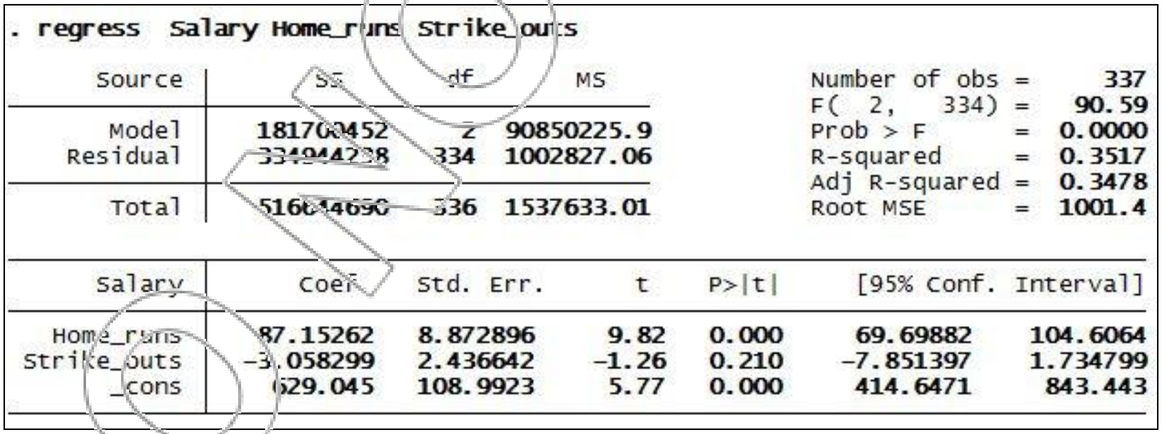
The original regression using just strike outs is shown in Figure 7.9.



```
. regress  Salary Strike_outs

      Source |       SS       df       MS              Number of obs =     337
-------------+------------------------------           F(  1,   335) =   65.92
       Model |  84949301.8     1  84949301.8           Prob > F      =  0.0000
    Residual |   431695388   335  1288642.95           R-squared     =  0.1644
-------------+------------------------------           Adj R-squared =  0.1619
       Total |   516644690   336  1537633.01           Root MSE      =  1135.2

------------------------------------------------------------------------------
      Salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 Strike_outs |    14.8636   1.830671     8.12   0.000     11.26254    18.46465
       _cons |   405.6697   120.8324     3.36   0.001     167.9838    643.3556
------------------------------------------------------------------------------
```

Figure 7.9: Salary vs. strike outs.

Watch what happens when we add the home runs variable to our model. We will see a major change in the coefficient on strike outs (see Figure 7.10).



```
. regress  Salary Home_runs Strike_outs

      Source |       SS       df       MS              Number of obs =     337
-------------+------------------------------           F(  2,   334) =   90.59
       Model |   181700452     2  90850225.9           Prob > F      =  0.0000
    Residual |   334944238   334  1002827.06           R-squared     =  0.3517
-------------+------------------------------           Adj R-squared =  0.3478
       Total |   516644690   336  1537633.01           Root MSE      =  1001.4

------------------------------------------------------------------------------
      Salary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   Home_runs |   87.15262   8.872896     9.82   0.000     69.69882    104.6064
 Strike_outs |  -3.058299   2.436642    -1.26   0.210    -7.851397    1.734799
       _cons |    629.045   108.9923     5.77   0.000     414.6471     843.443
------------------------------------------------------------------------------
```

[FigCap]Figure 7.10: Salary vs. home runs and strike outs.

The coefficient on strike outs has dropped from 14.86 to -3.06. What’s happening here? Which one is the ‘right’ coefficient? Well, they’re both right, but the proper number depends on the question you ask:

i. On average, how much does salary increase for every strike out?

ii. On average, for a player with a certain number of home runs, how much does salary increase for every strike out?

The answer to the first question is about $14,860, and the answer to the second is about -$3,060.

The direct effect of one more strike out is negative; that is, holding home runs constant, the owners would pay players less if they had more strike outs. What's important here is the existence of an indirect effect. Hitting a lot of home runs will make the owners happy enough to pay the player a higher salary, but trying to hit a home run will often lead to a strike out. So, more strike outs is associated with more home runs, which is associated with a greater salary. When the regression only includes the strike out variable, the coefficient has to carry the weight of the direct effect (which is negative) and the indirect effect (which is overwhelmingly positive) on salary. In other words, omitting the home-run variable from the regression biases the coefficient of the strike out variable. We will see this effect whenever related independent variables each have a measurable impact on the dependent variable.

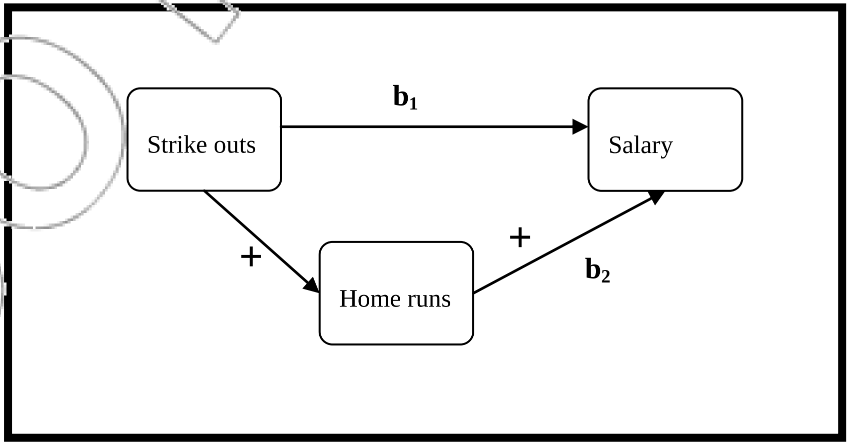


Figure 7.11: Influence diagram.

CALCULATING THE EXTENT OF THE BIAS

Compare two estimated regression equations, where we omit one of the variables in the second one:

$$y = b_0 + b_1 x_1 + b_2 x_2$$

$$y = b_0' + b_1' x_1$$

The **bias** on the coefficient of x_1 is defined to be b_1'-b_1. It turns out that this bias is given by the following:

$$b_1'\text{-}b_1=(\text{effect of } x_1 \text{ on } x_2)*(\text{effect of } x_2 \text{ on } y)$$

The effect of x_2 on y is given by b_2, and the effect of x_1 on x_2 is given by regressing x_2 on x_1:

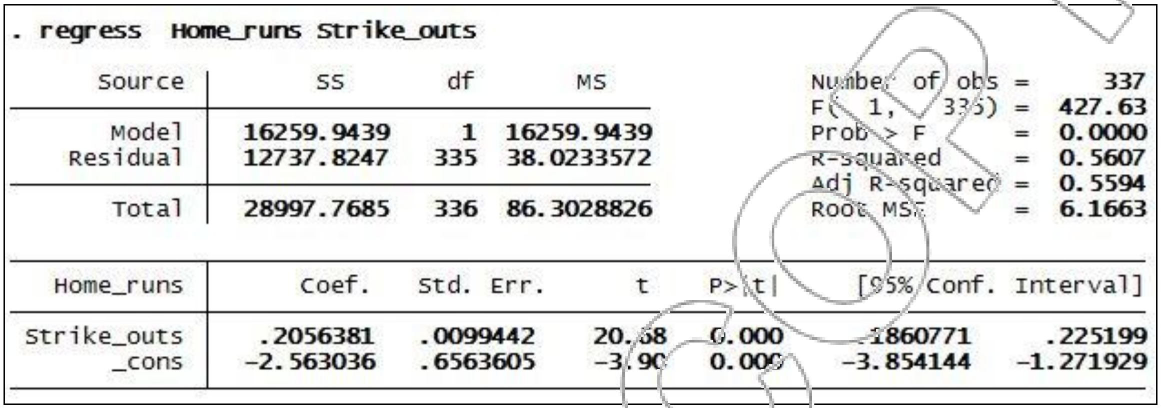
$$x_2 = c_0 + c_1 x_1$$

So, the exact formula is the following:

$$b_1'\text{-}b_1 = c_1 b_2$$

This formula remains valid if we have more than two *x* variables, provided we drop only one of them between the two regressions. The only thing that changes is that now the c_1 is the coefficient on x_1 in the **multiple** regression of the omitted variable on all the non-omitted variables.

As an illustration, we can determine the bias in the strike-outs case by using the previous regressions plus the one in Figure 7.12:



```
. regress  Home_runs Strike_outs

      Source |       SS       df       MS              Number of obs =     337
-------------+------------------------------           F(  1,   335) =  427.63
       Model |  16259.9439     1  16259.9439           Prob > F      =  0.0000
    Residual |  12737.8247   335  38.0233572           R-squared     =  0.5607
-------------+------------------------------           Adj R-squared =  0.5594
       Total |  28997.7685   336  86.3028826           Root MSE      =  6.1663

------------------------------------------------------------------------------
   Home_runs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 Strike_outs |   .2056381   .0099442    20.68   0.000     .1860771     .225199
       _cons |  -2.563036   .6563605    -3.90   0.000    -3.854144   -1.271929
------------------------------------------------------------------------------
```

Figure 7.12: Regression of home runs vs strike outs.

This new regression tells us that every additional strike out yields an average of 0.2056 home runs. The rule for determining the bias on the coefficient of strike outs from omitting home runs tells us to multiply the effect of strike outs on home runs times the effect of home runs on salary holding strike outs fixed (the coefficient on home runs from the regression in Figure 7.10) or 0.2056*87.1526 = 17.92. We can verify that this is the same as the change in the value of the strike-out coefficient when we go from the multiple regression with both variables to the simple regression with just strike outs: 14.86 – (-3.06) = 17.92.

Sign of the Bias

The omitted variable bias in this example was positive (omitting home runs caused an increase in the coefficient on strike outs) but that is not always the case. The influence diagram in Figure 7.11 gives us an idea how to generalize these results. In terms of the figure, the omitted variable bias on the coefficient of the variable in the upper-left box from omitting the variable in the lower box is given by the product of the two lower legs of the triangle.

If the signs of the relationships depicted by both lower legs are positive, then the bias will be positive as we saw in the strike-out example. Similarly, if both relationships have a negative sign, then the bias will be positive. For instance, consider a simple regression of the value of a house in Hawaii on its age. You might be surprised to find a positive coefficient here since newer houses are usually more valuable. However, this result is easily explained by taking into account omitted variable bias and the local real estate market. There is not much land in Hawaii, so the earliest houses were built in the best places like the beachfront. The omitted variable of "Distance to the beach" will have a negative relationship with the house's age and with its value. Though the direct impact of age is negative on the value of a house, the addition of the positive omitted variable bias can create an overall positive coefficient.

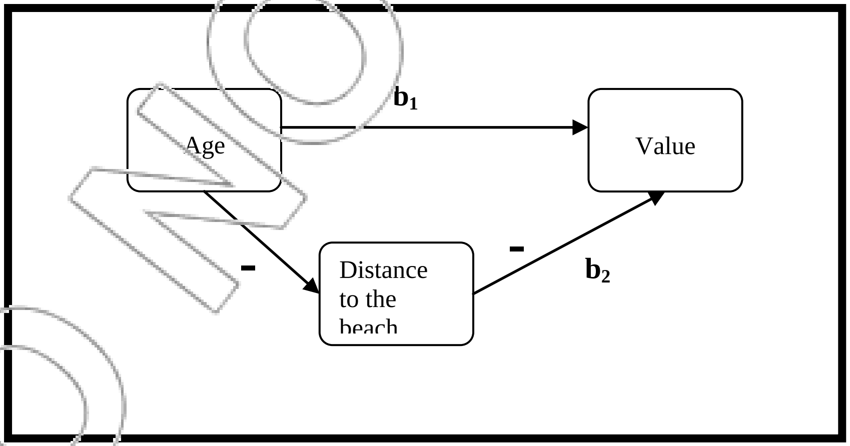


Figure 7.13: Influence diagram of real estate value.

What if one sign is positive, and the other one is negative? For instance, consider a regression of the number of priests in a city on the air quality, which has a negative coefficient. What might cause that result? Does dirty air cause people to become more religious? The omission of the variable **population size** would explain it. A city with dirty air is usually big (a negative relationship), and a city with many people living in it will usually need more clergy (a positive relationship). The product of these two effects creates a negative omitted variable bias on the coefficient of air quality. If this indirect effect is stronger than the direct effect of air quality on the number of priests, which in this case is probably near zero, then the coefficient in the simple regression will be negative.

SUMMARY

It is often useful to conduct hypothesis tests concerning sums and differences or general linear combinations of regression coefficients. The **Linear combinations of coefficients (klincom)** command in Stata can be used to carry out such tests. In the context of the Hot Dog case we used such a test to compare the combined effect of Ball Park's prices to the effect of Oscar Mayer's price on Dubuque's market share.

A multicollinearity problem arises when two or more independent variables are strongly related. In the Hot Dog case, the relationship was between two highly correlated price variables; however, correlation is a limited pair-wise concept, and the problem of multicollinearity is more general than this. Observing a lack of high correlation coefficients does not ensure a freedom from multicollinearity problems; therefore, variance inflation factors need to be used to detect multicollinearity problems accurately.

If a multicollinearity problem exists, then significant variables may have low t-ratios and high p-values. An F-test for joint significance must be conducted on the group of multicollinear variables to properly evaluate their significance if one or more independent variables appear insignificant according to the tests on the individual coefficients and some of these seemingly insignificant variables are involved in the multicollinearity. Nothing can be done to get rid of multicollinearity short of gathering new data where the strong linear relationships among independent variables are lacking.

The estimated regression coefficient on an independent variable may be biased by the omission of another independent variable that is related both to it and to the dependent variable. In many practical situations, you may suspect that such a variable may have been omitted from the analysis, but no data is available to allow you to include it. In such cases, being able to reason about the likely sign of the bias using the influence diagram can be helpful in understanding the potential impact and importance of the omission.

NEW TERMS

Multicollinearity — The term used to describe the presence of linear relationships among the independent variables

Hidden extrapolation — Making a prediction using values of the independent variables that are collectively far from the sample data though each x variable is individually within the sample data's range

Base variables — The variables in your regression you are not testing for joint significance

Added variables — The variables in your regression you wish to test for joint significance

Variance inflation factor (VIF) — A measure of how much the variance of the estimated regression coefficients are enlarged as compared to when the independent variables are not linearly related. Used to detect multicollinearity. A common rule is a VIF above 10 indicates strong multicollinearity involving that variable

Omitted variable bias The effect on a regression coefficient caused by omitting an important correlated variable from the model

NEW FORMULAS

F statistic, $F = ((SSE(x_1,\dots, x_q)/SSE(x_1,\dots, x_q, x_{q+1}, \dots, x_p))-1)*((n-p-1)/(p-q))$

p is the number of variables in the extended model, q is the number of variables in the base model, and p-q is the number of variables being tested.

The omitted variable bias on the coefficient of x_1 from omitting x_2 is

$$b_1'-b_1 = c_1b_2$$

where each of these values come from the following estimated regression equations:

- $y = b_0+b_1x_1+b_2x_2$
- $y = b_0'+b_1'x_1$
- $x_2 = c_0+c_1x_1$

NEW STATA AND EXCEL FUNCTIONS

STATA

User>Core Statistics>Bivariate Statistics>Correlations (correlate)

Equivalently, you may type **db correlate**. This command displays a correlation matrix with the estimated correlations between each pair of variables in the dataset. If any of the variables are non-numeric, Stata will report an error. To avoid this, you can specify the (numeric) variables for which you want Stata to calculate pairwise correlations in the "Variables" field of the correlate dialog box.

Alternatively, you can directly type the command **correlate *varlist***, where ***varlist*** corresponds to the names of the variables for which you want to calculate the correlations. Omitting ***varlist*** will generate a correlation matrix for all variables in the current Stata dataset (provided that all variables are numeric).

User>Core Statistics>Test Hypothesis, using most recent regression>Joint significance (testparm)

Equivalently, you may type **db testparm**. This command opens a dialog box that asks the user to select the **added variables** in the "Test coefficients of these variables" field. Choosing the "Jointly equal to zero" option will tell Stata to conduct an F-test, which we used to determine joint significance of the added variables in a regression with the base and added variables as the independent variables. Note that you need to have run a regression on your extended model before using this command. The Stata output will display the F statistic and p-value of a given F-test.

Alternatively, you can directly type the command **testparm *varlist***, where ***varlist*** contains the name(s) of the added variables. The native menu path in Stata is **Statistics>Postestimation>Tests>Test parameters**.

Ftail(n1, n2, f)

Typing **display Ftail(n1, n2, f)** into the Stata Command box will generate the p-value associated with a given F statistic, **f**. **n1** is the number of variables being tested (p-q), and **n2** is the degrees of freedom for the extended model with all the variables included (n-q-1).

User>Core Statistics>Test Hypotheses, using most recent regression>Linear combinations of coefficients (klincom)

Equivalently, you may type **db klincom**. This command opens a dialog box that asks the user to enter a linear expression of regression coefficients. Do so and then click **OK**, and Stata will conduct a hypothesis test with the null hypothesis "expression=0." Stata reports the test statistic and p-values corresponding to all three types of alternative hypothesis (i.e., "expression" <, ≠, > 0).

Alternatively, you can directly type the command **klincom *expression***.

Note that if you type **lincom *expression***[10] instead, Stata will execute its built-in linear combination of coefficients test rather than the customized **klincom** modification of **lincom**. The only difference is that the **klincom** command will display p-values corresponding to both one- and two-sided tests, while the **lincom** command only displays the p-value for the two-sided test.

[10] The corresponding menu path for this command is **Statistics>Postestimation>Linear combinations of estimates**. Equivalently, you may type **db lincom**.

User>Core Statistics>Model Analysis, using most recent regression>Variance Inflation Factors (vif)

Equivalently, you may type **db vif**. This command reports the variance inflation factors for each independent variable in the most recent regression. We can use this command to detect multicollinearity.

Alternatively, you can directly type the command **vif**.

EXCEL

FDIST

Typing **=FDIST(X, p-q, n-p-1)** into an empty cell returns the p-value associated with a given F statistic, **X**. p is the number of variables in the extended model, q is the number in the base model, and n is the sample size.

CASE EXERCISES

1. Show me even more money.

Running an agency that represents many professional athletes, you are often forced into serious contract negotiations. Having recently fired your assistant, you have decided to evaluate the data collected to support your argument that the player whose contract you are negotiating is currently underpaid. The data in the **strikeouts3**[11] file extends the previous dataset to include much more information.

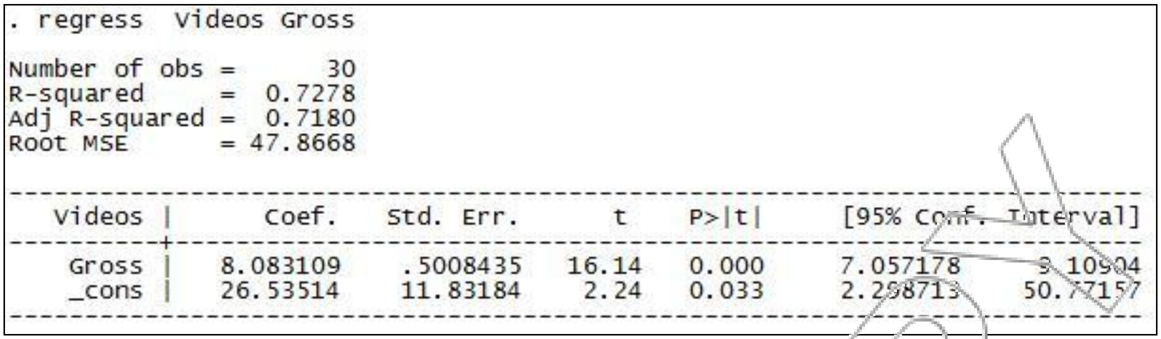
Start by conducting a regression using all of the data provided to predict salary. Do the signs of all of the coefficients make sense?

Next, remove each of the variables that are insignificant based on α = 0.05. Are the variables that you removed jointly significant? How can you tell?

2. Video sales

Your company has the rights to distribute home video of previously released movies. Your goal is to estimate the volume of DVDs you can expect to sell based on box office totals of the original movies. Data are available for 30 movies that indicate the box office gross (**Gross**, in millions of dollars) and the number of DVDs sold (**Videos**, in thousands).

[11] From "Pay for Play: Are Baseball Salaries Based on Performance?" by Mitchell R. Watnik. The Journal of Statistics Education, Volume 6, Number 2 (July 1998)



```
. regress  Videos Gross

Number of obs =      30
R-squared     =  0.7278
Adj R-squared =  0.7180
Root MSE      = 47.8668

------------------------------------------------------------------------------
      Videos |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Gross |   8.083109    .5008435    16.14   0.000     7.057178     9.10904
       _cons |   26.53514    11.83184     2.24   0.033     2.298713    50.77157
------------------------------------------------------------------------------
```

You are planning for the video release of *Matchstick Men* that grossed $36 million. In the Stata Data Editor, you enter 36 for **Gross** in a blank row, execute the command **db confint**, and get the following:

predicted	se_est_mean	se_ind_pred
317.53	13.89164	49.84192

a. Predict the DVD sales for *Matchstick Men*.

b. Construct a 95% prediction interval for the video sales of *Matchstick Men*.

c. Your firm has a truckload of films that were huge flops and grossed $0 each. What would you expect average video sales to be for these films known as the “flops”?

d. Based on your regression, can you prove at a 5% significance level that the average video sales of the flops will be greater than 10,000 copies per film?

3. B-school costs

The **bschools2002**[12] dataset contains information on the top business schools according to a 2002 *Business Week* magazine survey. Use all four numerical variables to develop a model that

[12] Merritt, Jennifer. ***Business Week***, 10/21/2002 Issue 3804, p84

explains the “estimated total costs” of attending the program. Does the coefficient of “base salary: median” make sense? What might be causing this unusual result?

4. Video libraries

A group of independently owned video stores in the south has formed a trade group to help support their survival in the face of competition from dominant national chains. The group of 29 store owners have collected data in the **videostores** file, which contains the average monthly sales, neighborhood population (in thousands), annual advertising expenses, and the number of DVD and VHS films in the libraries (films that have been available for over one year) of each store. A big problem facing these small stores is if they should update their collections of older films by adding DVD versions to their current library. Though they usually buy the new movies in both formats, the lower sales volumes at these small stores make the expense of an older DVD hard to justify. The typical store can break even if the DVD brings in more than 1 dollar per month.

Using all of the variables provided to you by the trade group:

a. Which of the four variables given seem to be significant predictors of sales?

b. On average, how much does one DVD add to the monthly sales of one of the stores?

c. Provide a 95% confidence interval for your estimate.

d. Should the stores upgrade their DVD libraries?

CASE INSERT 2

COLONIAL BROADCASTING

In this case, we will use our regression skills to help run a broadcasting company. The Colonial Broadcasting Company case describes the problem of Barbara Warrington, vice president of Programming at Colonial Broadcasting Company, who has to decide which television movies to broadcast and when to schedule them.

The assignment is to answer all questions in part A of the case except question 7a and all questions in part B except question 12.

In the regression output in the case, some numbers appear within parentheses indicating a negative number. That is, (8) means -8. All questions can be answered without running any additional regressions. However, you are free to do any supplementary analysis using the data contained in the **colonial** file.

In answering question 11, you will think you need to know the standard error of prediction, and you will be right. However, the regression output in the case only provides the standard error of regression. So, for convenience only, you may use the standard error of regression to approximate the standard error of prediction in your answer.

The Colonial Broadcasting Company case (parts A and B)[1] is located in the packet of cases bundled to the back of this text.

[1] Colonial Broadcasting Co., Harvard Business School Case, Product #9-894-011.

CHAPTER 8

THE ADVERTISING CASE:

HETEROSKEDASTICITY AND LOGARITHMS

This chapter presents a brief overview of natural logarithms and demonstrates their use as a technique to model curvature in regression and as a method for removing heteroskedasticity or non-constant variance. Special concerns when making predictions using regressions with logarithmic dependent variables are discussed. An example relating advertising expenditures to sales is explored. The detection and implications of heteroskedasticity are explained. Case Exercise 1 reexamines the hot dog case from Chapter 7 with these new tools and issues in mind.

8.1 A Primer on Logarithms in Regression

Logarithms are used extensively in statistics. In particular, log-linear regression models are a useful alternative to the standard linear form. They work well in various applications where some of the assumptions of the standard linear regression are not satisfied. Moreover, the coefficients of the independent variables in a logarithmic regression are easy to interpret, and the whole equation is easy to use for prediction.

Log forms of regression are used at least as much, if not more often, than the linear form. So, we need to have a good understanding of what they mean and how they work. To achieve this goal, we describe the **main properties** of the logarithm function (the so-called natural logarithm, **ln** in Stata or **LN** in Excel), and show how the logarithmic transformation of variables can be used in regressions. We will talk about different log regression forms (log-log and semi-log), and the **interpretation of coefficients** in these regressions. Then we will highlight the differences between linear and logarithmic regressions as far as **prediction** with these regressions is concerned. Finally, we will introduce an important practical motivation for using log-regressions: logs often “cure” heteroskedasticity. A more in-depth analysis of heteroskedasticity including detection, effects, and fixes is the final subject of the chapter.

PROPERTIES OF THE NATURAL LOGARITHM FUNCTION (ln)

ln(x) is a function that can be evaluated for any positive x value. We show the graph of the function below (Figure 8.1, generated in Excel). To get the graph, we created a column of different x-values (ranging from .0018 to 20), generated their logs (by typing = LN(A2) in cell B2, etc.), and generated the graph with the chart-wizard.

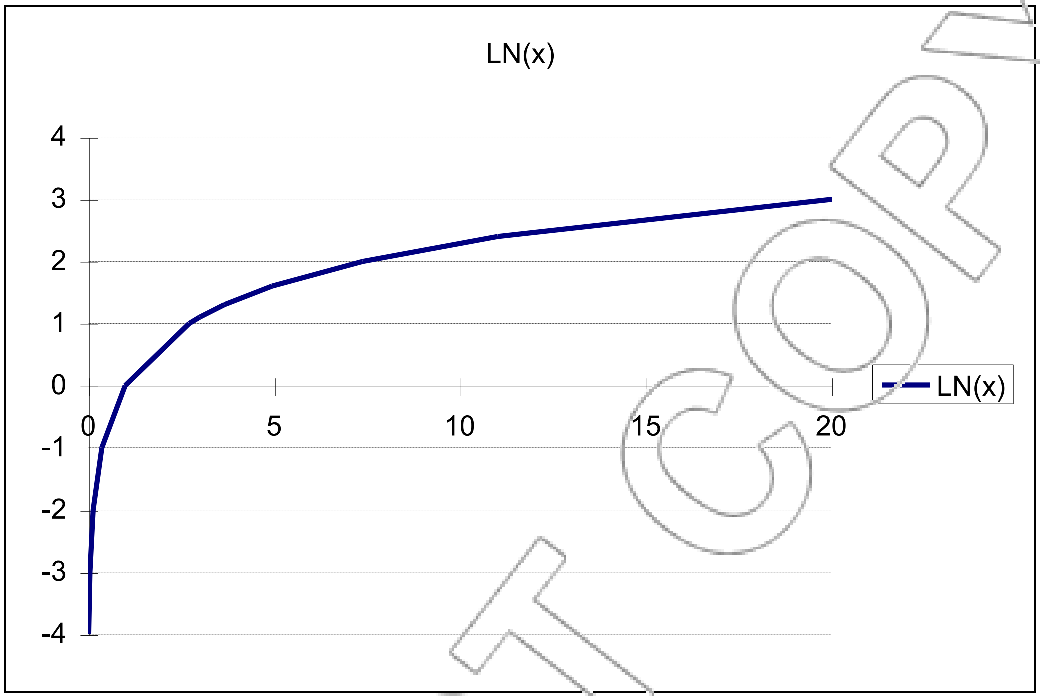


Figure 8.1: Graph of ln(X) vs. X.

The function is increasing everywhere, ln(1) = 0, and, as x approaches 0, ln(x) tends to negative infinity.[1] The logarithm is a concave function in that it increases more slowly as x increases (i.e., the slope decreases as x increases).

An interesting property of the logarithm function is that if you keep multiplying x by a constant (for example, if you double it starting from one, i.e., 1, 2, 4, 8, 16), then the logarithm will increase by a constant increment. In the example, ln(1) = 0, ln(2) = 0.693, ln(4) = 1.386, ln(8) = 2.079, ln(16) = 2.773; the increment is about 0.693 or the log of the multiplier, 0.693 = ln(2).

[1] We will use Stata's **ln(x)** function to do logarithmic calculations. To calculate ln(1), for example, you can type **display ln(1)** in the Stata Command box.

In general, if you increase a number by a fixed proportion (say, by 15 percent, i.e., you multiply it by 1.15), then the logarithm of the number will increase by the logarithm of the multiplier (in the example, by 0.1398 = ln(1.15)).

The logarithm function transforms the **proportional increments** ("doubling" or "increasing by 15%") into **additive increments** ("adding ln(2) = 0.693" or "adding ln(1.15) = 0.1398"). In other words, the logarithm function transforms **growth rates** into (additive) **growth**.

Perhaps more interesting, the following rule of thumb can be used for translating small percentage changes in *x* into absolute changes in ln(*x*).

Every 1% change in *x* corresponds to (approximately) a 0.01 change in ln(*x*).

That is, a *k*% change in *x* corresponds to a $0.01*k$ change in ln(*x*), for any *k* not too large. For example, a 5% increase from 20 results in 21; if you take logs, the difference between ln(21) and ln(20) is equal to ln(21)-ln(20) ≈ 3.04-2.99 = 0.05.

The ln function has many other interesting and related properties. For example, the logarithm of a product, ln(2*3), is equal to the sum of the logarithms of the two factors, ln(2)+ln(3). Also, ln(x^a) = a*ln(*x*), and ln(1/*x*) = -ln(*x*).

Many examples show where logarithms play an important role in the world. In music, the position of a key on the keyboard is a logarithmic function of its pitch's frequency. Our senses, in general, measure things in logs (this is called Fechner's law): "As stimuli are increased by multiplication, sensation increases by addition." Logs come up in financial computations, too.

Suppose that you put $1 in the bank, and a year later receive $1.20 (quite a good deal). What interest rate does this gain correspond to if interest is compounded continuously? The answer is r = ln(1.2) = 0.1823, or 18.23%.

The inverse of the natural logarithm function is the **exponential function**, exp (in Stata). If you have the value for the logarithm of a variable, then, to get the variable's value, you "exponentiate" it. That is, exp(ln(*x*)) = *x* for any positive number *x*.

8.2 Logarithmic Regressions: Forms and Interpretation of the Coefficients

Recall that in the standard linear regression setting we assume the following:

$$(1) \qquad Y = \beta_o+\beta_1X+\text{error term.}$$

Here, we are saying that a one-unit increase in X causes Y to increase by β_1 units, on average. For example, if X is price in dollars and Y is sales of wheat in thousands of tons, β_1 is the number of thousands of tons that average wheat sales change by when the price is increased by one dollar.

We examine two logarithmic regression forms when you have a single independent variable. One is called the **semi-log** specification, and the other the **log-log** specification. In the semi-log specification, you create a new variable, lnY = ln(Y), and regress it against X. In the log-log specification, you regress lnY against lnX = ln(X).

That is, the semi-log regression model can be written as follows:

$$\text{(SL)} \qquad \ln Y = \beta_o+\beta_1 X+\text{error term.}$$

Here, the interpretation of the coefficient β_1 is that when X increases by 1 unit. lnY changes by β_1 units, on average. Because of the interpretation of logs given above, we can say that a one-unit increase in X is associated with approximately a (β_1*100)% change in Y.

For example, let the equation be lnY = 1-0.03*X. Each unit increase in X leads to a 0.03 decrease in lnY, which corresponds to a 3% decrease in Y. (We had to multiply 0.03 by one hundred to get 3, and then we added “percent”.)

The log-log regression model with a single X variable is as follows:

$$\text{(LL)} \qquad \ln Y = \beta_o+\beta_1 \ln X+\text{error term.}$$

Some X variables cannot appear in a log-log regression because they take non-positive values. A good example is when X is a dummy: You cannot take the log of a dummy because it sometimes equals 0.

The interpretation of the coefficient in (LL) is interesting: A 1% increase in X will imply a β_1% change in Y. Why? A 1% increase in X corresponds to (approximately) a 0.01 increase in lnX = ln(X). According to (LL), a 0.01 increase in lnX will lead to a β_1*0.01 change in lnY. This change, in turn, corresponds to (approximately) a β_1% change in Y.

For example, let the equation be lnY = 1-3*lnX. Then a 1% increase in X leads to a 0.01 increase in lnX, which implies a 0.03 decrease in lnY. This corresponds to a 3% decrease in Y. Therefore, a 1% increase in X leads to a 3% decrease in Y. Here, we do not multiply the coefficient by 100 in contrast to what we had to do in the semi-log case.

The natural interpretation of the coefficient of X in the (LL) regression is that it relates a percentage increase in X to a percentage change in Y. Contrast this with the interpretation of the coefficient in a linear regression (L), which relates a unit increase in X to a unit change in Y.

You might recall from microeconomics that the percentage response in a quantity to a percentage change in another quantity is called the **elasticity**. Thus, in equation (LL), we are assuming the elasticity of Y with respect to X is β_1. Examples include where Y is sales, X is price, and β_1 is the price elasticity of demand; where Y is sales, and X is income, and β_1 is the income elasticity of demand; and where Y is cost, and X is output, and β_1 is the output elasticity of cost. For this reason, the form (LL) is widely used and of practical importance.

In a multiple regression, you may have some X variables in logs and some others in their original linear "measurement units:"

$$\boxed{\ln Y = \beta_o+\beta_1 \ln X_1+\beta_2 X_2+\ldots+\text{error term.}}$$

Such a mixed semi-log/log-log regression form may be necessary to accommodate dummy variables in a log-log regression, for example. Remember, you cannot take ln of a dummy or other variable that sometimes has zero or negative values. The interpretation of the coefficients follows just as above. Holding the other included variables fixed, a 1% increase in X_1 will change

Y by β_1%. Holding the other included variables fixed, a unit increase in X_2 will change Y by approximately (β_2*100)%.

8.3 Prediction With Logarithmic Regressions

When you transform some variables using logs and run a logarithmic regression, remember you are no longer working with the original X,Y data. This affects how you do forecasting in two ways.

First, when you are using a log-log model, lnX is the independent variable. This means that if you want to predict when X = 100, you do not enter 100 in Stata's Data Editor. Rather, the X in the regression is ln(X). Thus, you must remember to type in the value **4.6051702** (=ln(100)) in the appropriate cell in the data editor. (Note that when computing logarithmic values it is a good idea to keep more decimal places than usual as they can make a difference when converting back to the original units. For example, exp(4.6051702)=100, but exp(4.605)≈99.98.)

The second important thing is that if you are using lnY as the dependent variable (e.g., in the SL model or the LL model), what the **Prediction, using most recent regression (confint)** command will give you is a prediction, a confidence interval, and a prediction interval for lnY and not for Y. Since this is not typically what you want, you must reconstruct the prediction for Y, the CI, and the PI. To do this, you must exponentiate Stata's prediction output so you are getting Y and not lnY. This must be done for the fitted value (i.e., the prediction) and the ends of the confidence and prediction intervals. In addition to this, it turns out that exponentiating introduces a downward bias in the CI and in the estimate for the **average** value of Y (but not for the estimate of an individual value of Y). Typically, this bias is small in practice, but it can be large and you should get in the habit of correcting for it. The way you do this is to multiply through by exp(s^2/2) after exponentiating, where s is the **standard error of the regression** which is found in the **Root MSE** row in the Stata regression output. The expression exp(s^2/2) is called the correction factor. This bias is absent from the PI or when estimating an individual value of Y. Therefore, you must not use the correction factor in calculating the PI or your estimated individual value of Y.

8.4 Ad Sales: Using Logarithmic Regressions

We will study an interesting application of logs in the Ad Sales case that uses the data in the file **adsales**. This dataset contains observations for the sales of a product (variable **sales**) and advertising expenditures for the same product (variable **expend**). Each are measured in thousands of dollars. Should we anticipate a linear relationship between sales and advertising or do diminishing returns exist? In other words, it is likely that each additional dollar spent advertising may not have as much of an impact as the previous dollar? The scatterplot in Figure 8.2 suggests diminishing returns from advertising.

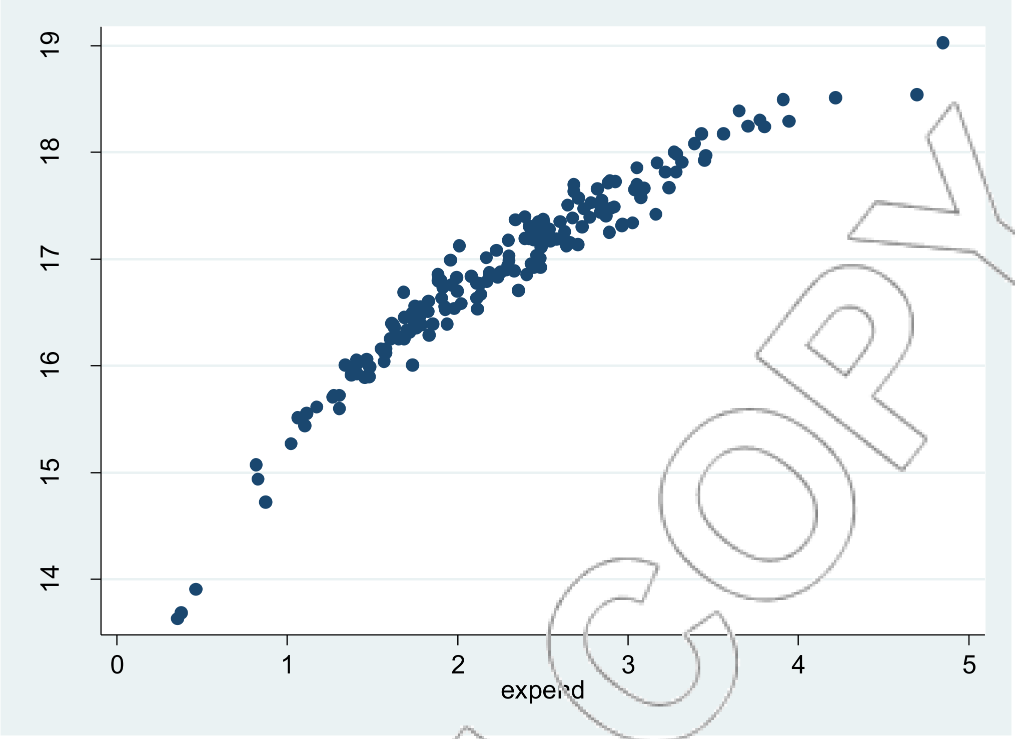


Figure 8.2 Scatterplot of sales vs. expend.

A log-log model might be appropriate. To see this, you may use the residual plot techniques introduced in Chapter 6 to diagnose curvature problems. If you regress sales against expend and then plot the residuals versus the predicted values, you will see distinct curvature in that plot. This means that the linear model is inadequate. We have seen three types of non-linear models thus far: quadratic, semi-log, and log-log. In order to implement them, create three new columns that contain the natural logarithms of variables **expend** and **sales** and the square of **expend** respectively. Label them as **lnexpend**, **lnsales** and **expendsquared**.[2] By trying each of the three non-linear models and examining the plots of residuals versus predicted values, you may verify that the log-log model appears to be the one that best captures the curvature in the relationship (and so removes the curvature from the residual plot).

[2] You can generate these variables in Stata by typing the following commands: 1) **generate lnexpend=ln(expend)**; 2) **generate lnsales=ln(sales)**; and 3) **generate expendsquared=expend^2**.

Run the regression for **lnsales** against **lnexpend**. Suppose we want to obtain the predicted individual and average values of sales and confidence and prediction intervals using a 95% confidence level when spending $2,000 on advertising (expend = 2). First, calculate ln(2) (=**.69314718**), then open Stata's Data Editor and type or paste this value, **.69314718**, in cell lnexpend[174] (i.e., row 174 and column lnexpend). Minimize or close the Data Editor. Then, click **User>Core Statistics>Prediction, using most recent regression (confint)** or type **db confint**. Click **OK**, and Stata will give you, in row 174, the predicted value and confidence and prediction intervals with 95% confidence level for lnsales when lnexpend = ln 2 = 0.69314718. To get the predicted average value for **sales** when expend = 2, you can type **generate pred_avg_sales=exp(predicted)*exp((e(rmse)^2)/2)** in the Stata Command box (i.e., exponentiate the prediction for lnsales and then multiply by the correction factor; **e(rmse)** is where Stata stores s, the value of the standard error of the regression (or Root MSE)). Open the Data Browser, and you will find the predicted average sales when ad spending (expend) = 2 in cell pred_avg_sales[174]. The resulting number should be 16.71939 or $16,719.39. To get the predicted individual value for sales when expend = 2, you can type the command **generate pred_indiv_sales=exp(predicted)**. Open the data browser and look at the cell pred_indiv_sales[174]. The resulting number should be 16.71864 or $16,718.64.

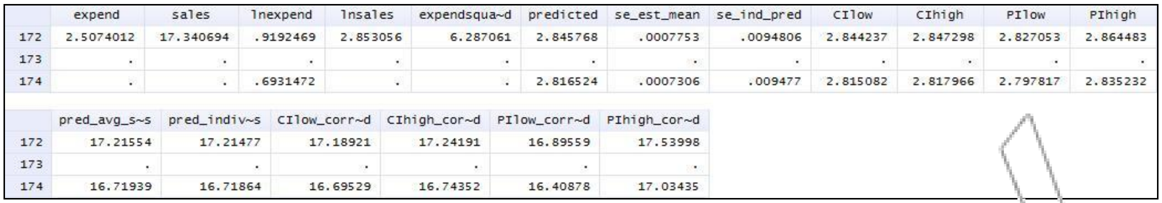
To obtain the corrected confidence interval, you can type the following commands: 1) **generate CIlow_corrected=exp(CIlow)*exp((e(rmse)^2)/2)** and 2) **generate CIlhigh_corrected=exp(CIhigh)*exp((e(rmse)^2)/2)**. You will obtain the 95% confidence interval for average sales when expend = 2 as (16.69529, 16.74352) or ($16,695.29, $16,743.52) (in cells CIlow_corrected[174] and CIlhigh_corrected[174], respectively).

To obtain the correct prediction interval, you can type the following commands: 1) **generate PIlow_corrected=exp(PIlow)** and 2) **generate PIhigh_corrected=exp(PIhigh)**. You will obtain the 95% prediction interval for sales when expend = 2 as (16.40878, 17.03435) or ($16,408.78, $17,034.35) (in cells PIlow_corrected[174] and PIhigh_corrected[174], respectively). Notice that we did not use the correction factor in calculating the prediction interval.

When you are done, rows 172 to 174 of your data sheet will look like Figure 8.3. If you want to calculate the confidence and prediction intervals for any other confidence level, open the **Prediction, using most recent regression (confint)** dialog box again and type the confidence level that you want in the "Confidence level in %" field. Values in the **pred_avg_sales** and **pred_indiv_sales** columns will remain unchanged. However, to get the correct CI and PI, you will have to regenerate the variables **CIlow_corrected**, **CIhigh_corrected**, **PIlow_corrected**, and **PI_high_corrected**. To do so, for example, you can type the command **replace CIlow_corrected=exp(CIlow)*exp(e(rmse)^2/2)** after you have rerun the confint prediction command with the newly specified confidence level.[3]

You may also type in other values of lnexpend in the data editor. To get the appropriately transformed prediction, CI, and PI in this case, use the **Prediction, using most recent regression (confint)** command again after you have entered new values of lnexpend. Then, regenerate the variables **pred_avg_sales**, **pred_indiv_sales**, **corrected_CIlow**, **corrected_CIhigh**, **corrected_PIlow**, and **correctedPI_high** by typing **replace…** instead of **generate…** in the respective commands that you used to generate these variables originally. For example, to regenerate the variable **pred_indiv_sales**, you can type **replace pred_indiv_sales=exp(predicted)**.

[3] Similarly, you can type the following commands to regenerate **CIhigh_corrected**, **PIlow_corrected**, and **PI_high_corrected**: 1)**replace CIhigh_corrected=exp(CIhigh)*exp(e(rmse)^2/2)**, 2)**replace PIlow_corrected=exp(PIlow)**, and 3) **replace PI_high_corrected=exp(PIhigh)**.



	expend	sales	lnexpend	lnsales	expendsqua~d	predicted	se_est_mean	se_ind_pred	CIlow	CIhigh	PIlow	PIhigh
172	2.5074012	17.340694	.9192469	2.853056	6.287061	2.845768	.0007753	.0094806	2.844237	2.847298	2.827053	2.864483
173	.	.	.	.	.	.	.	.	.	.	.	.
174	.	.	.6931472	.	.	2.816524	.0007306	.009477	2.815082	2.817966	2.797817	2.835232

	pred_avg_s~s	pred_indiv~s	CIlow_corr~d	CIhigh_cor~d	PIlow_corr~d	PIhigh_cor~d
172	17.21554	17.21477	17.18921	17.24191	16.89559	17.53998
173	.	.	.	.	.	.
174	16.71939	16.71864	16.69529	16.74352	16.40878	17.03435

Figure 8.3: Prediction for sales with expend = 2.

8.5 Introduction to Heteroskedasticity

Finally, we should talk about an important reason why log-regressions are useful that is separate from their use in modeling curvature as in the Ad Sales application. A key reason for using logarithmic regressions is simple: by taking the logarithm of Y and regressing it on the X variables, which may be in linear units or in logs, we are often able to reduce heteroskedasticity (non-constant error variance).

Why? Suppose the relationship between Y and X is such that average Y = β_o+β_1X; however, an individual observation's deviation from the average (the "error term") is proportional to Y. For simplicity, imagine the individual Y_i (at any given level of X_i) is within ±2% of the average Y at X_i. This structure is heteroskedastic. The standard error of the regression is not constant but instead increases proportionally with Y.

Now see what happens when we create lnY = ln(Y), and regress this variable against X or lnX. That is, we can use a semi-log or a log-log specification. A ±2% error in Y will become a ±0.02 error in lnY. The new error term is not increasing with Y anymore; the error has become homoskedastic.

This example exhibits what often happens in practice: a heteroskedastic regression, where the error term is approximately proportional to Y, can be transformed into a homoskedastic regression by transforming the dependent variable into logarithms. (We need not transform X for this purpose.) To further illustrate the effect of logs on a regression, we show three versions of the same data using three different scatterplots (with a fitted line): the first plot shows Y against X (the relation is visibly heteroskedastic); the second one is lnY against X, and the third one is lnY against lnX.

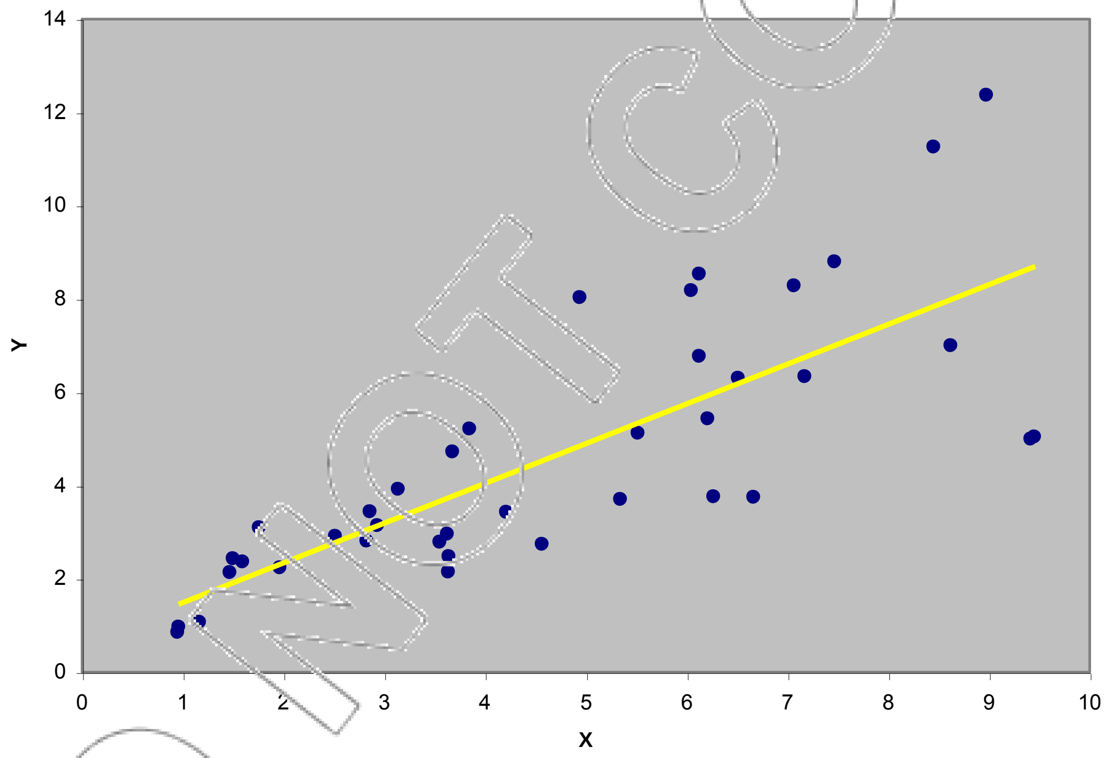


Figure 8.4: Y vs. X.

In Figure 8.4, Y against X appears to be linear but heteroskedastic. The errors are getting larger as Y increases. Note the “cone-shaped” cloud of data points.

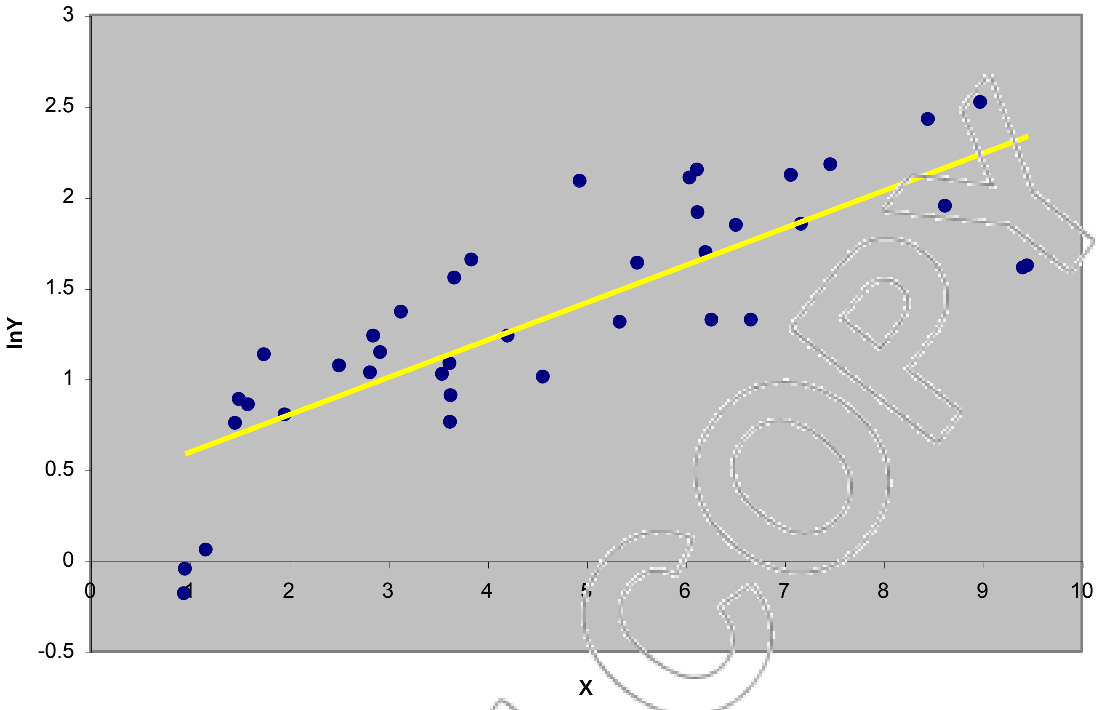


Figure 8.5 lnY vs. X.

In the second scatterplot (see Figure 8.5), the variance of the error term seems to be roughly stable, and so the heteroskedasticity is gone, but there is noticeable curvature. This is not surprising: If Y is indeed linear in X, then lnY will be non-linear in X (the logarithmic transformation of Y introduces curvature).

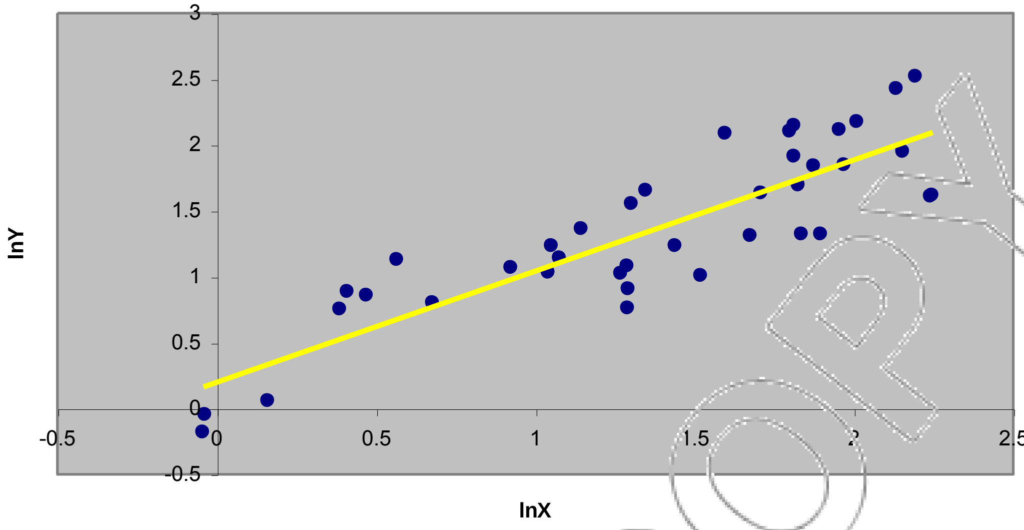


Figure 8.6: lnY vs. lnX.

The third plot (see Figure 8.6) shows that the heteroskedasticity is gone, and the curvature introduced by the semi-log model is gone, too, in this log-log model. This is beautiful.

The situation illustrated in these three scatterplots is not always the case when we find heteroskedasticity in the linear specification, but it is fairly typical. A log-transformation of the dependent variable often resolves heteroskedasticity, and at least one of the possible log-regressions (LL or SL) often works in terms of linearity. In the scatterplots, the SL specification exhibited curvature, and the LL specification did not. However, there are many examples in which the reverse is true and LL exhibits curvature. In other examples, both models effectively capture the curvature in the data.

In Section 8.7, we will explore heteroskedasticity, its detection, effects, and possible fixes, in more depth.

Summary for logarithms in regression

As we stated earlier, a *k* % change in (any variable) X corresponds to approximately a *k**0.01 change in its logarithm, lnX, for any *k* not too large. This property of the logarithm is useful in guiding the interpretation of coefficients in a log regression. It also allows us to eliminate heteroskedasticity when the error term is approximately proportional to the dependent variable: we take the logarithm of Y and regress it against X or lnX.

The two forms of logarithmic regression we examined are semi-log (lnY against X) and log-log (lnY against lnX). In the semi-log case, we multiply the coefficient on X by 100 to get the percentage change in Y as a result of a unit increase in X, holding all other included variables constant. In the log-log case, the coefficient is the elasticity of Y with respect to X (the percentage change in Y for a 1% increase in X), holding all other included variables constant. If a variable takes on zero or negative values, then we cannot take its logarithm.

When using a logarithmic regression for prediction, we must exponentiate the fitted lnY to get the prediction for an individual Y (and the same applies to the prediction interval). For the estimated mean of Y (predicting an average Y), we have to exponentiate the predicted lnY and multiply it by the **correction factor**, exp(s^2/2), where *s* is the standard error of regression. The same applies to the calculation of a confidence interval for average Y: exponentiate the two limits and multiply them by the correction factor.

8.6 An Optional Mathematical Digression

Two comments for the more mathematically inclined:

1. Another look at the change in lnX for a change in X using derivatives:

Those of you who took calculus may remember that the derivative of the natural logarithm function is d$ln(x)$/dx = 1/x. In other words, the slope of the (natural) logarithm curve is 1/x at x.

What does this mean? The slope tells us that for a small Δx increase in x, the function $ln(x)$ will increase by $\Delta ln(x) \approx (1/x)*\Delta x = \Delta x/x$. In other words, the absolute change in the logarithm of x is approximately the percentage change in x (the approximation works best for small Δx changes).

2. Another look at the logarithmic form.

The inverse of the ln function, the exp function, can be written as exp(x) = e^x, where e = 2.7183…. This famous constant is known as the basis of the natural logarithm, or Euler’s number. Exponentiating both sides of the equations SL and LL, we get the following equations:

$$\text{(SL')}\quad Y = e^{\beta_0+\beta_1 X}$$

$$\text{(LL')}\quad Y = e^{\beta_0} X^{\beta_1}$$

We have omitted the error term from these expressions for simplicity. (The error term would be multiplicative: There would be a factor e^{error} multiplying the right hand sides of SL' and LL'.) You can see the relationship between Y and X is non-linear in either model. Moreover, these regression forms make precise the meaning of the coefficient on the X variable in the SL and LL specifications.

Consider the SL specification and increase X by one. For concreteness, suppose X increases from 0 to 1. As a result, Y will change by a factor of e^{β_1}; in the example, it goes from e^{β_0} to $e^{\beta_0 + \beta_1}$. So, the percentage change in Y is $100*(e^{\beta_0 + \beta_1}-e^{\beta_0})/e^{\beta_0} = ((e^{\beta_1}-1)*100)$, which, for small β_1, approximately equals a (β_1*100)% change. (You can check: $e^{\beta_1} \approx 1+\beta_1$ for small β_1.)

The interpretation given earlier for the coefficient in the LL specification is exact: In the LL regression, β_1 is the elasticity of Y with respect to X.

8.7 Heteroskedasticity: Detecting, Effect on Results, Possible Fixes

Four basic assumptions are needed for the regression model to give us the best estimates: linearity, constant error variance, independent errors, and normal errors. The second of these assumptions is the assumption that the error term has the same variance for all observations:

Regression Assumption (homoskedasticity): $\text{Var}(\varepsilon_i) = \sigma^2$ for all i.

The purpose of this section is to show you two methods for checking whether this assumption is satisfied in any particular application, to tell you what goes wrong when this assumption is violated, and to suggest possible ways of fixing violations.

Detecting a Violation: There are at least two useful ways to detect variations (heteroskedasticity) in the error variances. The first technique is to run the regression and examine a plot of the residuals versus the predicted values. What should we expect to see on this graph? If our regression assumptions are satisfied and the error term for each observation has the same variance, then the predicted value we look at should not affect the vertical spread (a way of visualizing variance) of the residuals. Thus, the vertical spread in points on the graph should remain approximately the same all the way across.

In contrast, if the graph of residuals versus predicted values is cone-shaped or otherwise varies in a systematic way in the vertical spread of the residuals, this indicates a violation of our constant variance assumption. Below is an example of a plot of residuals versus predicted values that displays a spread in the residuals that increases as the predicted value increases (see Figure 8.7). This pattern is often seen when analyzing data on income levels, prices, or asset values.

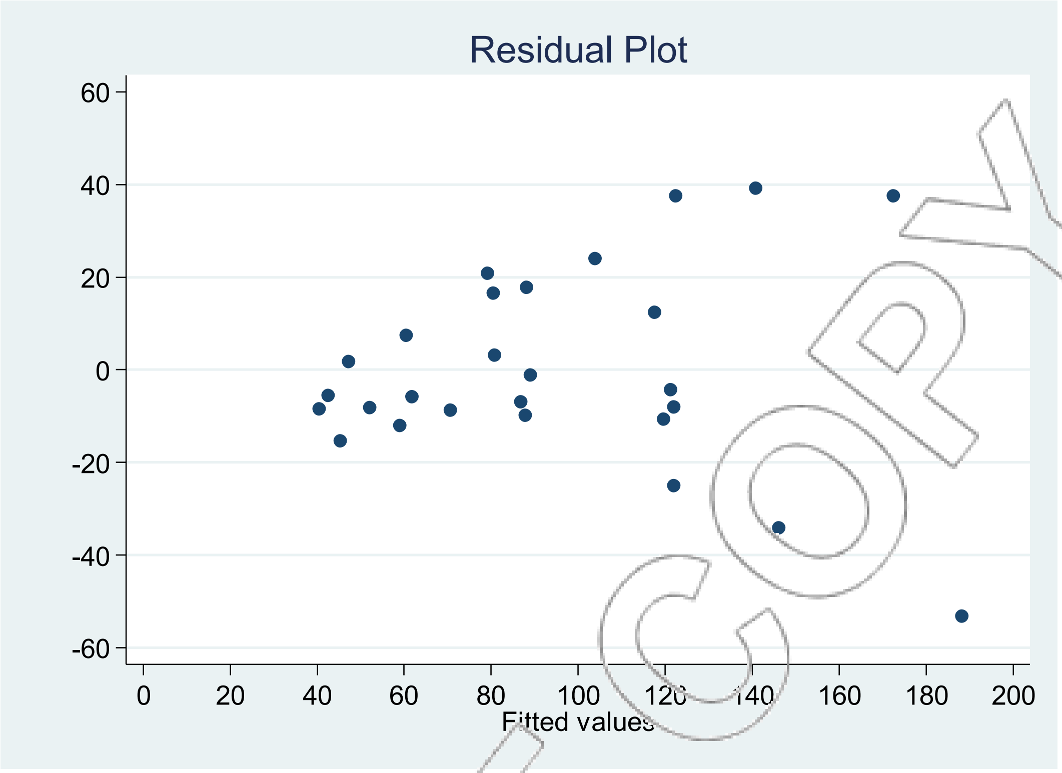


Figure 8.7: Residual plot with heteroskedasticity.

Though examining the graph of the residuals versus the predicted values can be useful, it can be difficult to see if clear evidence of non-constant variance exists through graphical methods. To avoid some of these problems, more quantitative techniques are available for detecting non-constant error variance. One of the easiest to implement is a version of the Breusch-Pagan Test (named after its inventors). This consists of a hypothesis test where the null hypothesis is that Var(ε_i) is constant (homoskedastic) and the alternative hypothesis is that Var(ε_i) varies with the predicted values (y-hat's) in a linear way. Stata performs this test and produces the p-value for us. To do this, first run a regression. Then, click **User>Core Statistics>Model Analysis, using most recent regression>Breusch-Pagan heteroskedasticity test (hettest)** or type **db hettest**.[4] The p-value for this test will be the value corresponding to **Prob > chi2**. A low p-value suggests

[4] The corresponding typed command is **hettest**.

rejecting the null and a high p-value suggests not rejecting it. Therefore, a small p-value (usually below .1 or .05) is strong evidence of heteroskedasticity.

Effect of a Violation: Suppose we discover the constant error variance assumption has been violated. What are the consequences? The estimates of our regression coefficients remain unbiased, but the calculated standard deviations and interval estimates are no longer good estimates. Thus, we will no longer have a good measure of the accuracy of our estimates and predictions. Without a good measure of accuracy, we will not know how much to rely on our estimates in making decisions, we will not be able to judge if we need to gather more data, and we will not be able to conduct correct hypothesis tests to measure the strength of our findings. What can be done to remedy this?

Possible Fixes: Transforming the variables using logarithms (in semi-log or log-log form) if variance increases in the fitted values often helps. To see if it does, transform the variables, run the transformed regression, examine the residuals versus predicted values, and run the Breusch-Pagan Test again. Transformation using logarithms has worked if a serious indication of non-constant variance no longer occurs. More advanced techniques than we will cover, such as Weighted Least Squares, may help in situations where data transformations do not. (An advanced reference describing this procedure is Chapter 10.1 in *Applied Linear Regression Models, 4th ed.* by Neter, Kutner, Nachtsheim, and Wasserman.) Other advanced methods include procedures for calculating standard errors (and the associated interval estimates and hypothesis tests) that are robust to heteroskedasticity. In Stata, robust standard errors can be calculated instead of the usual standard errors when running a regression by using the options on the **SE/Robust** tab of the **regress** dialog box.

NEW TERMS

Elasticity
The percentage response in one quantity to a percentage change in another

Semi-Log (SL) Model
A regression model in which the dependent variable is transformed using the natural logarithm function ln and the independent variable(s) are not

Log-Log (LL) Model
A regression model in which the dependent and independent variables are transformed using the natural logarithm function ln

Correction factor
The value, $\exp(s^2/2)$, used to correct for a downward bias in regression estimates of average Y (including confidence intervals for average Y) induced by using ln(Y) as the dependent variable

Heteroskedasticity
Non-constant variance. This violates the assumptions of the regression model

Breusch-Pagan Test
A statistical test used to detect heteroskedasticity in a regression. Low p-values of this test indicate heteroskedasticity is present

NEW FORMULAS

Properties of Logarithms

$$\ln(x*y) = \ln(x) + \ln(y)$$

$$\ln(x^a) = a*\ln(x)$$

$$\ln(1/x) = -\ln(x)$$

Correction Factor = exp(s^2/2) where s is the standard error of regression

NEW STATA AND EXCEL FUNCTIONS

STATA

User>Core Statistics>Model Analysis, using most recent regression>Breusch-Pagan heteroskedasticity test (hettest)

Equivalently, you may type **db hettest**. This command computes the Breusch-Pagan Test p-value that may be used to detect heteroskedasticity in the most recent regression model.

Alternatively, you can bypass the dialog box by directly typing the command **hettest**.

ln

Typing **display ln(X)** into the Stata Command box returns the natural logarithm of the number X as long as X is positive.

exp

Typing **display exp(X)** into the Stata Command box exponentiates the number X and displays the result. Exponentiating is the mathematical opposite or inverse of the natural log function. exp(X) = e^X where e is a special mathematical constant having the property that ln(e) = 1.

EXCEL

LN

Typing **=LN(X)** into an empty cell returns the natural logarithm of the number X as long as X is positive. Typing **=LN(A2)** into an empty cell returns the natural logarithm of the number contained in cell A2.

EXP

Typing =EXP(X) into an empty cell exponentiates the number X. Typing =EXP(A2) into an empty cell exponentiates the number in cell A2.

CASE EXERCISES

1. Hot Dog revisited

We return to the market for supermarket hot dog dominance. Previously, we investigated some weekly scanner data from grocery stores on Dubuque's market share and price and the prices of two competitors: Oscar Mayer and Ball Park. We used these data to investigate how Dubuque's market share depends on these prices. We saw how multicollinearity affected our findings. Now we are prepared to be on the lookout for heteroskedasticity (non-constant variance).

Keeping this in mind, we would like to use the data in the **hotdog** file to help Dubuque answer some further questions:

a. If Dubuque prices at $1.65, Oscar Mayer prices at $1.75, and Ball Park prices at $1.50 for regular and $1.60 for beef franks, what is Dubuque's expected market share?
b. If, at these prices, we observe Dubuque with a 1.5% market share, would this give us reason to think the market had changed? What if Dubuque had a 4% market share?
c. At these prices, should Dubuque raise or lower its price? You may assume the size of the hot dog market is roughly fixed at 12,000 hot dog packages per week and Dubuque has a cost per unit produced of $1.30/ package. Does it matter how competitors would react to this change?

2. Office networks

A tech support company, Net Geeks, is bidding on a major contract to provide networking support to a firm that owns a chain of tax preparation consultancies across the country. In preparing its bid, Net Geeks has acquired the data contained in the **email** file, which lists the average number of daily internal emails and the number of computers for a sample of 24 of the tax firm’s offices. One key question in determining their bid involves the expected number of internal emails in an office with 20 computers; specifically, Net Geeks needs to know the probability that any particular office with 20 computers will have an average daily internal email volume below 200. Your job is to develop the best regression model to answer this question and use it to respond to the following questions:

a. What is the best estimate for the average daily internal email volume for an office with 20 computers?
b. Provide a 95% prediction interval for this estimate.
c. Estimate the probability that the average daily internal emails at a particular office with 20 computers will be under 200.
d. What can you say about the validity of the estimate in part c?
e. Estimate the probability that the mean number of average daily internal emails for offices with 20 computers will be under 200.

3. Super staffing

Your company is currently building a new factory, which will employ 1,200 workers. You are confronted with the question of how many supervisors (supers) to hire for this plant to supervise the workers and to ensure a well-organized production process. You have employee data (**Factory**) from your other factories, namely the number of supervisors and workers at these facilities.

Construct a linear regression of supers vs. workers.

a. Mathematically, what does the coefficient on workers tell us about our staffing needs?
b. Estimate the number of supers needed for our new factory and provide a 95% prediction interval for your estimate.
c. Are there any problems in using this regression to answer part b?

Construct a regression of lnsupers vs. workers.

d. Mathematically, what does the coefficient on workers tell us about our staffing needs?
e. Estimate the number of supers needed for our new factory and provide a 95% prediction interval for your estimate.
f. Are there any problems in using this regression to answer part e?

Construct a regression of lnsupers vs. lnworkers.

g. Mathematically, what does the coefficient on workers tell us about our staffing needs?
h. Estimate the number of supers needed for our new factory and provide a 95% prediction interval for your estimate.
i. Are there any problems in using this regression to answer part h?
j. Which of the three regressions above is the best one to use for this scenario? Explain.

4. Big movies revisited

Movie studios spend a great deal of energy determining which films will be successful. A major hit or flop can have a measurable effect the bottom line of companies as big and diverse as Disney and Time Warner. The **bigmovies**[5] file contains information on the major films of 1998 that we briefly examined in Chapter Two. Use this information to develop a model that predicts total domestic gross for a film based on the following independent variables:

Best Actor The number of actors or actresses in the movie who were listed in Entertainment Weekly's list of the 25 Best Actors and the 25 Best Actresses of the 1990s

Top Dollar Actors The number of actors or actresses appearing in the movie who were among the top 20 actors and top 20 actresses in average box office gross per movie in their careers at the beginning of 1998 and had appeared in at least 10 movies at that time

Summer A dummy variable indicating if the movie was released during the summer season (May 31 to Sept 5 inclusive) (= 1 if released during summer, = 0 otherwise)

Holiday A dummy variable indicating if the movie was released on a holiday weekend (President's Day, Memorial Day, Independence Day, Labor Day, Thanksgiving, Christmas Day, New Year's Day) (= 1 if released on a holiday weekend, = 0 otherwise)

Christmas A dummy variable indicating if the movie was released during the Christmas season (December 18th – 31st) (= 1 if released during the Christmas season, = 0 otherwise)

[5] Source: The Internet Movie Database, http://www.imdb.com.

Opening Screens	The number of movie screens the film was shown on during the film's first weekend of general release

a. Construct a linear model using total domestic gross as the dependent variable.

b. Use the Model Analysis function in Stata to check the assumptions of the regression model.

Now add a new column of data titled lntotalgross that contains the natural logarithm of the total domestic gross.

c. Construct a semi-log model using lntotalgross as the dependent variable.

d. Use the Model Analysis function in Stata to check the assumptions of the regression model.

e. Choose the better model from the two above and use it to predict the total gross of a movie opening on 2,600 screens with no big or top-dollar actors on a non-holiday weekend during the summer. Provide a 90% prediction interval for your estimate.

CHAPTER 9

SODA SALES AND HARMON FOODS: DEALING WITH TIME AND SEASONALITY

We will use two forecasting cases in this chapter to demonstrate different techniques for modeling seasonality. Quarterly data in the soda case display a seasonal pattern as summer sales outpace winter sales. We use multiple dummy variables to additively model and measure the seasonal impact on sales. Next, the longer Harmon Foods HBS case uses a multiplicative seasonality model to forecast sales of its breakfast cereal. The case introduces the technique of lagging independent variables to model lingering effects. Finally, we will explore different techniques for analyzing time series data including the Cochrane-Orcutt method and the Auto Regressive Integrated Moving Average (ARIMA) model.

9.1 Soda Sales

INTRODUCTION

You have been asked by Cesca, Inc., to forecast future sales of Dada Soda. The data are in the **soda** file. It consists of quarterly Dada Soda sales figures for the last four years (see Figure 9.1).[1] Quarter 1 is the beginning of a year and is, therefore, a winter quarter.

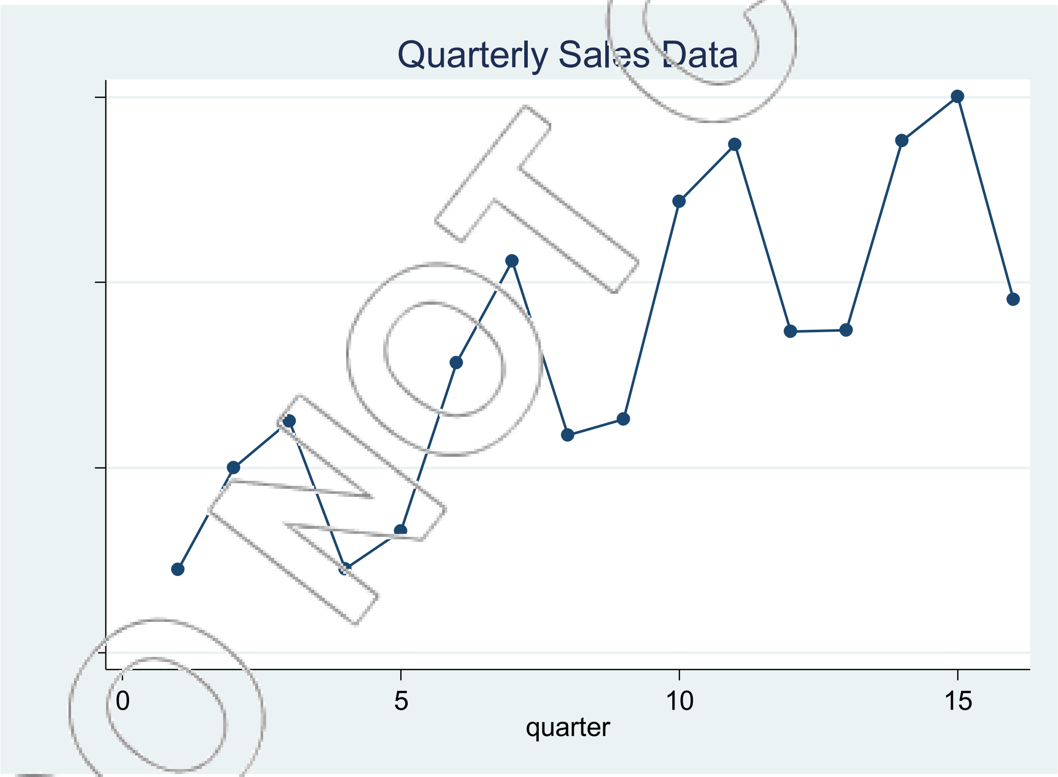


Figure 9.1: Quarterly sales for Dada Soda.

[1] To generate this graph, click **User>Core Statistics>Bivariate Statistics>Bivariate Plots (twoway)** or type **db twoway** to open the **twoway** dialog box. Click **Create…** to specify your dependent and independent variables, and select **Connected** in the "Basic plots: (select type)" field. The direct command for this example is **twoway connected sales quarter**.

Two things are apparent from the graph: Sales are growing over time, and a strong seasonal factor exists. Suppose we ignore the seasonality and regress sales against the quarter variable, i.e., draw a best-fit line through the graph (see Figure 9.2).

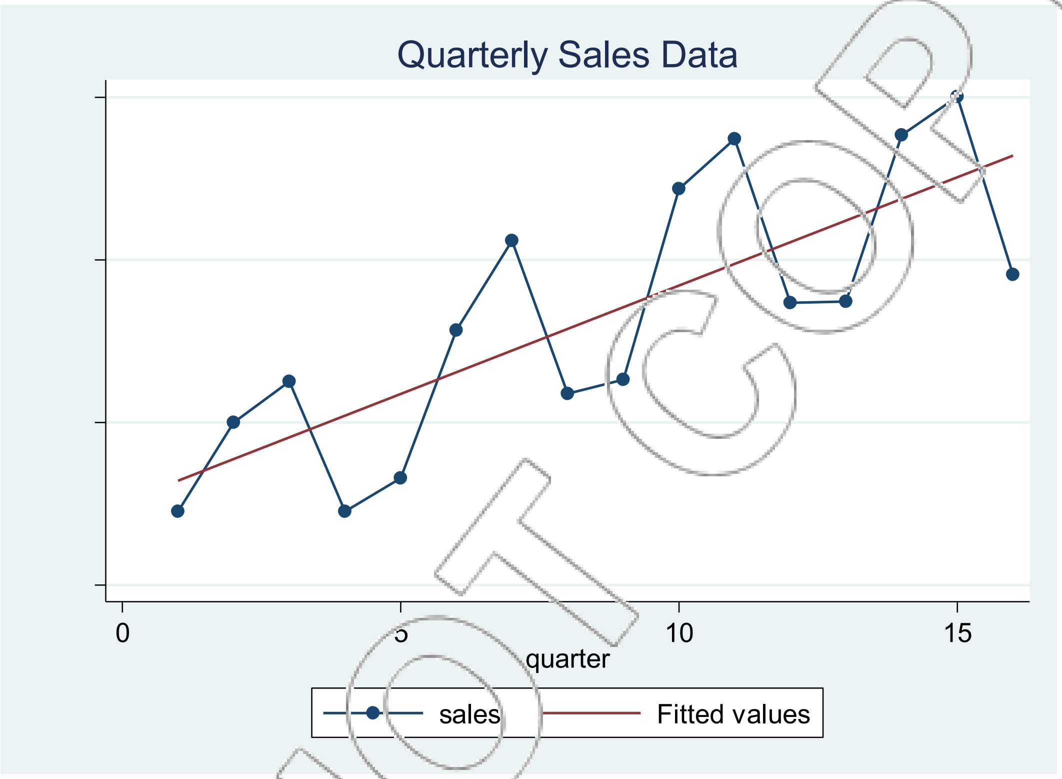


Figure 9.2: Quarterly sales for Dada Soda with regression line.

This procedure enables us to estimate future sales growth by extrapolation since the coefficient on the X variable (quarter) represents average sales growth per quarter in the last four years. The estimated coefficient on quarter is 6668.61 so predicted sales growth is 4×6668.61 = 26,674.44 units per year. However, there are two problems: One is practical and the other is technical, but still important. The practical problem is that it would be useful to have an estimate of the seasonal effects as well as of the average sales growth. At the moment, the regression is predicting sales will increase every quarter, and that is not the case: From year to year, sales are going up, but, for example, they consistently decrease from summer to fall in a given year. Solving this practical problem takes care of the technical one, so we will go through the solution first and explain what the technical problem was at the end.

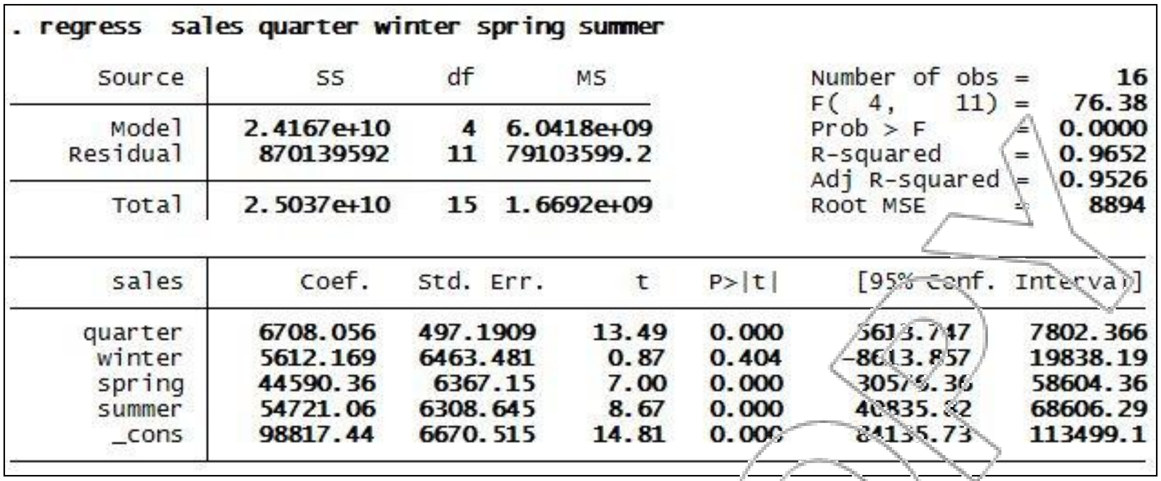
INTRODUCING SEASONAL DUMMIES

We need to introduce dummy variables to take account of the effect of the different seasons. To cope with the four seasons, we will need three dummy variables because one season will function as a benchmark to which we will compare the other three. We choose to include one for each of winter, spring, and summer, so our extended dataset looks like Figure 9.3.

quarter	sales	winter	spring	summer
1	122520	1	0	0
2	149931	0	1	0
3	162481	0	0	1
4	122630	0	0	0
5	132818	1	0	0
6	178325	0	1	0
Etc…				

Figure 9.3: Dada Soda data.

Now we will run the new regression and discuss what the coefficients tell us (see Figure 9.4).



```
. regress  sales quarter winter spring summer

      Source |       SS       df       MS              Number of obs =      16
-------------+------------------------------           F(  4,    11) =   76.38
       Model |  2.4167e+10     4  6.0418e+09           Prob > F      =  0.0000
    Residual |   870139592    11  79103599.2           R-squared     =  0.9652
-------------+------------------------------           Adj R-squared =  0.9526
       Total |  2.5037e+10    15  1.6692e+09           Root MSE      =    8894

------------------------------------------------------------------------------
       sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     quarter |   6708.056   497.1909    13.49   0.000     5613.747    7802.366
      winter |   5612.169   6463.481     0.87   0.404    -8613.857    19838.19
      spring |   44590.36    6367.15     7.00   0.000     30576.36    58604.36
      summer |   54721.06   6308.645     8.67   0.000     40835.82    68606.29
       _cons |   98817.44   6670.515    14.81   0.000     84135.73    113499.1
------------------------------------------------------------------------------
```

Figure 9.4: Regression of Dada Soda with seasonal dummy variables.

INTERPRETING THE DUMMY COEFFICIENTS

As always when dealing with dummy variables, we work out what the equation means by going through the different qualitative states, i.e., the different seasons, one at a time. For example, in fall, we know that all three dummies equal zero and the regression equation from Figure 9.4 reads as follows:

$$\text{sales} = 98817 + 6708 \text{ quarter}$$

If we compare fall one year to fall the next year, this equation will apply to both but the quarter variable has increased by four, so it predicts that fall quarter sales should increase by 4×6708 = 26,832 units per year. If we look at summer instead, we know that the summer dummy equals 1 and both the others equal 0, so the regression equation from Figure 9.4 reads as follows:

$$\begin{aligned}\text{sales} &= 98817+6708 \text{ quarter}+54{,}721(1)\\ &= 153{,}538+6708 \text{ quarter}\end{aligned}$$

Again, this tells us that if we compare yearly summer quarter sales, we should expect an increase in sales of 4×6708 = 26,832 units per year. The same will apply if we look at spring and winter, so the first conclusion is that once we have controlled for seasonality, the predicted annual increase in sales is 26,832 units. In addition, we can predict how sales will change quarterly. Suppose we move from summer to fall. The quarter variable increases in value by 1, giving an extra 6,708 units, but the summer dummy changes from 1 to 0 so we lose 54,721 units, a net decrease of 48,013 units. Things are a little more difficult when (for example) we move from winter to spring. The quarter variable goes up by 1 as before, the winter dummy goes from 1 to 0, and the spring dummy goes from 0 to 1, so the net effect is +6,708-5,612+44,590 = 45,686 units.

We have, therefore, managed to resolve the changes into a quarterly seasonal effect, and a yearly growth trend. The R-squared has increased from around 60% to over 95%, which suggests this multiple regression fits the data better than the regression without the seasonal terms did. However, R-squared is not the appropriate way to compare the fit of two regressions that have the same dependent (Y) variable but different numbers of independent (X) variables. A better measure for such a comparison is something called the **adjusted R-squared**. It is reported on the Stata output directly below the R-squared. The purpose of the adjusted R-squared is to adjust the measure of a regression’s fit to account for the extra degrees of freedom that adding additional X variables absorbs. In this example, even after this adjustment there is a large improvement in variation explained by the regression with the seasonal dummy variables as demonstrated by the large increase in the adjusted R-squared. Finally, we will discuss the technical problem mentioned earlier.

SEASONALITY AND AUTOCORRELATION

The regression model makes a number of assumptions about the distribution of the error terms (i.e., the distribution of Y around its average given the values of the independent (X) variables). One of these is the rather mysterious sounding assumption that "the errors are independent." Look again at Figure 9.2. For any particular quarter, the estimated error term is the distance from the fitted line to that quarter's data point.[2] "Independence" means that knowing the size of one quarter's error does not say anything about the next quarter's error. But that isn't true here. If you tell me this quarter's sales were "well above average," i.e., well above the fitted line, then I can guess this quarter is summer, next quarter will be fall, and the fall quarter's sales will likely be well below the fitted line because of the seasonality in soda sales. This phenomenon of the failure of independence is known as autocorrelation and, much like the heteroskedasticity studied in Chapter 8, interferes with the statistical inference we do using regression. When it is present, our estimated coefficients are still unbiased estimates, but the estimated standard deviations are not, so we cannot use confidence intervals or hypothesis tests unless we correct this problem, which we did here by adding the seasonal dummy variables. We discuss autocorrelation more generally in Section 9.4, including a method for detecting it and removing it.

SUMMARY

We saw how seasonal dummies may be used to "de-trend" time series data, enabling us to estimate a yearly growth trend and seasonal effects. This also solved the problem of autocorrelation in the data.

[2] It is an estimated error term because it is calculated using the estimated regression line. The true error term is how far the data point lies from the true regression line.

9.2 Seasonality: Using Seasonal Indices in Forecasting

The Dada Soda case shows us one way to account for seasonal variations in our data. In that case, the sales seemed to vary consistently over the four quarters or seasons. We captured this variation in our estimated regression prediction by including dummy variables for the different seasons. By using these intercept dummy variables to capture the seasonal effects, we were implicitly assuming the seasonal effect was **additive**. In other words, we only allowed the season to move the regression line up or down by a constant, and we did not allow the season to change the slope of the line. In practical terms, we assumed that the summer, winter, spring, and fall effects were each of a fixed size. The effects would be identical if we were selling 1 million cases or if we were selling 100 million cases.

Sometimes, we may want to use a different model of seasonal effects, one where the effect of the season is expressed as a percentage of the number of sales. In other words, the summer effect might be to increase sales by 10%. With this model, the effect of summer at the 1-million-case level is to add about 100,000 cases; at the 100-million-case level, it would add 10 million cases. This percentage-based model is known as a **multiplicative** model of seasonality in contrast to the additive model above. Why multiplicative? Because we can express each season's effect (month's effect, day-of-the-week's effect, etc.) by a **seasonal index**, which is a number multiplied by our regression results to get a prediction.

For example, in the Harmon Foods, Inc. case (see Section 9.3), the seasonal index for January shipments is 113. This number should be interpreted as saying that, all else equal, shipments in January will be 113% (or 1.13 times) the average of all months' shipments. We say all else equal because we know other factors such as a time trend or advertising affect shipments.

So, how can you use these seasonal indices in combination with regression to make forecasts?

Step 1: Deseasonalize the Y variable by dividing each observation by its corresponding seasonal index (converted from percentages if necessary). In the Harmon Foods, Inc. case, this means dividing January shipments by 1.13, February shipments by 0.98, etc.

Step 2: Build a regression model as usual (ignoring seasons) with the deseasonalized data as your Y variable.

Step 3: Use your estimated regression model to get a predicted deseasonalized value for the time period of interest.

Step 4: Multiply this predicted value by the appropriate seasonal index to get a prediction. You should multiply any interval estimates by the seasonal index as well.

That's all there is to it. If the seasonal effect works in percentage terms, the multiplicative model and seasonal indices will be appropriate; if the seasonal effects are of a fixed absolute size, the additive model will be a better choice.

Seasonally adjusted data are data that have been deseasonalized. For example, many economic statistics such as unemployment, retail sales, and housing starts are usually reported in a deseasonalized form. How are seasonal indices estimated? Some statistics packages can do this procedure for you. In fact there are many ways, some quite complicated, to estimate seasonal effects. For a taste of how part of the U.S. government does it, go to the Bureau of Labor

Statistics web site at http://stats.bls.gov/, search for the term “seasonal adjustment,” and explore some of the links.

Often, as in the Harmon Foods, Inc. case, seasonal indices previously estimated by others (in this case, an industry group) using a large set of historical and industry-wide or country-wide data are provided; thus, these indices do not need to be estimated from your data. You need only use them in your analysis.

9.3 The Harmon Foods, Inc., Case

The Harmon Foods, Inc. case is located in the packet of cases bundled to the back of this text.

QUESTIONS TO PREPARE:

1. Using only the data giving monthly shipments of Treat (and possibly a time trend, but no variables that allow for seasonal or monthly cycles), provide a forecast for shipments of Treat in January 1988. Give a 95% prediction interval for this forecast. This forecast shows what one can do without the rest of the data in the dataset and without seasonal information.
2. Develop and estimate a model you think makes the most sense to use for forecasting monthly shipments of Treat cereal. How did you arrive at this model?
3. Use the model you developed above to forecast shipments for January 1988 assuming that 200,000 consumer packs are shipped in that month and $120,000 in dealer allowances are provided. Give a 95% prediction interval for your forecast.

4. Use your estimated model to comment on the impact and effectiveness of consumer promotions and dealer promotions.
5. What improvements, if any, would you recommend to the product manager in terms of the timing and amounts of dealer promotions and consumer promotions in the future?

9.4 Regression Analysis of Time Series Data

Most of the datasets that we have encountered in previous chapters are so-called **cross-sectional samples**: We have some data on a population (e.g., car buyers, newspaper subscribers) at a fixed point in time, and analyze the relationship among various variables in the sample (e.g., price and income, Sunday and daily circulations). Time plays no role in these analyses. In other datasets, notably in the Harmon Foods and Dada Soda cases, we have consecutive observations of several variables (sales of the product and marketing efforts). These data are called **time series data.**

When we work with a time series dataset and build a regression model to explain a dependent variable, we should immediately consider including two types of variables among the explanatory variables: a **time index** (a variable that increases by one every period, representing a linear trend) and **seasonal dummies** (variables that allow us to represent seasonal variations in the dependent variable).[3] Another lesson that we learned in the Harmon Foods case is that, in the regression, we can easily incorporate the idea that our current actions matter for the future by using **lagged explanatory variables**.

[3] To set an existing time variable (e.g., the variable **quarter** from the **soda** file) as a time index in Stata, you can type the command **tsset *varname***. See the list of new Stata functions at the end of the chapter for more details.

We mentioned earlier a new problem that may arise when we run a regression using a time series dataset. We may encounter the problem of **autocorrelated residuals**: The error terms that represent the difference between the actual observations of the Y variable and the theoretical regression line may not be independent (completely random) over time. This is the case, for example, if the shocks that affect the dependent variable are persistent over time.

Suppose the Y variable represents the sales of our product. If sales this week were higher than expected due to a random event (e.g., good weather, a favorable review in the local paper), it is likely that we will be "lucky" next week as well since weather tends to persist, information about the review will diffuse among our potential customers, etc.

Autocorrelation of the residuals has the same consequences as heteroskedasticity: The standard errors (of the coefficients, the estimated mean, the regression and the prediction) become unreliable. In particular, in the most common forms of autocorrelation, the standard errors on the coefficients will be underestimated, resulting in p-values for the coefficients that appear to be lower than they are in reality. As a result, we may conclude that a coefficient is significant when in reality it is not. In a time series regression, one must be exceptionally wary of this possibility.

Another issue in time series regressions is if we can include the lagged dependent variable among the regressors. If residuals are autocorrelated, then the inclusion of lagged Y among the X-variables will cause bias in the coefficients and must be avoided.

To see this, consider the following example. Suppose (as in the Harmon Foods case) we have a time series dataset, where our dependent variable is **Sales** and the explanatory variables measure marketing efforts (e.g., number of **Coupons** issued, cash **Incentives** provided to dealers). It is reasonable to believe that promotions have different immediate and delayed effects (e.g., consumers stock up on the product when there is a discount). Moreover, **Sales** in previous periods may affect our current sales, e.g., satisfied customers tend to become repeat customers. You may think a variable like **Sales_1** (**Sales** lagged one period) could successfully represent the effects of our past actions (promotions and the resulting sales) on our current sales.

However, it may be wrong to regress **Sales** on **Coupons**, **Incentives**, a **time index**, **seasonal dummies**, and **Sales_1**. Why? **Sales_1** may be correlated with the error term in this regression because **Sales_1** contains last period's error term and errors may be autocorrelated. For example, the error term may reflect the effects of a newspaper review on **Sales**, and that effect is likely to be persistent. The error term essentially stands for all variables omitted from the regression, and we know coefficients become biased when an included variable (**Sales_1**) is correlated with omitted variables. Therefore, if error terms are autocorrelated then including **Sales_1** leads to biased coefficients. Instead of including the lagged dependent variable (**Sales_1**), you should include **lagged explanatory variables** to represent the idea that our past actions (marketing efforts) matter for current **Sales**.

Several simple and intuitive tests exist to detect specific forms of autocorrelation in the residuals. For example, after having run the regression of the dependent variable on the appropriate explanatory variables, you can regress the residuals (the difference between the actual and the fitted values of Y for each observation) on past values of the residuals (lagged residuals) and see if the coefficient on the lagged residuals is significant.

In Stata, you can generate the residuals (from your most recent regression) from the custom menu by clicking **User>Core Statistics>Model Analysis, using most recent regression>Residuals, outliers and influential observations (inflobs)** or typing **db inflobs** or, to get the residuals alone, through the standard menus by clicking **Statistics>Postestimation>Predictions, residuals, etc** or typing **db predict**. Choose **Residuals (equation-level scores)** from the "Prediction" field and type the name that you want in the "New variable name" field.[4] (We will name our residuals **residuals** for illustrative purpose here.) To create once-lagged residuals, you can type the command **generate residual_1=residuals[_n-1]** (the [_n-1] command indicates that the nth value in the **residual_1** column is taken from the n-1st value in the **residuals** column.) Then, perform a simple regression of residuals on lagged residuals. (You can run this regression with or without a constant; both produce a valid test for first-order autocorrelation given large samples.) If the slope coefficient is significant, this indicates first-order autocorrelation. This procedure is called the Cochrane-Orcutt test.

A cure for this autocorrelation is relatively simple using the Cochrane-Orcutt method. Suppose you find autocorrelation in the Cochrane-Orcutt test: The coefficient on lagged residuals in the regression of residuals, call it ρ, is significant. Transform each observation (the Y and X variables) as follows. For each observation at $t = 2,3,\ldots$, create $Y^*_t = Y_t - \rho Y_{t-1}$; similarly, create $X^*_t = X_t - \rho X_{t-1}$.[5] (The first observation is dropped because no observation occurs before it.) Now regress Y* on the transformed explanatory variable(s), X*. This new regression usually does not exhibit autocorrelated residuals; if it does, then the procedure of transforming the variables can be repeated. The coefficients on all the X* variables will be the same as the coefficients on the corresponding original X variables. However, the coefficients will have the right standard errors and p-values because autocorrelation in the residuals has been eliminated. We can rely on the new p-values for determining which variables are significant.

[4] Alternatively, you can type the direct command **predict *varname*, residuals** after running a regression.
[5] To generate Y^*_t in Stata, you can type the following command after obtaining the coefficient ρ by regressing **residuals** on **residual_1**: **generate *varname* = Y**$_t$**-_b[residual_1]* Y**$_t$**[_n-1]**. ***varname*** is the name that you would give for **Y***$_t$, **Y**$_t$ is the name of your Y variable, and **_b[residual_1]** is where Stata stores the estimated coefficient ρ from the regression of **residual** on **residual_1**. X^*_t can be generated similarly.

In Stata, you can correct for autocorrelation more easily by using Stata's built-in Prais-Winsten and Cochrane-Orcutt regression. To do this, click **Statistics>Time series>Prais-Winsten regression** or type **db prais**. Specify your dependent variable and independent variable(s). Check the boxes corresponding to "Cochrane-Orcutt transformation" and "Stop after the first iteration (twostep)."[6] Click **OK**, and Stata will report the estimated coefficient(s) on the X* variable(s). Note that these estimates agree with those produced using the manual procedure described above.[7] Stata also lists the estimated coefficient on lagged residuals next to **rho**. Note that Stata estimates ρ by regressing residuals on lagged residuals without a constant.

If you do not check the "Cochrane-Orcutt transformation" box, Stata will run the default Prais-Winsten regression instead, where it keeps and transforms the first observation into $Y^*_1 = \sqrt{1-\rho^2}\ Y_1$; (Likewise for X^*_1.) For t>1, Y^*_t and X^*_t are transformed using the method described in the previous paragraph. The difference between using the Prais-Winsten method and the Cochrane-Orcutt method is small when you have large samples.

Finally, if you do not check the "Stop after the first iteration (twostep)" box, Stata will automatically repeat the transformation procedure until the estimate of ρ becomes stable. Both iterative methods are theoretically equally valid.

Another test for autocorrelation is the Durbin-Watson test. In Stata, you can click **User>Core Statistics>Model analysis, using most recent regression>Default Durbin-Watson statistic**

[6] Alternatively, you can directly type the command **prais *depvar indepvars*, corc twostep**.

[7] There is one small difference. When you estimate the model $Y^*_t=\beta_0(1-\rho)+\beta_1 X^*_t+u^*_t$ using Stata's built-in Cochrane-Orcutt transformation, the reported estimated constant is an estimate of β_0. On the other hand, if you estimate this model using the manual transformation described above, the reported estimated constant will correspond to the estimate of $\beta_0(1-\rho)$.

(ddw) or type **db ddw**. Stata reports the Durbin-Watson d-statistic, which can range between 0 and 4 and should be close to 2 if there is no autocorrelation. Positive autocorrelation tends to lower the value of the d-statistic, while negative autocorrelation raises the value.

SUMMARY

Everything described thus far belongs to what we can call the "traditional econometric analysis" of time series data. We can apply the same regression techniques that we use for cross-sectional analyses. The only differences relative to a cross-sectional regression are the following:

1. New candidates for regressors like a **time index**, **seasonal dummies** and **lagged x variables**
2. The potential problem of **autocorrelated residuals** (resulting in incorrect standard error estimates)

9.5 Time Series Analysis

We can also use a different approach to analyzing time series data, called **time series analysis**. Time series analysis, in its purest form, ignores ordinary explanatory variables and, instead, focuses on estimating the dynamic behavior of the dependent variable alone. In other words, time series analysis is the science (and sometimes art) of extrapolation from a series of numbers, Y_1, Y_2, …, Y_T , without using any X variables except time and seasonality.

For example, one simple method of extrapolation (forecasting Y_{T+1} based on Y_1, Y_2, …, Y_T) is linear trend extrapolation. You can do this by regressing Y against a time index. Another method, exponential trend extrapolation, is carried out by regressing ln(Y) on a time index. To make both models fit better, we can enrich them each with seasonal dummies. In what follows, we discuss more sophisticated, but similarly atheoretical (i.e., no underlying model or theory) methods.

There are at least three reasons for interest in such simplistic, naïve methods of forecasting. First, in practice, collecting data on potential explanatory variables to carry out a proper regression analysis is sometimes too expensive; the only data readily available may be a series of observations regarding the dependent variable. Second, even if we can obtain the extra information and build a proper regression model, time series forecasts are cheap, require little effort to produce and can serve as a useful benchmark for comparison purposes; running a time series analysis may uncover patterns that we will explain using regression methods. Third, a sophisticated time series forecast (for example, the ARIMA model, which we will describe below) may well outperform an unsophisticated (or incorrectly specified) econometric model. In the 1970s and 1980s, time series models became popular after several studies showed the superiority of ARIMA models over standard econometric models in particular applications.

Econometric methods have since improved (e.g., in handling autocorrelation) and are generally preferred over extrapolation methods when available.

The ARIMA model of time series analysis (also called the Box-Jenkins method after its inventors in 1970) has two building blocks: autoregression (AR) and moving average (MA).

A variable Y is a p^th-order autoregressive series, AR(p) for short, if it can be written in the following way:

$$Y_t = \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + \ldots + \Phi_p Y_{t-p} + \varepsilon_t$$

$\Phi_1, \Phi_2, \ldots, \Phi_p$ are the parameters of the AR(p) process, and ε_t is an independent error term.

In other words, the current value of Y only depends on its past values (up to p lags). A variable Y is a q^th-order moving average series, MA(q) for short, if it can be written in the following way:

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}$$

$\theta_1, \theta_2, \ldots, \theta_q$ are the parameters of the MA(q) process, and the ε terms are independent errors. In other words, the current value of Y is a weighted sum of current and past (unobservable) disturbances.

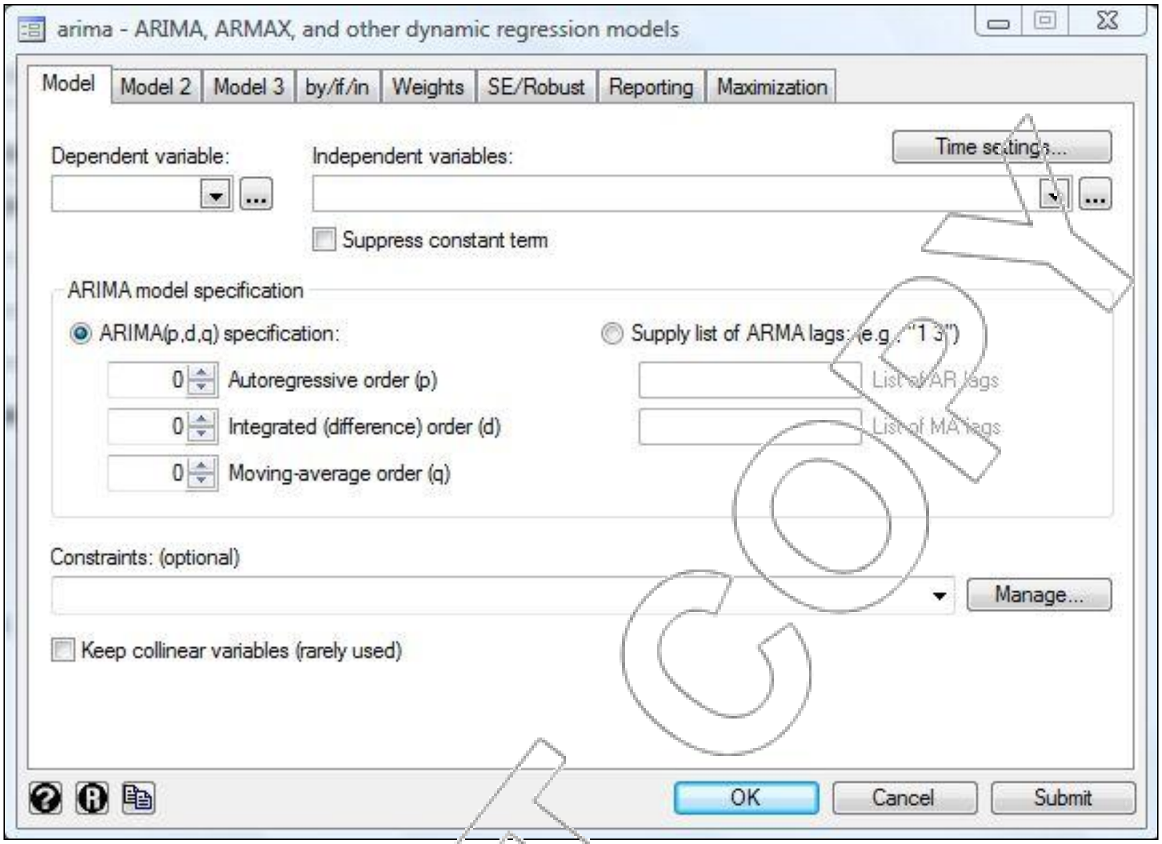
The ARIMA(p,d,q) model is more general than AR or MA. First, we difference the original series d times. Differencing a series means that we replace Y_t with $Y_t - Y_{t-1}$; that is, we consider the increments of the series instead of the series itself. We call the original Y series an ARIMA(p,d,q) process if, after differencing it d times, the resulting series Y* can be written in the following way:

$$Y^*_t = \Phi_1 Y^*_{t-1} + \Phi_2 Y^*_{t-2} + \ldots + \Phi_p Y^*_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

ARIMA(p,d,q) can be thought of as a model where the d^{th} difference of Y follows an AR(p) process such that the error term is MA(q).

There is no reason as to why a variable Y should follow an ARIMA process. ARIMA is not supported by any formal economic theory; it is a general class of random processes widely used in practice for forecasting without using explanatory variables. For example, if Y is generated by the famous “random walk” process, then it is ARIMA with $p = 0$, $d = 1$, and $q = 0$. If one decides to model Y as an ARIMA(p,d,q) process with a given p, d, and q, then a computer program (such as Stata) can estimate the parameters Φ_1, Φ_2, …, Φ_p , and θ_1, θ_2, …, θ_q. Given these parameters, you can forecast future values or see how the past (observed) values of Y fit the ARIMA model.

In Stata, you can compute the parameters of a general ARIMA(p,d,q) process by clicking **Statistics>Time series>ARIMA and ARMAX models** or typing **db arima**. This will open the following dialog box:


arima - ARIMA, ARMAX, and other dynamic regression models
Model
Model 2
Model 3
by/if/in
Weights
SE/Robust
Reporting
Maximization
Dependent variable:
Independent variables:
Time settings...
□ Suppress constant term
ARIMA model specification
● ARIMA(p,d,q) specification:
○ Supply list of ARMA lags: (e.g., "1 3")
0 Autoregressive order (p)
0 Integrated (difference) order (d)
0 Moving-average order (q)
List of AR lags
List of MA lags
Constraints: (optional)
Manage...
□ Keep collinear variables (rarely used)
OK
Cancel
Submit

Select your dependent variable and independent variable(s) from the respective drop-down lists. Check the box next to "Suppress constant term" and enter corresponding values for p, d, and q in the "ARIMA(p,d,q) specification" field. Click **OK**, and Stata will display its ARIMA regression result.[8] Under the **Coef.** column, you can find Stata's estimates for Φ_p and θ_q in the **ar Lp.** and **ma Lq.** rows, respectively.

The main practical question that remains is how to choose the parameters p, d, and q for an ARIMA model and forecast. Time series analysts would probably say that this is the "art" part of forecasting. The most important guideline is to keep these parameters as low as possible (parsimony). In general, choose d, the number of times the series is differenced, to make the

[8] Alternatively, you can directly type the command **arima *depvar indepvar*, noconstant arima(p,d,q)**. Omit ***indepvar*** if you are not including any explanatory variable.

series stationary, which means that the mean, variance, and other properties of Y* must not depend on time. Usually d = 1 or d = 2 suffices.

To find the “right” parameters p and q, time series analysts usually look at a diagram called a correlogram. To create this diagram, for all k = 1,2,…, we compute ρ_k, the correlation coefficient between Y* and Y* lagged k times, and plot ρ_k against k. The correlogram should fall off to numbers close to zero as k increases; otherwise, Y* is not stationary and needs to be differenced further. A correlation coefficient ρ_k on the correlogram is called significant if it is greater in absolute value than $2/\sqrt{T}$, where T is the number of observations.

The pattern on the correlogram suggests the appropriate numbers for p and q. For example, if ρ_1 (respectively, ρ_1 and ρ_2) are significant but the subsequent ρ_k values look random, then Y* is an MA(1) (respectively, MA(2)) process. If the correlogram declines geometrically, then Y* can be modeled as an AR(1) process. If it exhibits a wave, then AR(2) or a higher order AR process is required. If the correlogram appears to decline geometrically but the sign of ρ_1 does not match the signs of the rest of the ρ_k values, then ARIMA(1,d,1) is suggested.

We summarize ARIMA by working out an example. The **Kodak** file contains the annual gross revenues of Eastman Kodak Co. between 1975 and 1999 (in billions of constant 1982 dollars). Plotting the data in Figure 9.5, there is no visible trend, so we do not difference the series (d=0).

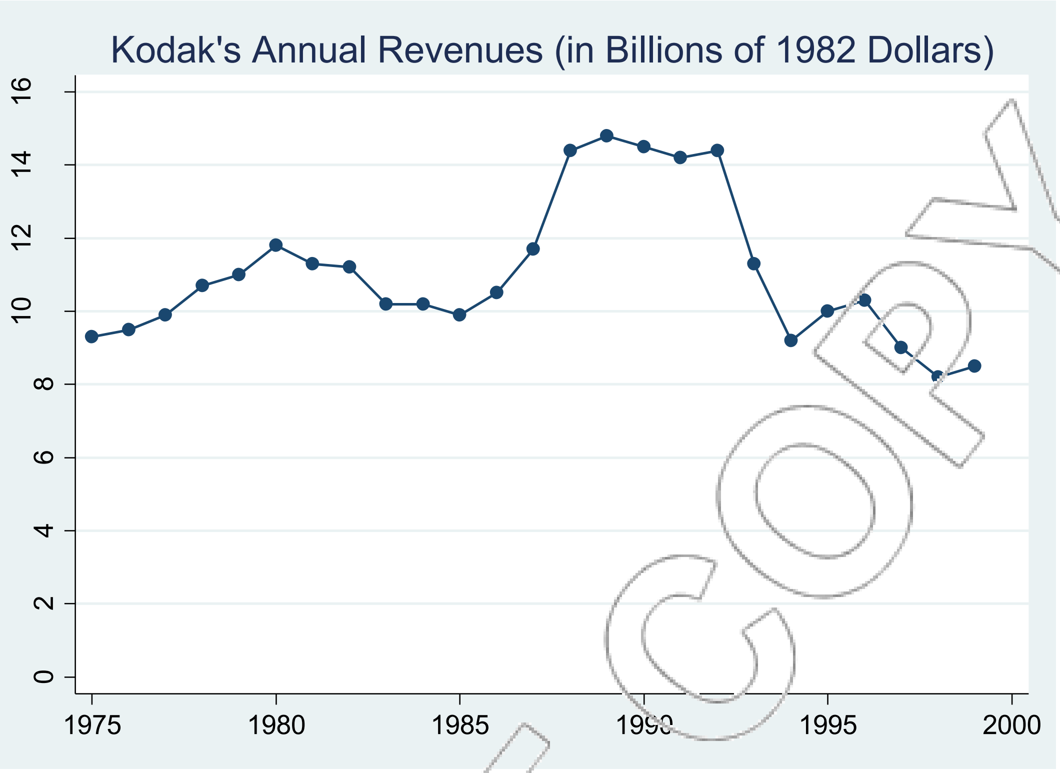


Figure 9.5. Kodak's annual revenues.

Next, we look at the correlogram in Figure 9.6. This graph can be generated in Stata by clicking **Graphics>Time-series graphs>Correlogram (ac)** or typing **db ac**, choosing **Revenue** as the variable, and typing **7** in the "Number of autocorrelations to compute" field.[9]

[9] Alternatively, you can directly type the command **ac Revenue, lags(7)**. Note that instead of plotting ρ_k, the correlation coefficient between Y* and Y* lagged k times, against k, the **ac** command plots the autocorrelations of Y* against its lags. Although autocorrelation is defined as the correlation between a time series variable and its lags, it is calculated using a slightly different formula than the standard formula used to calculate the correlation coefficient between two generic variables. You can refer to Stata's PDF manuals for the respective formulas that Stata uses to calculate ρ_k and autocorrelations.

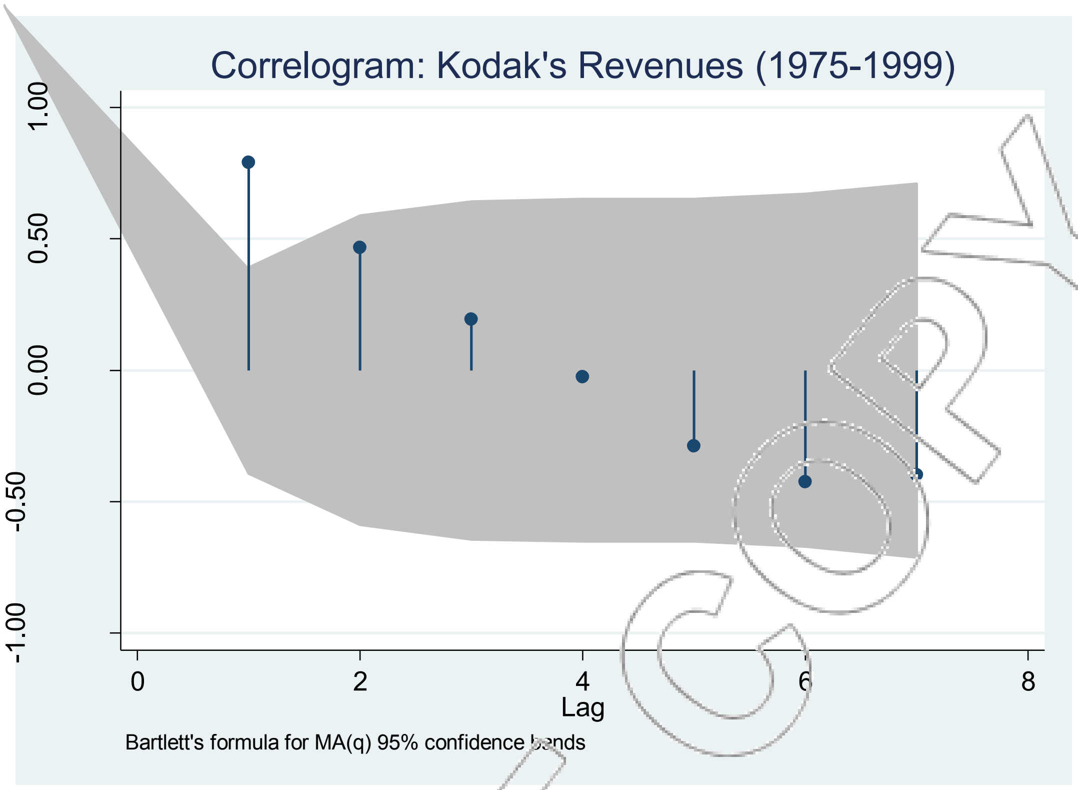


Figure 9.6: Correlogram of Kodak revenues.

The decline in ρ_k appears to be steady (and approximately linear); the first two ρ_k values are significant (greater than $2/\sqrt{25} = .4$ in absolute value), but the rest do not appear to be random either (this is not an MA process). An AR(1) process seems to be appropriate. When we run the AR(1) regression in Stata, we find that the estimated AR(1) process can be written in the following way.[10]

$$\boxed{\textbf{Revenue} = 1.63692 + .8501455{*}\textbf{Revenue\_1}.}$$

[10] To run this regression, you can directly type the command **regress Revenue L1.Revenue** after declaring **Year** as your time index variable using the **tsset** command. **L1.Revenue** equals **Revenue** lagged one period. In the boxed AR(1) equation, **Revenue_1=L1.Revenue**.

To check how well the AR(1) process fits the data, we can estimate Kodak's revenues for the years 1976–1999 and calculate the **mean absolute deviation** (MAD) from the actual observations. To do this in Stata, you can type the following commands after running the AR(1) regression: 1) **predict residual, residuals**; 2) **generate abs_residual=abs(residual)**; and 3) **summarize abs_residual**. The **Mean** value is the MAD and turns out to be 0.761931, or about $0.76 billion. As a comparison, the average level of **Revenue** in the sample is about $11 billion (both in constant 1982 dollars).

SUMMARY

Though there is no theoretical reason why a particular variable might follow a linear or exponential trend, the techniques we have seen are useful. Predicting future performance using these methods has its drawbacks. However, the advantages mentioned earlier (including the value of the ARIMA model when the only data available are for the dependent variable and for establishing a baseline) make knowledge of this approach worthwhile.

NEW TERMS

Additive model
A regression model using dummy variables to account for seasonality. Each season is assumed to have a fixed effect on the dependent variable

Multiplicative model
A regression model which assumes each season affects the dependent variable by a certain percentage

Seasonal index
An index used to seasonalize and deseasonalize the dependent variable and predictions in a multiplicative seasonality model

Time series data
Consecutive observations of a set of variables

Time index	A variable that increases by one every time period. Used to model a linear trend over time
Lagged variables	Variables that use values from a previous time period to explain outcomes in the current time period
Autocorrelated residuals	A problem where the error terms are not independent
Cochrane-Orcutt test	A test for autocorrelation using the residuals
Linear trend extrapolation	A time series method used to model linear trends in Y over time
Exponential trend extrapolation	A time series method used to model linear trends in ln(Y) over time
ARIMA or Box-Jenkins method	A time series method employing autoregression (AR) and moving average (MA) techniques
Stationary	A model where the properties of Y* do not depend on time
Correlogram	A diagram used to determine the proper time series parameters

NEW STATA FUNCTIONS

Statistics>Time series>Setup and utilities>Declare dataset to be time-series data

Equivalently, you may type **db tsset**. This command opens the **tsset** dialog box. You can select the variable that you want to declare as a time index from the "Time variable" field.

Alternatively, you can directly type the command **tsset *varname***.

To create a generic time index using observation numbers, you can enter the following commands: 1) **generate *newvar*=[_n]**, and 2) **tsset *newvar***.

Statistics>Time series>Prais-Winsten regression

Equivalently, you may type **db prais**. This command opens the **prais** dialog box, where you can ask Stata to implement either the Cochrane-Orcutt transformation or the Prais-Winsten transformation to correct for autocorrelation by checking/unchecking the "Cochrane-Orcutt transformation" box. Check the "Stop after the first iteration (twostep)" box if you want Stata to transform your variables only once, as described at the end of Section 9.4. Note that you need to declare a time index variable using the **tsset** command before running the **prais** command.

Alternatively, you can directly type the command **prais *depvar indepvars*, corc twostep**. Omitting the **corc** option will implement the Prais-Winsten transformation instead.

User>Core Statistics>Model analysis, using most recent regression>Default Durbin-Watson statistic (ddw)

Equivalently, you may type **db ddw**. Click **OK** in the ensuing **ddw** dialog box, and Stata will report the Durbin-Watson d-statistic with which you can use to detect autocorrelation in the residuals. The d-statistic ranges between 0 and 4 and should be close to 2 if there is no autocorrelation. Positive autocorrelation tends to lower the value of the d-statistic, while negative autocorrelation raises the value.

Statistics>Time series>ARIMA and ARMAX models

Equivalently, you may type **db arima**. This command opens the **arima** dialog box, where you can specify the dependent variable, independent variable(s) (if any), and the order numbers for p, d, q according to your model. Stata reports the estimated values for Φ_p and θ_q in the **ar Lp.** and **ma Lp.** rows, respectively. Note that you need to declare a time index variable by using the **tsset** command before running an ARIMA(p,d,q) regression.

Alternatively, you can directly type the command **arima *depvar indepvars*, arima(p,d,q)**.

Graphics>Time-series graphs>Correlogram (ac)

Equivalently, you may type **db ac**. This command opens the **ac** dialog box, where you can select the variable for which you want to generate a correlogram. You can specify the number of lags in the “Number of autocorrelations to compute” field. Note that you need to declare a time index variable using the **tsset** command before generating a correlogram.

Alternatively, you can directly type the command **ac *varname*, lags(#)**.

CASE EXERCISES

1. Harmon Foods

Read the Harmon Foods case and prepare answers to the five questions listed in Section 9.3 of this chapter.

2. Paradise tax

The governor of the state of Hawaii is bound by the state constitution to budget no more funds than the amount projected by the State Council on Revenues. Part of this revenue is from the transient accommodations tax, which is a hotel tax. Forecasting the tax revenues from this and other tourism taxes are important to the state as well as the major businesses operating in the tourism industry. The data in the **hawaiiTAT**[11] file contains information from 1990 through the summer of 2003 regarding the quarterly collection of this tax as well as statistics such as visitor days (the number of days spent by visiting tourists each quarter) and the average daily room rate.

[11] Derived from http://www2.hawaii.gov/DBEDT/.

Furthermore, a seasonal index based on visitor arrivals by plane (no tourists swim or drive to the islands though a tiny percentage arrives by boat) has been constructed as well.

Develop an additive and a multiplicative model to forecast the state's collection of the transient accommodations tax. Which model do you feel is the better choice to make a prediction for the fall of 2003 when the room rates are expected to average $133 per night with 14,000,000 visitor days? Provide estimates from each model and justify your choice.

3. Restaurant Planning

The owners of Blue Stem, an upscale restaurant in a trendy area of Chicago, have gathered data on its nightly receipts. Over the year, the restaurant occasionally offers a free dessert promotion to ticket holders from the theater next door. The promotions occur mostly on the weekends, which are the most popular nights for dining out. The restaurant would like to separate the promotion effect from the weekend effect, so it can determine if the promotion is worthwhile. The data are available in the **bluestem**[12] file.

An industry group has provided a nightly index reflecting the relative popularity of different nights for higher end restaurants in the city.

Develop two models, one using additive and one using multiplicative techniques, to test the effectiveness of the promotion. In each case, report how much, on average, the promotion boosts revenues on a Saturday night.

[12] Source: Linda Hall, Co-owner Blue Stem Restaurant

CASE INSERT 3

NOPANE ADVERTISING STRATEGY

In this case, we will look at the advertising strategy for a drug, Nopane. The brand manager is faced with the choice of advertising level, copy, and region in the face of intense competition. The assignment is to read the case and answer the following questions. For the first three, you can use the regressions included with the case; however, you will need to conduct your own analysis using Stata to respond to the additional questions.

Questions to Prepare

1. What does Regression 1 in the case say about the merits of “emotional” vs. “rational” copy? What does Regression 3 say about the two types of copy? What is the interpretation of the coefficient on copy in Regression 1? Regression 3?

2. Assuming Alison Silk’s hypothesis is correct, which of the regressions is most relevant for choosing an advertising strategy? Why?

3. Answer question 2, assuming instead that Stanley Skamarycz’s hypothesis is correct.

4. Given the data from the case (in the **nopane** file), what national advertising strategy (i.e., which copy and which one of the three levels of ad spending) would you advocate? Each additional unit sold per 100 prospects over a six-month period yields a profit (net of

production and delivery costs, but not net of advertising costs) of $10. Provide support for your position.

5. Instead of a single national campaign, Ms. Silk knows it would be possible (though more costly) to have one campaign for the East and West Coast states and another for the middle of the country. Comment on the desirability of splitting up the campaign.

Hints: Remember omitted variable bias. For questions 4 and 5, you may want to think about using dummy variables and/or slope dummy variables.

The Nopane Advertising Strategy case is located in the packet of cases bundled to the back of this text.[1]

[1] Nopane Advertising Strategy, Harvard Business School Case, Product #9-893-005.

CASE INSERT 4

THE BASEBALL CASE

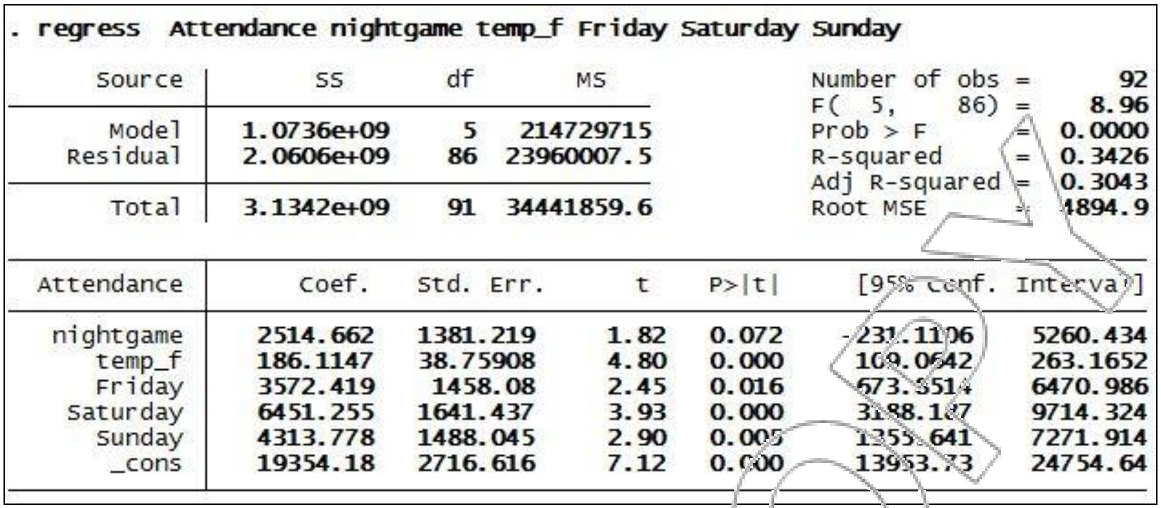
Singha Field is home to the BK Lions professional baseball team. The team's new marketing director, Noelle Amsley, has been trying to develop a better understanding of the key drivers of attendance at the ballpark to increase ticket revenues, optimize concession inventories and staffing, and schedule the timing of promotional giveaways.

The stadium is capable of holding almost 41,000 fans. The exact number is hard to pin down due to the sale of standing-room-only tickets and VIP ticket comping. The data for this case are included in the file **baseball case**.

PART A: REGRESSION ANALYSIS

Noelle's first model uses three concepts to predict attendance: time of day, temperature, and day of the week. Specifically, she has a dummy variable for **night** games, the day's high **temperature**, and three dummies indicating if the game takes place on a **Friday**, **Saturday**, or **Sunday**, respectively.

1. Use Regression 1 to estimate attendance for a Sunday afternoon game where the temperature is 82 degrees.

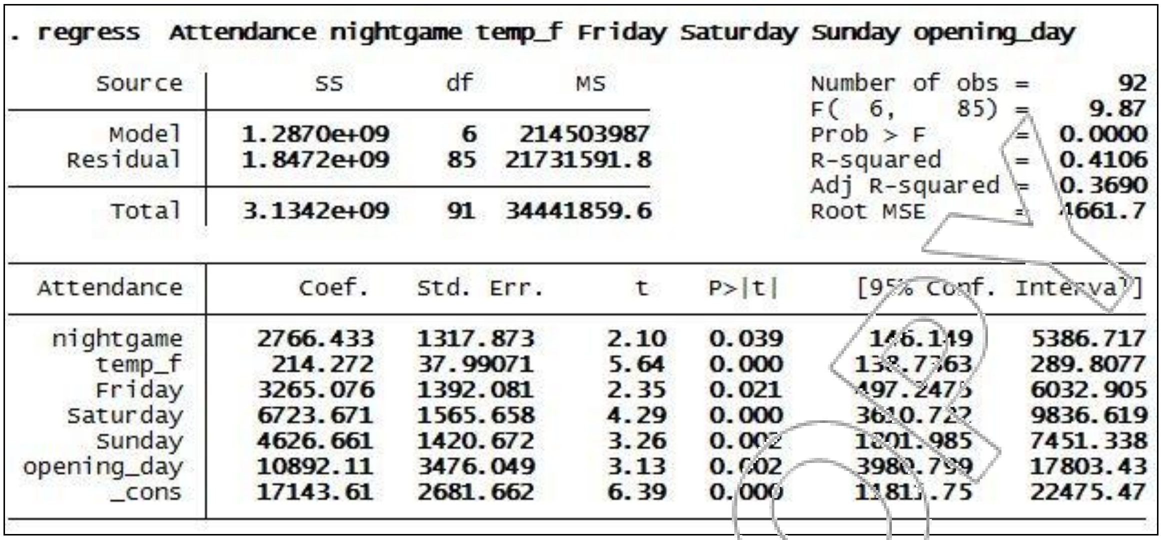


```
. regress  Attendance nightgame temp_f Friday Saturday Sunday

      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F(  5,    86) =    8.96
       Model |  1.0736e+09     5   214729715           Prob > F      =  0.0000
    Residual |  2.0606e+09    86  23960007.5           R-squared     =  0.3426
-------------+------------------------------           Adj R-squared =  0.3043
       Total |  3.1342e+09    91  34441859.6           Root MSE      =  4894.9

------------------------------------------------------------------------------
  Attendance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   nightgame |   2514.662   1381.219     1.82   0.072    -231.1106    5260.434
      temp_f |   186.1147   38.75908     4.80   0.000     109.0642    263.1652
      Friday |   3572.419    1458.08     2.45   0.016     673.8514    6470.986
    Saturday |   6451.255   1641.437     3.93   0.000     3188.187    9714.324
      Sunday |   4313.778   1488.045     2.90   0.005     1355.641    7271.914
       _cons |   19354.18   2716.616     7.12   0.000     13953.73    24754.64
------------------------------------------------------------------------------
```

Regression 1

A quick look at the model analysis output from Stata (clicking **User>Core Statistics>Model Analysis, using most recent regression>Residuals, outliers, and influential observations (inflobs)** or typing **db inflobs**) shows six outliers among the 92 data points. Two of them are day games on very cold weekdays where the model predicts the lowest possible turnout. However, these particular games nearly sold out. Noelle kicks herself: They're both the opening day of the season, a special game for baseball fans.

Adding a new dummy variable called **opening_day** that equals one on the first home game of the season and zero otherwise produces Regression 2.



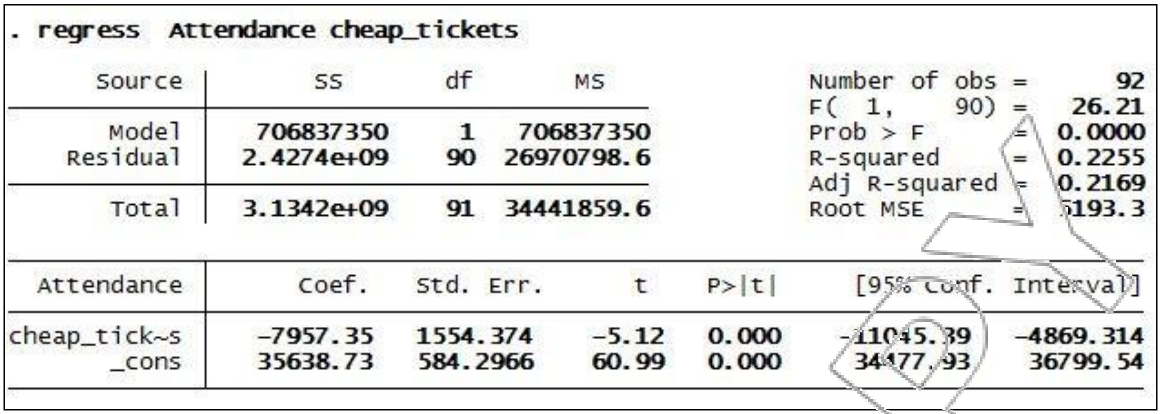
```
. regress  Attendance nightgame temp_f Friday Saturday Sunday opening_day

      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F(  6,    85) =    9.87
       Model |  1.2870e+09     6   214503987           Prob > F      =  0.0000
    Residual |  1.8472e+09    85  21731591.8           R-squared     =  0.4106
-------------+------------------------------           Adj R-squared =  0.3690
       Total |  3.1342e+09    91  34441859.6           Root MSE      =  4661.7

------------------------------------------------------------------------------
  Attendance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   nightgame |   2766.433   1317.873     2.10   0.039      146.149    5386.717
      temp_f |    214.272   37.99071     5.64   0.000     138.7363    289.8077
      Friday |   3265.076   1392.081     2.35   0.021     497.2475    6032.905
    Saturday |   6723.671   1565.658     4.29   0.000     3610.722    9836.619
      Sunday |   4626.661   1420.672     3.26   0.002     1801.985    7451.338
 opening_day |   10892.11   3476.049     3.13   0.002     3980.799    17803.43
       _cons |   17143.61   2681.662     6.39   0.000     11811.75    22475.47
------------------------------------------------------------------------------
```

Regression 2

2. Use Regression 2 to estimate the attendance for a Sunday afternoon game where the temperature is 82 degrees and it is not opening day.

3. Compare your results from questions 1 and 2. Explain why your estimate changes between the two models.

The team management recently began using a more sophisticated pricing structure to improve its revenues. Instead of charging the same set of prices for every game, there are two different pricing schemes: full-price tickets and cheap tickets. For games where management anticipates a lower level of interest, it charges the cheap ticket prices in order to stimulate demand. Regression 3 shows the significant effect of **cheap_tickets** on attendance, but the coefficient is confusing to Noelle. She had expected the sign to be positive. Shouldn't the lower prices *increase* attendance?



```
. regress  Attendance cheap_tickets

      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F(  1,    90) =   26.21
       Model |   706837350     1   706837350           Prob > F      =  0.0000
    Residual |  2.4274e+09    90  26970798.6           R-squared     =  0.2255
-------------+------------------------------           Adj R-squared =  0.2169
       Total |  3.1342e+09    91  34441859.6           Root MSE      =  5193.3

------------------------------------------------------------------------------
  Attendance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
cheap_tick~s |   -7957.35   1554.374    -5.12   0.000    -11045.39   -4869.314
       _cons |   35638.73   584.2966    60.99   0.000     34477.93    36799.54
------------------------------------------------------------------------------
```

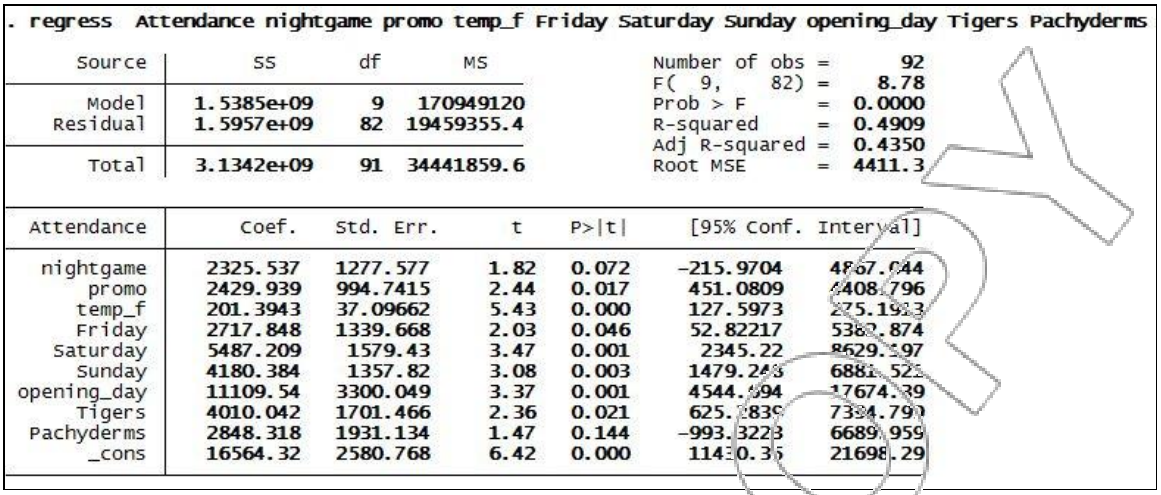
Regression 3

4. Do these results violate the law of demand that says all else being equal, a lower price should increase the quantity demanded?

Noelle's colleague, Andrew Groden, is interested in learning how two other factors are driving attendance: promotional giveaways such as free hat day; and popular opponents, such as the team's historic rivals, the ML Tigers, as well as their cross-town rivals, the Pachyderms. To test these factors' significance, Noelle has added three dummy variables called **promo**, **Tigers**, and **Pachyderms**, which are added to her earlier regression to produce Regression 4. She quickly informs Andrew that the first two are significant, but the Pachyderms do not seem to be a big draw to the ballpark.

Andrew disagrees. "It's just because those games were all scheduled on days that were already popular. Five of the six times they played were on Fridays or the weekends, and all of the games were in the summer when the weather is usually perfect! Those games increased the interest in the games, but there just weren't enough seats available in the ballpark to see the effect."

5. Does Andrew's theory sound reasonable? Why would a team schedule games against a popular rival, knowing that it did not need to encourage attendance on those dates?

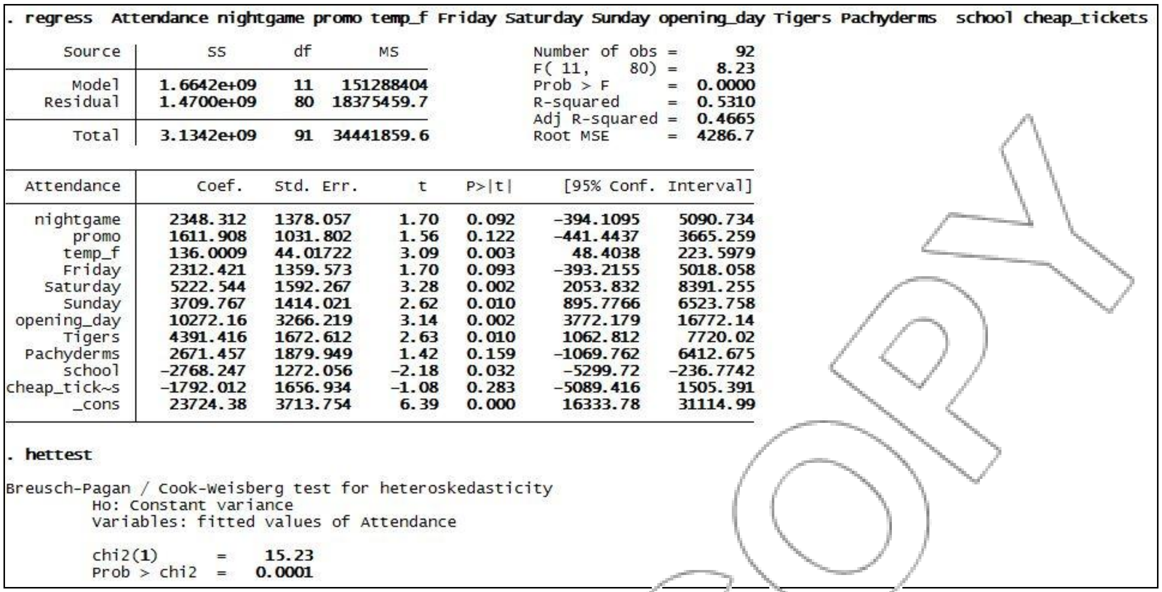


```
. regress  Attendance nightgame promo temp_f Friday Saturday Sunday opening_day Tigers Pachyderms

      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F(  9,    82) =    8.78
       Model |  1.5385e+09     9   170949120           Prob > F      =  0.0000
    Residual |  1.5957e+09    82  19459355.4           R-squared     =  0.4909
-------------+------------------------------           Adj R-squared =  0.4350
       Total |  3.1342e+09    91  34441859.6           Root MSE      =  4411.3

------------------------------------------------------------------------------
  Attendance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   nightgame |   2325.537   1277.577     1.82   0.072    -215.9704    4867.044
       promo |   2429.939   994.7415     2.44   0.017     451.0809    4408.796
      temp_f |   201.3943   37.09662     5.43   0.000     127.5973    275.1913
      Friday |   2717.848   1339.668     2.03   0.046     52.82217    5382.874
    Saturday |   5487.209    1579.43     3.47   0.001      2345.22    8629.197
      Sunday |   4180.384    1357.82     3.08   0.003     1479.243    6881.525
 opening_day |   11109.54   3300.049     3.37   0.001     4544.694    17674.39
      Tigers |   4010.042   1701.466     2.36   0.021     625.2839    7394.799
  Pachyderms |   2848.318   1931.134     1.47   0.144    -993.3223    6689.959
       _cons |   16564.32   2580.768     6.42   0.000     11430.35    21698.29
------------------------------------------------------------------------------
```

Regression 4

Regression 5 adds two more variables to Noelle's model. One is **school**, which equals one whenever the local public school system is in session (keeping thousands of potential fans away from many games) and zero otherwise. The other variable she adds is **cheap_tickets**, as was used in Regression 3.

6. Is the variable **cheap_tickets** significant in this regression? Interpret the coefficient and its significance in the context of this new regression.



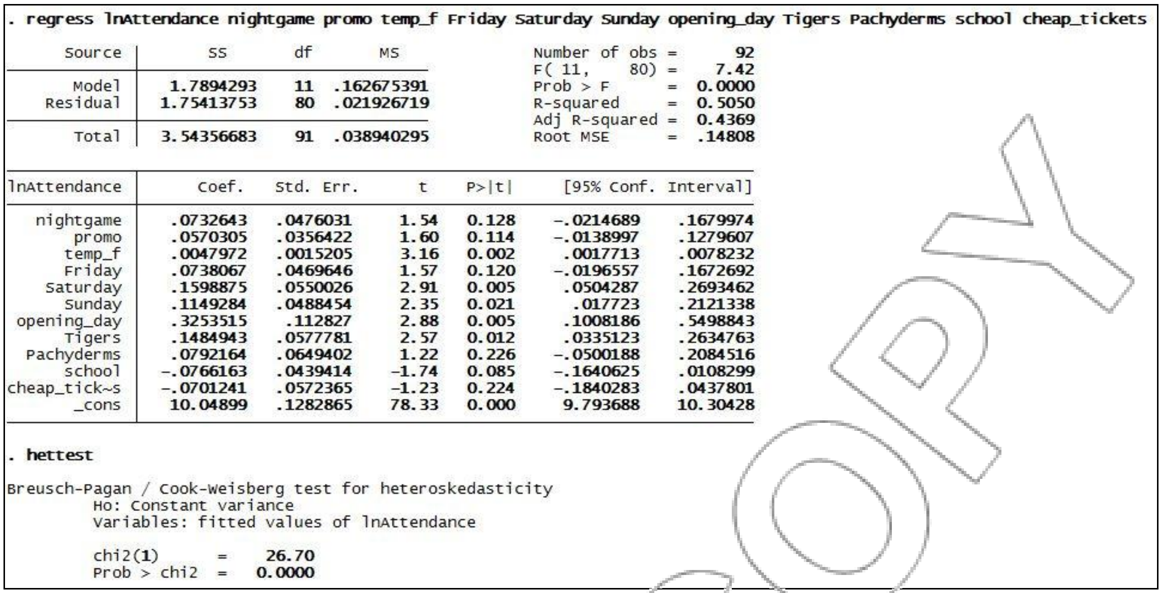
```
. regress  Attendance nightgame promo temp_f Friday Saturday Sunday opening_day Tigers Pachyderms  school cheap_tickets

      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F( 11,    80) =    8.23
       Model |  1.6642e+09    11   151288404           Prob > F      =  0.0000
    Residual |  1.4700e+09    80  18375459.7           R-squared     =  0.5310
-------------+------------------------------           Adj R-squared =  0.4665
       Total |  3.1342e+09    91  34441859.6           Root MSE      =  4286.7

------------------------------------------------------------------------------
  Attendance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   nightgame |   2348.312   1378.057     1.70   0.092    -394.1095    5090.734
       promo |   1611.908   1031.802     1.56   0.122    -441.4437    3665.259
      temp_f |   136.0009   44.01722     3.09   0.003     48.4038    223.5979
      Friday |   2312.421   1359.573     1.70   0.093    -393.2155    5018.058
    Saturday |   5222.544   1592.267     3.28   0.002     2053.832    8391.255
      Sunday |   3709.767   1414.021     2.62   0.010     895.7766    6523.758
 opening_day |   10272.16   3266.219     3.14   0.002     3772.179    16772.14
      Tigers |   4391.416   1672.612     2.63   0.010     1062.812     7720.02
  Pachyderms |   2671.457   1879.949     1.42   0.159    -1069.762    6412.675
      school |  -2768.247   1272.056    -2.18   0.032     -5299.72   -236.7742
cheap_tick~s |  -1792.012   1656.934    -1.08   0.283    -5089.416    1505.391
       _cons |   23724.38   3713.754     6.39   0.000     16333.78    31114.99
------------------------------------------------------------------------------

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of Attendance

         chi2(1)      =    15.23
         Prob > chi2  =   0.0001
```

Regression 5

7. Use Regression 5 to make a forecast of attendance for a Saturday night game against the Tigers that is not on opening day. Also, the temperature is 89 degrees, there are full-price tickets, a promotional giveaway, and school is out of session. Provide a 95% prediction interval for your answer. Do you have any concerns about your forecast?

PART B: NON-LINEARITIES

Noelle has been studying Regression 5. She is concerned about the Breusch-Pagan Test, which indicates a heteroskedasticity problem with the model. She becomes more concerned after conducting a semi-log model, Regression 6, which failed to fix the problem. Noelle suspects that a linear model may not be the most appropriate fit to the data; in particular, she is worried about the large number of games that are pushing the stadium’s capacity limits.



```
. regress lnAttendance nightgame promo temp_f Friday Saturday Sunday opening_day Tigers Pachyderms school cheap_tickets

      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F( 11,    80) =    7.42
       Model |   1.7894293    11  .162675391           Prob > F      =  0.0000
    Residual |  1.75413753    80  .021926719           R-squared     =  0.5050
-------------+------------------------------           Adj R-squared =  0.4369
       Total |  3.54356683    91  .038940295           Root MSE      =  .14808

------------------------------------------------------------------------------
lnAttendance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   nightgame |   .0732643   .0476031     1.54   0.128    -.0214689    .1679974
       promo |   .0570305   .0356422     1.60   0.114    -.0138997    .1279607
      temp_f |   .0047972   .0015205     3.16   0.002     .0017713    .0078232
      Friday |   .0738067   .0469646     1.57   0.120    -.0196557    .1672692
    Saturday |   .1598875   .0550026     2.91   0.005     .0504287    .2693462
      Sunday |   .1149284   .0488454     2.35   0.021      .017723    .2121338
 opening_day |   .3253515    .112827     2.88   0.005     .1008186    .5498843
      Tigers |   .1484943   .0577781     2.57   0.012     .0335123    .2634763
  Pachyderms |   .0792164   .0649402     1.22   0.226    -.0500188    .2084516
      school |  -.0766163   .0439414    -1.74   0.085    -.1640625    .0108299
cheap_tick~s |  -.0701241   .0572365    -1.23   0.224    -.1840283    .0437801
       _cons |   10.04899   .1282865    78.33   0.000     9.793688    10.30428
------------------------------------------------------------------------------

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnAttendance

         chi2(1)      =    26.70
         Prob > chi2  =   0.0000
```

Regression 6

Both linear and logarithmic models are unbounded, meaning they don't have an upper limit. Regression 1, for instance, predicts more than 42,000 fans for a Saturday afternoon game with a temperature of around 88 degrees (not unreasonable for a summer day) even though that exceeds the capacity of the stadium by more than a thousand people. A regression of lnAttendance using the same independent variables predicts more than 43,000 fans.

The problem as Noelle sees it is that none of the models she has learned about seems right for the pattern she observed in the dataset: attendance getting closer and closer to a maximum value as "conditions" improve. Taking temperature as the independent variable, Noelle plots Attendance versus Temperature with two different fits. These fits include one linear and one curving up toward the capacity. These plots are seen in Figures 1 and 2.

Looking at Figure 2 gives Noelle an idea. Though a semi-log model, Y = a·e^{bX} does not have a maximum when the constant a is positive, it does have a minimum. Y will never fall below zero.

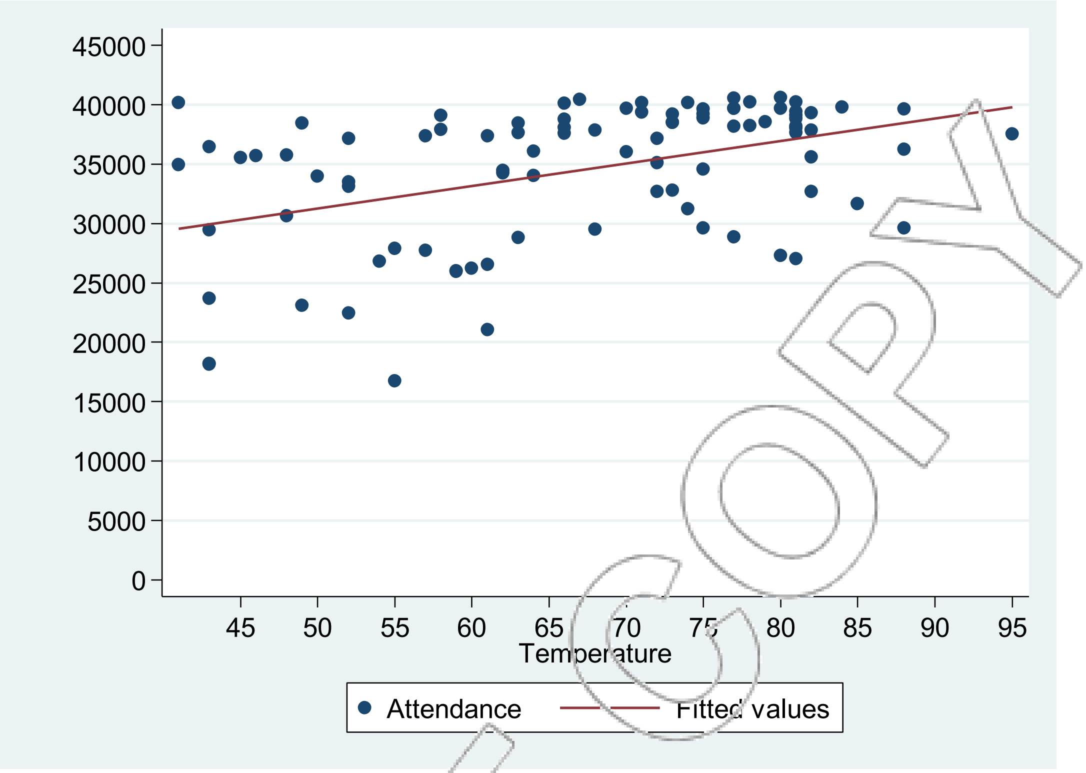

45000
40000
35000
30000
25000
20000
15000
10000
5000
0
45
50
55
60
65
70
75
80
85
90
95
Temperature
● Attendance
Fitted values

Figure 1

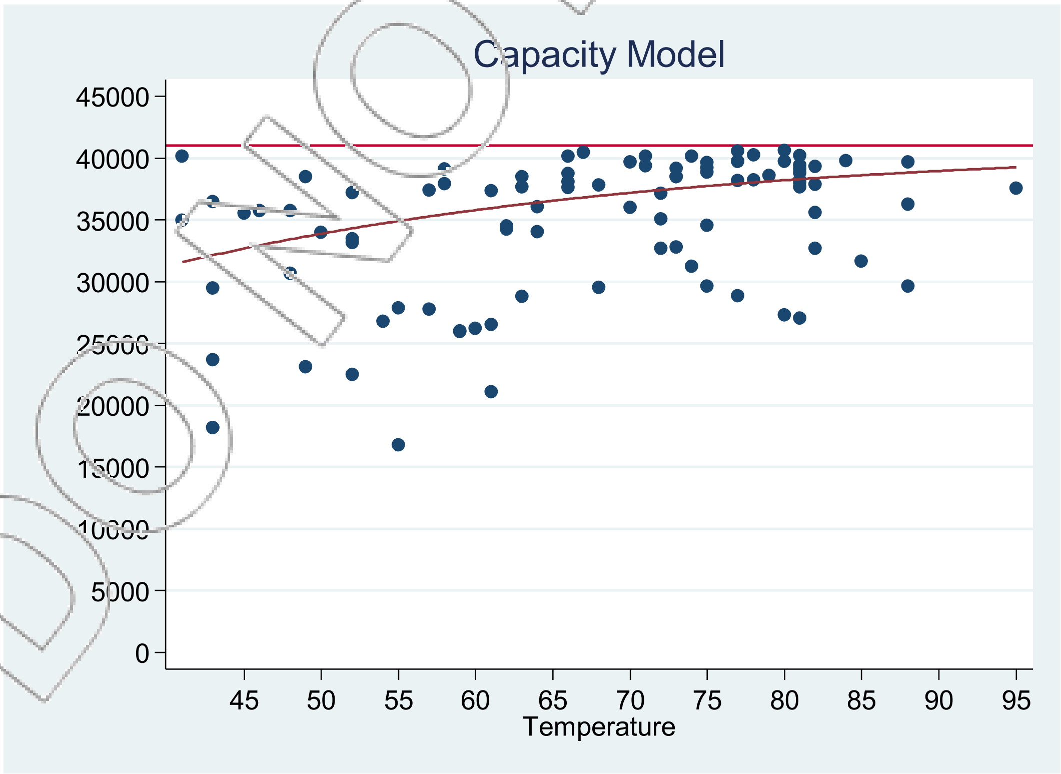

Capacity Model
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
45
50
55
60
65
70
75
80
85
90
95
Temperature

Figure 2

Flipping Figure 2 upside-down by plotting Empty Seats versus Temperature gives Noelle the graph in Figure 3, which looks just like the kind of graph where a semi-log model fits perfectly! Taking a log of the empty seats and plotting it versus Temperature gives her Figure 4. Empty seats were computed using 41,000 as the capacity. Regression 7 uses the same dependent variable but adds the entire collection of independent ones as Noelle had done previously.

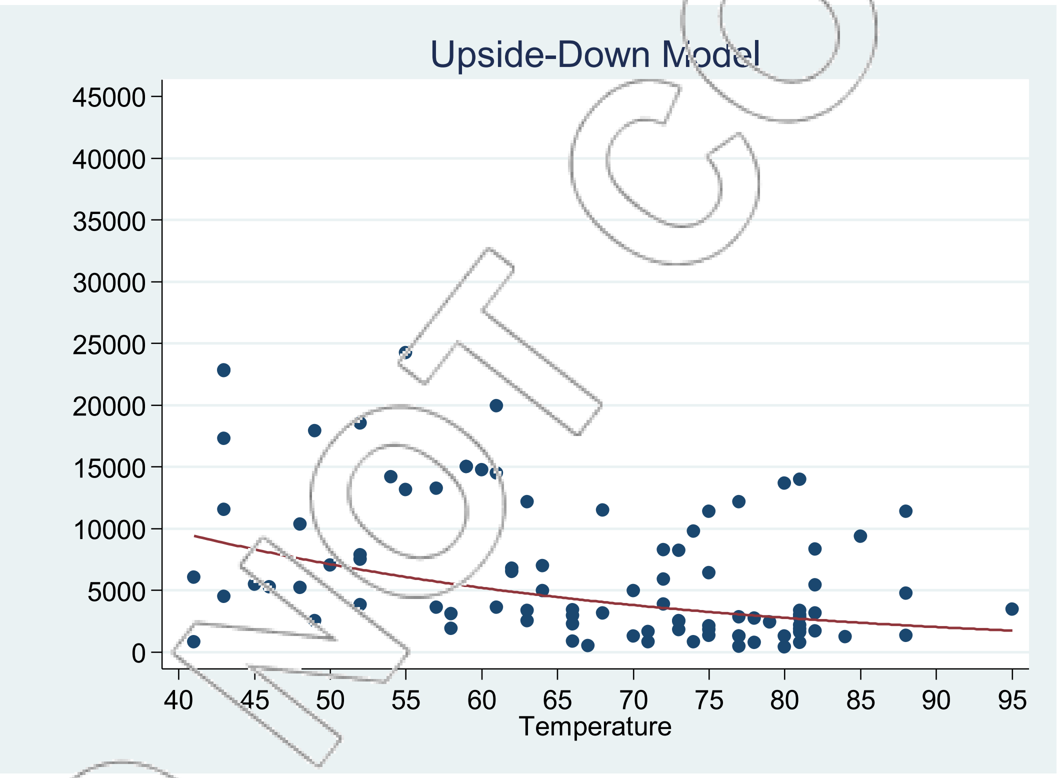


Figure 3

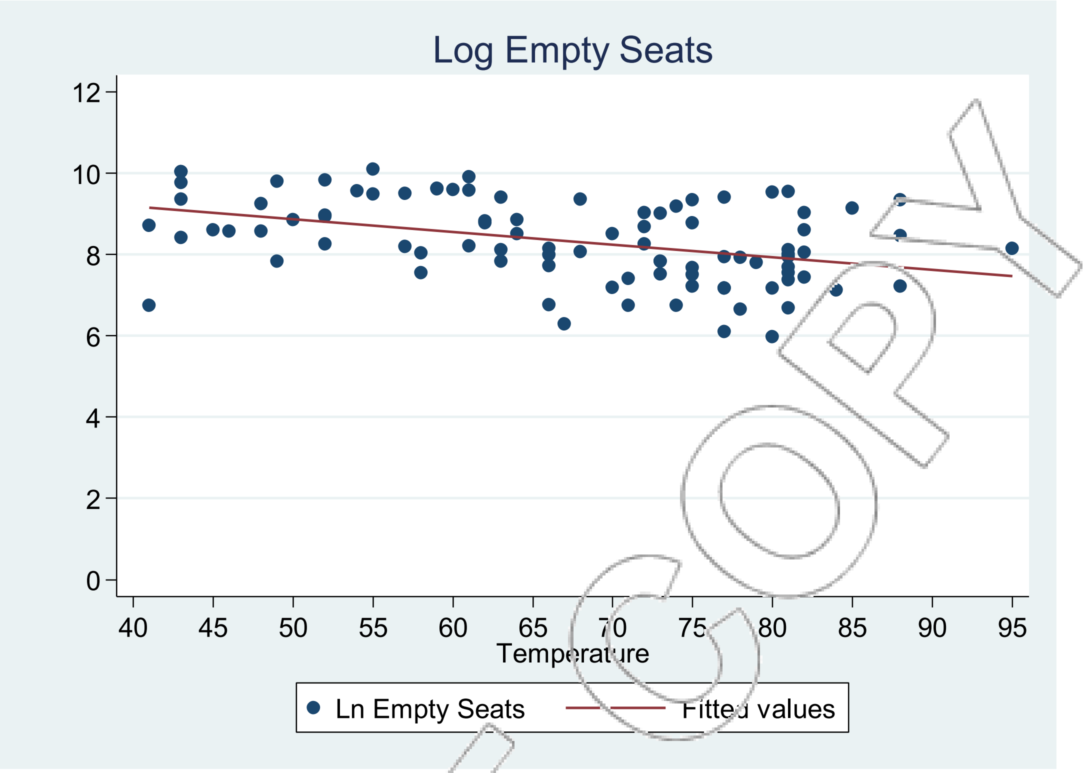
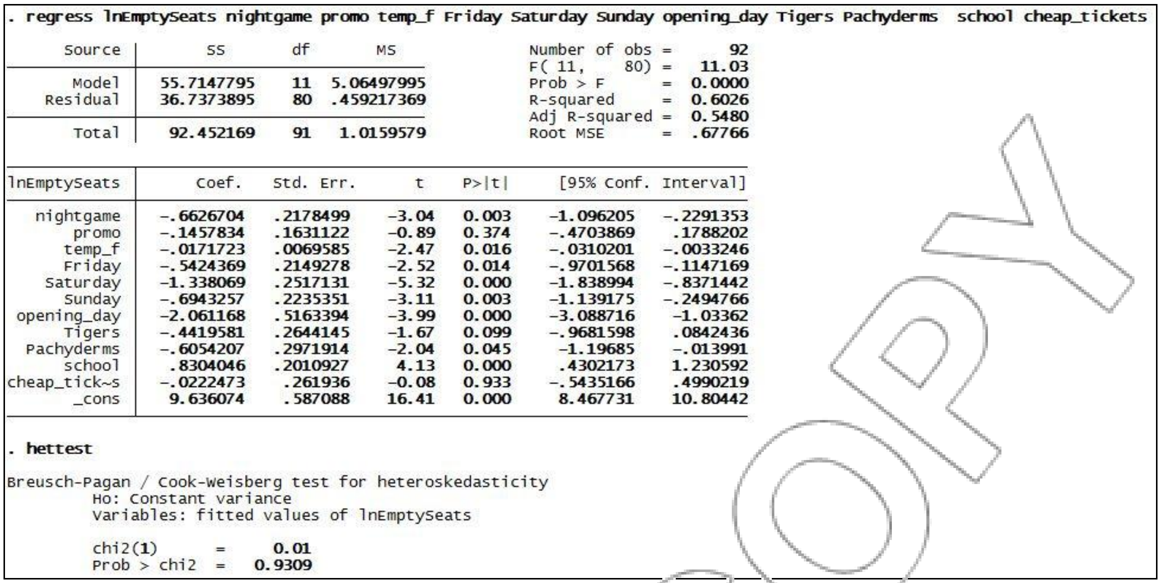


Figure 4

8. How does the semi-log model of empty seats used in Regression 7 compare to the models used in Regressions 5 and 6? Briefly discuss the pros and cons of using this last model.

9. Use Regression 7 to predict attendance for a Saturday night game against the Tigers that is not opening day. Also, the temperature is 89 degrees, there are full-price tickets, a promotional giveaway, and school is out of session. In addition to a single attendance number, provide a 95% prediction interval for your answer.



```
. regress lnEmptySeats nightgame promo temp_f Friday Saturday Sunday opening_day Tigers Pachyderms  school cheap_tickets

      Source |       SS       df       MS              Number of obs =      92
-------------+------------------------------           F( 11,    80) =   11.03
       Model |  55.7147795    11  5.06497995           Prob > F      =  0.0000
    Residual |  36.7373895    80  .459217369           R-squared     =  0.6026
-------------+------------------------------           Adj R-squared =  0.5480
       Total |   92.452169    91   1.0159579           Root MSE      =  .67766

------------------------------------------------------------------------------
lnEmptySeats |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   nightgame |  -.6626704   .2178499    -3.04   0.003    -1.096205   -.2291353
       promo |  -.1457834   .1631122    -0.89   0.374    -.4703869    .1788202
      temp_f |  -.0171723   .0069585    -2.47   0.016    -.0310201   -.0033246
      Friday |  -.5424369   .2149278    -2.52   0.014    -.9701568   -.1147169
    Saturday |  -1.338069   .2517131    -5.32   0.000    -1.838994   -.8371442
      Sunday |  -.6943257   .2235351    -3.11   0.003    -1.139175   -.2494766
 opening_day |  -2.061168   .5163394    -3.99   0.000    -3.088716    -1.03362
      Tigers |  -.4419581   .2644145    -1.67   0.099    -.9681598    .0842436
  Pachyderms |  -.6054207   .2971914    -2.04   0.045     -1.19685    -.013991
      school |   .8304046   .2010927     4.13   0.000     .4302173    1.230592
cheap_tick~s |  -.0222473    .261936    -0.08   0.933    -.5435166    .4990219
       _cons |   9.636074    .587088    16.41   0.000     8.467731    10.80442
------------------------------------------------------------------------------

. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnEmptySeats

         chi2(1)      =     0.01
         Prob > chi2  =   0.9309
```

Regression 7

APPENDIX: A STATA MINI-MANUAL

This Stata mini-manual is a complement, not a substitute, for the other resources available for you in learning Stata. Do not worry if some of the terminology used in this manual is unfamiliar. The purpose here is to instruct you on the mechanics of using Stata, not in understanding the statistics: this is what the text is all about!

GETTING STARTED WITH STATA

Loading the Core Statistics Custom Menu

Stata is statistical software that enables you to do easily many of the statistical calculations required for this course. It is quite a powerful and flexible program, and is likely to meet your statistics needs not only throughout your education, but also throughout your career. To start Stata, you can either:

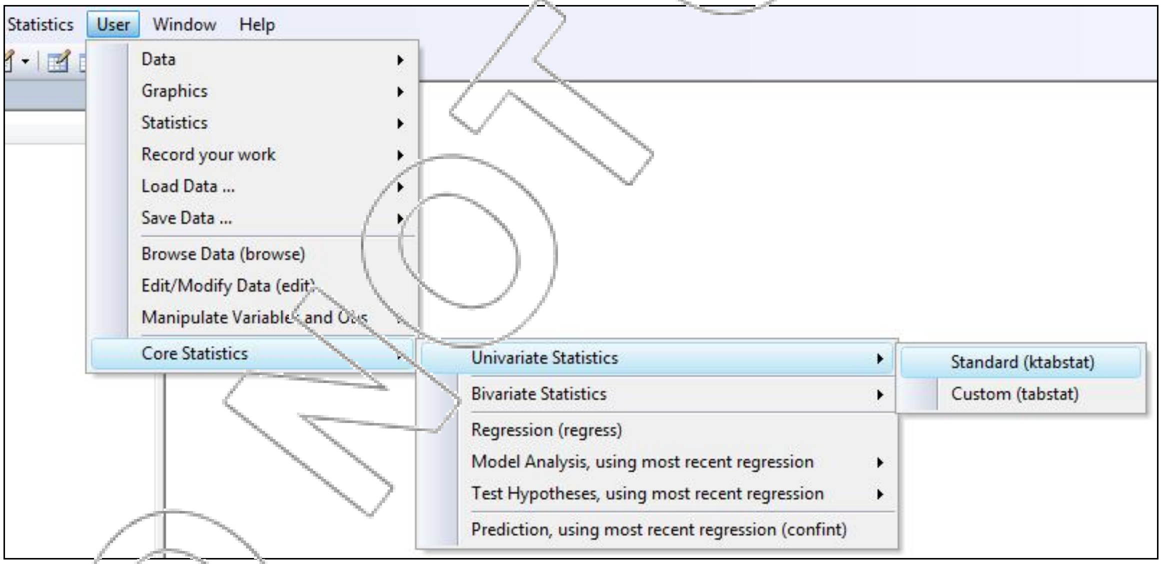
1. Double-click to open a Stata **.do** file (of commands), or
2. Double-click to open a Stata **.dta** file (of data), or
3. Double-click (or otherwise start) the Stata executable.

In this text, we will be making extensive use of the custom Core Statistics menu and will assume throughout the text that you have loaded it into Stata. To load this menu, follow these steps: Run Stata. There will be a command line. Type the command

do http://kellogg.northwestern.edu/stata/menu.do and hit enter (you will need an internet connection). In the dialog box that appears, check “Core Statistics” only. After the dialog box executes, the custom menu will be installed under the **User Menu** in Stata. It only needs to be installed once, and will appear there each time you start Stata.

Using Menus, Dialog Boxes and Typed Commands

Throughout this manual, commands on the main menu and sub-menus will be separated by the > sign. For example, clicking **User>Core Statistics>Univariate Statistics>Standard (ktabstat)** means doing this:



You can click on **User**, then **Core Statistics**, then **Univariate Statistics**, then **Standard (ktabstat)** (once each), or click on **User**, hold the mouse button down as the sub-menus pop up, and release the button when you have gotten to **Standard (ktabstat)**.

You can also open most Stata command dialog boxes by typing **db *dialogboxname*** in the Command box. For example, typing **db ktabstat** will open the **ktabstat** dialog box.

A third and commonly used alternative for carrying out Stata commands is to type commands directly into the Command box. This method is most efficient for frequently used commands that have few options (e.g., running a regression). For more complicated tasks, such as generating a graph with customized title, legends, scales, etc., it is generally easier to use the dialog box instead. Note that whenever you use a dialog box to run a command, Stata will display the corresponding direct command at the top of the output. When this text lists a direct command (such as **regress *depvar indepvars***), the italicized portion refers to the following:

***depvar* -** the name of a dependent variable

***indepvar(s)* -** the name(s) of (an) independent variable(s)

***newvar* -** the name that you want to give to a new variable

***oldvar* -** the name of an existing variable

***varlist* -** a list of variable names separated by spaces

***varname* -** the name of a variable

***varX* -** the name of an X variable

***varY* -** the name of a Y variable

Logging Your Work

It is generally a good idea to record the work that you have done in Stata so you can refer to it in the future if necessary. You can use Stata's **log** command to store all of your commands and outputs in a plain text file. To start logging your work, click **User>Record your work>Open Log (log using)** or type **db log**. (Stata's native menu option is **File>Log>Begin….**) Type the name that you want for your new log file, select **Log (*.log)** as the file type, and click **Save**. After you have started a log file, all output in Stata's Results window will be recorded.

If you want to record your work using an existing log file, you can open the **log** dialog box, double-click on the desired file, and select "Append to existing file" in the ensuing Stata Log Options window.

Alternatively, you can type the direct command **log using *newfilename*.log** to create a new log file. Stata will store this file in the default data folder unless you specify the directory in which you want to save your log file (in this case the direct command would be **log using directory\*newfilename*.log**). The direct command for appending to an existing log file is **log using *filename*.log, append**. However, it is generally easier to open or create a log file by using the **log** dialog box instead.

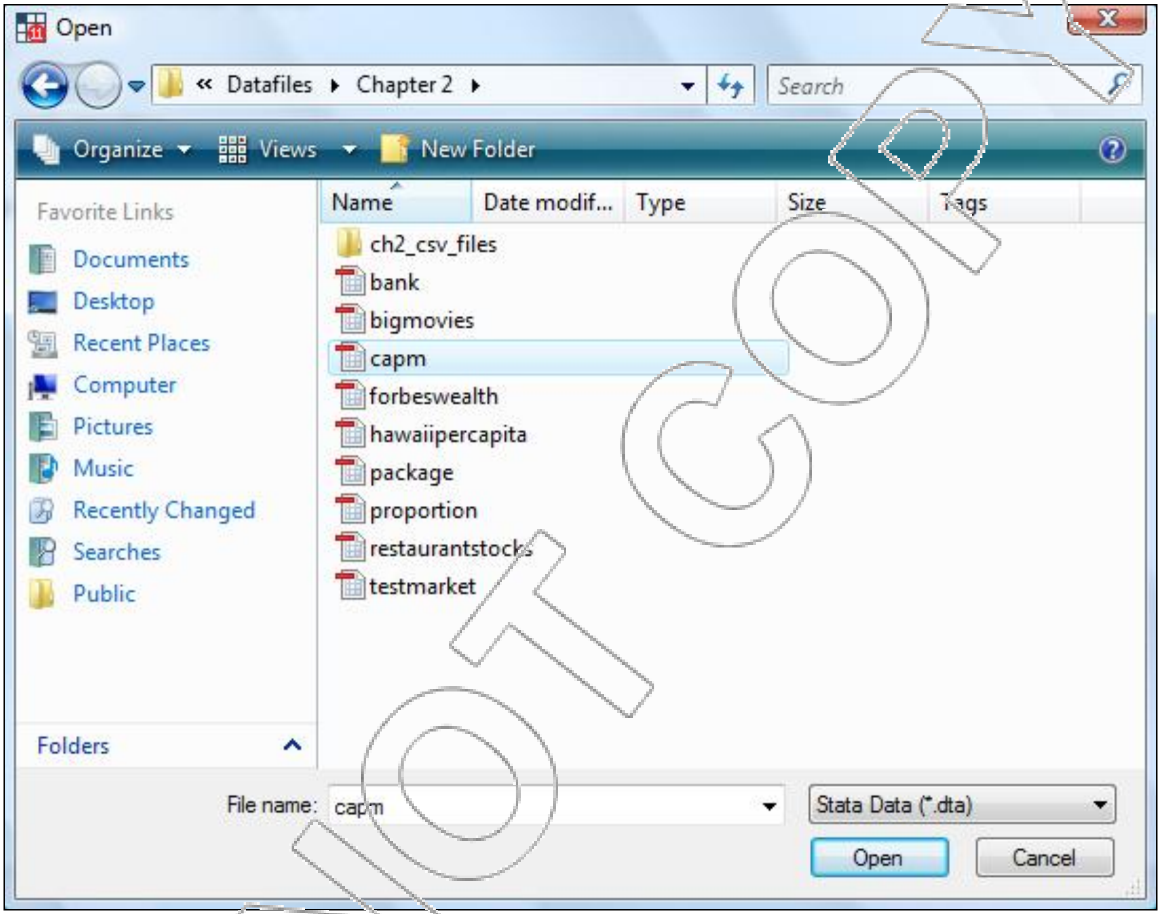
To stop logging your work, you can click **User>Record your work>Close Log (log close)** or **File>Log>Close**. You may also type the direct command **log close**. Any open log will be closed automatically when you exit Stata.

Opening/Starting a Data File

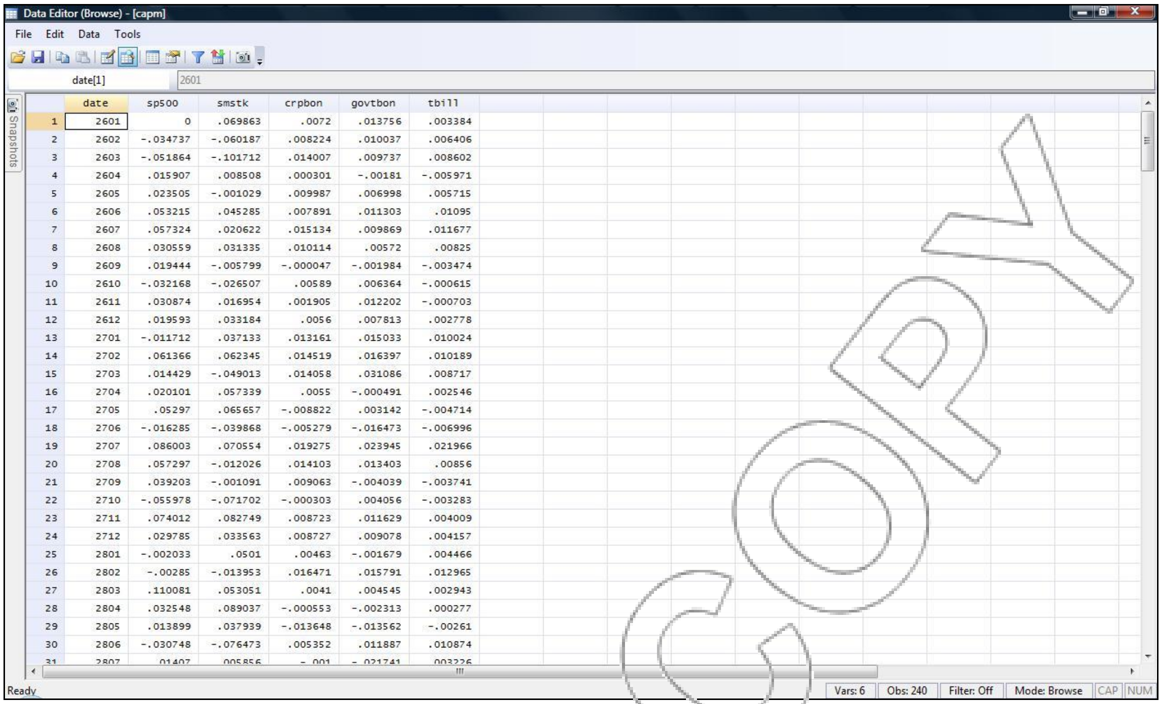
When Stata starts, it will have an empty data sheet in the Data Editor. This is where you enter all the data that you wish to analyze. Usually, you will want to load a data file into Stata. To do this, click **User>Load Data…>Stata Dataset (use)** or type **db use**.[1] You will see a window like the one below. Choose the folder that your data file is in, choose the data file and click **Open**. For

[1] Alternatively, you can click **File>Open…**

example, in the following window, you can import the **capm.dta** dataset into Stata by clicking **Open**.



Once your data are in place, the **Data Browser** (or **Data Editor**) should look like this:



date[1] 2601

	date	sp500	smstk	crpbon	govtbon	tbill
1	2601	0	.069863	.0072	.013756	.003384
2	2602	-.034737	-.060187	.008224	.010037	.006406
3	2603	-.051864	-.101712	.014007	.009737	.008602
4	2604	.015907	.008508	.000301	-.00181	-.005971
5	2605	.023505	-.001029	.009987	.006998	.005715
6	2606	.053215	.045285	.007891	.011303	.01095
7	2607	.057324	.020622	.015134	.009869	.011677
8	2608	.030559	.031335	.010114	.00572	.00825
9	2609	.019444	-.005799	-.000047	-.001984	-.003474
10	2610	-.032168	-.026507	.00589	.006364	-.000615
11	2611	.030874	.016954	.001905	.012202	-.000703
12	2612	.019593	.033184	.0056	.007813	.002778
13	2701	-.011712	.037133	.013161	.015033	.010024
14	2702	.061366	.062345	.014519	.016397	.010189
15	2703	.014429	-.049013	.014058	.031086	.008717
16	2704	.020101	.057339	.0055	-.000491	.002546
17	2705	.05297	.065657	-.008822	.003142	-.004714
18	2706	-.016285	-.039868	-.005279	-.016473	-.006996
19	2707	.086003	.070554	.019275	.023945	.021966
20	2708	.057297	-.012026	.014103	.013403	.00856
21	2709	.039203	-.001091	.009063	-.004039	-.003741
22	2710	-.055978	-.071702	-.000303	.004056	-.003283
23	2711	.074012	.082749	.008723	.011629	.004009
24	2712	.029785	.033563	.008727	.009078	.004157
25	2801	-.002033	.0501	.00463	-.001679	.004466
26	2802	-.00285	-.013953	.016471	.015791	.012965
27	2803	.110081	.053051	.0041	.004545	.002943
28	2804	.032548	.089037	-.000553	-.002313	.000277
29	2805	.013899	.037939	-.013648	-.013562	-.00261
30	2806	-.030748	-.076473	.005352	.011887	.010874
31	2807	.01407	.005856	-.001	-.021741	.003226

Ready Vars: 6 Obs: 240 Filter: Off Mode: Browse CAP NUM

There are other ways to input data into Stata. In a blank Data Editor, you may copy and paste or type in data manually. Often, you may have data already entered in a spreadsheet that you want to import into Stata. To import data from an Excel spreadsheet, for example, you can do one of the following:

1. Directly copy and paste the entire dataset from your Excel spreadsheet into Stata's Data Editor. Before copying the data, you should first format your spreadsheet so that the first row contains variable names. When you paste your data into the Data Editor, click "Treat first row as variable names" in the Paste Clipboard Data prompt. Click **File>Save** in the Data Editor or click **User>Save Data…>Stata Dataset (save)**[2] in the Stata main window to save your dataset as a **.dta** file.
2. Save your Excel spreadsheet as a comma separated file by clicking **File>Save As…** in Excel. Select **CSV (Comma delimited)** as your file type and click **Save**. Next, open Stata and click **User>Load Data…>ASCII (text) data created by a**

[2] Alternatively, you may type **db save**.

spreadsheet (insheet) (or type **db insheet**).[3] Select **Comma Separated Values (*.csv)** from the file type drop-down list. Browse for your file, choose **Comma-delimited data** from the "Delimiter" field, check the box next to "Preserve variable case" and click **OK**. Open the Data Browser to verify that your data has been imported correctly.

A few things to keep in mind when you are converting and importing a **.csv** file:

1. You need to format an Excel spreadsheet properly before saving it as a comma delimited file. The first row in your spreadsheet should contain variable names, and there can be no empty rows or columns within your data. Your dataset should not contain non-numeric symbols such as commas and the dollar sign. When you have missing data, you should leave the appropriate cell(s) blank instead of entering placeholders such as N/A.
2. Stata does not allow space(s) within a variable name. For example, a variable with the name **Avg Temp** in Excel will be imported as **AvgTemp** into Stata.
3. Stata stores the names of all imported variables in lowercase unless you check the "Preserve variable case" box in the **insheet** dialog box.
4. If you choose **Use default** as your "Storage type" in the **insheet** dialog box, Stata will store any variable that contains decimal values as a float variable. Because a float variable has about 7 digits of accuracy, and because Stata may store a value of 5.6 as 5.5999999, you may encounter rounding discrepancies as you work with datasets converted using the default float storage type. One solution to this problem is to select the **Force double** storage type when importing a **.csv** file. This option keeps variables with decimal values accurate up to 16 digits.

[3] Stata's native menu option is **File>Import> ASCII data created by a spreadsheet**.

Exporting a Data File

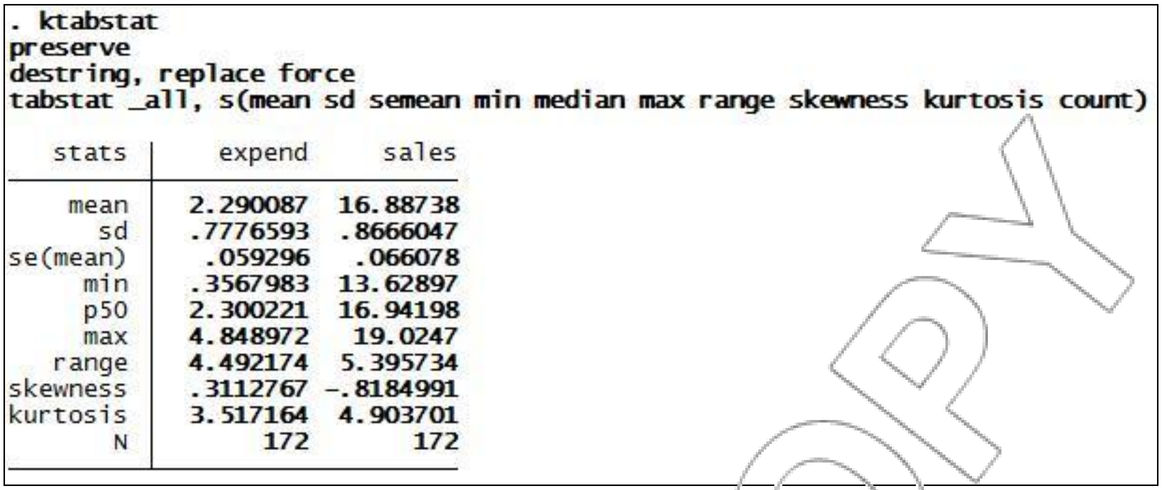
Sometimes you may need to export a datasheet from Stata to another spreadsheet program such as Excel. To do so, you can use one of the following methods:

1. Open your **.dta** file in Stata. Go to the Data Editor and select the entire dataset. Copy and paste the dataset into Excel and save the spreadsheet in a desired file format such as **.xls** or **.csv**.
2. Open your **.dta** file in Stata. Click **User>Save Data…>ASCII (text) data readable by a spreadsheet (outsheet)** or type **db outsheet**.[4] Click on the **Save As…** button to specify the name and location for your data file and choose **Comma Separated Values (*.csv)** as the file type. Select **Comma-separated (instead of tab-separated) format** in the "Delimiter" field and click **OK**. You can open the new **.csv** file in Excel to verify that your dataset has been exported correctly.

Basic Statistics and Critical Values

With Stata, you can easily obtain some basic statistical quantities. As an example, open the **adsales.dta** data file. Click **User>Core Statistics>Univariate Statistics>Standard (ktabstat)** (or type **db ktabstat**) to generate useful summary statistics for each variable in the file.[5] The output looks like that in Figure A.1.

[4] Stata's native menu path is **File>Export>Comma- or tab-separated data**.
[5] Alternatively, you may directly type the command **ktabstat**.



```
. ktabstat
preserve
destring, replace force
tabstat _all, s(mean sd semean min median max range skewness kurtosis count)

   stats |    expend     sales
---------+--------------------
    mean |  2.290087  16.88738
      sd |  .7776593  .8666047
se(mean) |   .059296   .066078
     min |  .3567983  13.62897
     p50 |  2.300221  16.94198
     max |  4.848972   19.0247
   range |  4.492174  5.395734
skewness |  .3112767 -.8184991
kurtosis |  3.517164  4.903701
       N |       172       172
------------------------------
```

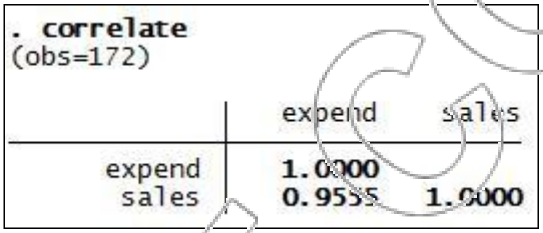
Figure A.1: Univariate statistics for the adsales.dta data.

If you want Stata to calculate statistics other than the ones included in the **ktabstat** command, or if you want Stata to display basic statistics only for specific variables in your dataset, you can click **User>Core Statistics>Univariate Statistics>Custom (tabstat)** or type **db tabstat**) instead.[6] This command allows you to select up to eight statistics that you want Stata to display for your specified variable(s). The direct command is **tabstat *varlist*, s(…)**, where you can specify the names of summary statistics in the **s(…)** portion of the command. For the complete list of summary statistics, type **help tabstat** into the Stata Command box and refer to the Options>statistics section. Note that the **tabstat** command will not work for string, or non-numeric, variables. Therefore, if there is any string variable present in your dataset, it is generally easier to use the **ktabstat** command instead, as it is programmed to convert string variables to numeric variables temporarily prior to calculating summary statistics. Your original dataset will not be affected by this temporary conversion.

To find the correlation coefficients between all pairs of variables in your dataset, you can click **User>Core Statistics>Bivariate Statistics>Correlations (correlate)** (or type **db correlate**),

[6] Stata's native menu option is **Statistics>Summaries, tables, and tests>Tables>Table of summary statistics (tabstat)**.

leave the “Variables” field empty, and click **OK**. Stata’s native menu option is **Statistics>Summaries, tables, and tests>Summary and descriptive statistics>Correlations and covariances**, and the direct command is **correlate**. If there are some non-numeric variables in your data, **correlate** will return an error message. If you want Stata to compute correlation coefficients for selected variables (e.g., the non-numeric ones) only, you can specify those variables in the **correlate** dialog box.[7] Again, using the **adsales.dta** data, we produce the output in Figure A.2.



```
. correlate
(obs=172)

             |   expend    sales
-------------+------------------
      expend |   1.0000
       sales |   0.9555   1.0000
```

Figure A.2: Correlations for adsales.dta data.

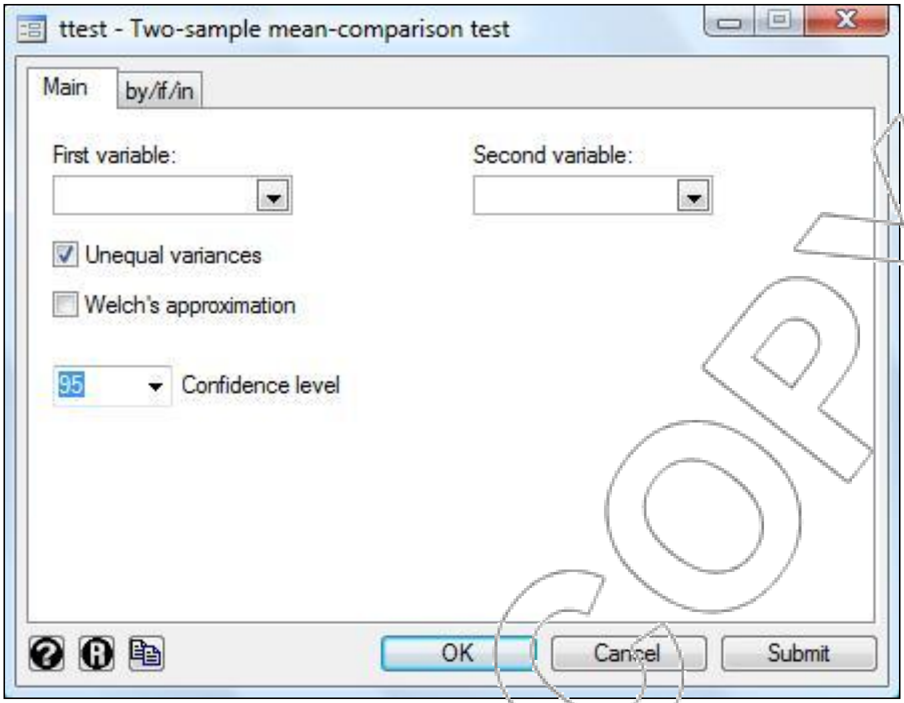
Here, 0.9555 is the correlation between expend and sales.

To perform a **1-Sample t-test** in Stata, you can click **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>One-sample mean-comparison test** (see Chapter 2).[8] To compare the means of two populations using a **2-Sample t-test**, click **Statistics>Summaries, tables, and tests>Classical tests of hypotheses>Two-sample mean-comparison test** (see Chapter 2).[9] We will usually assume the variances of the variables in a 2-sample t-test are different so you will check the box next to “Unequal variances.” The dialog box for a 2-sample t-test looks like this:

[7] The direct command is **correlate *varlist***.

[8] The direct command is **ttest *varname* == #**.

[9] The direct command is **ttest *varname1* == *varname2*, unpaired unequal**.



First variable:

Second variable:

☑ Unequal variances

□ Welch's approximation

95 Confidence level

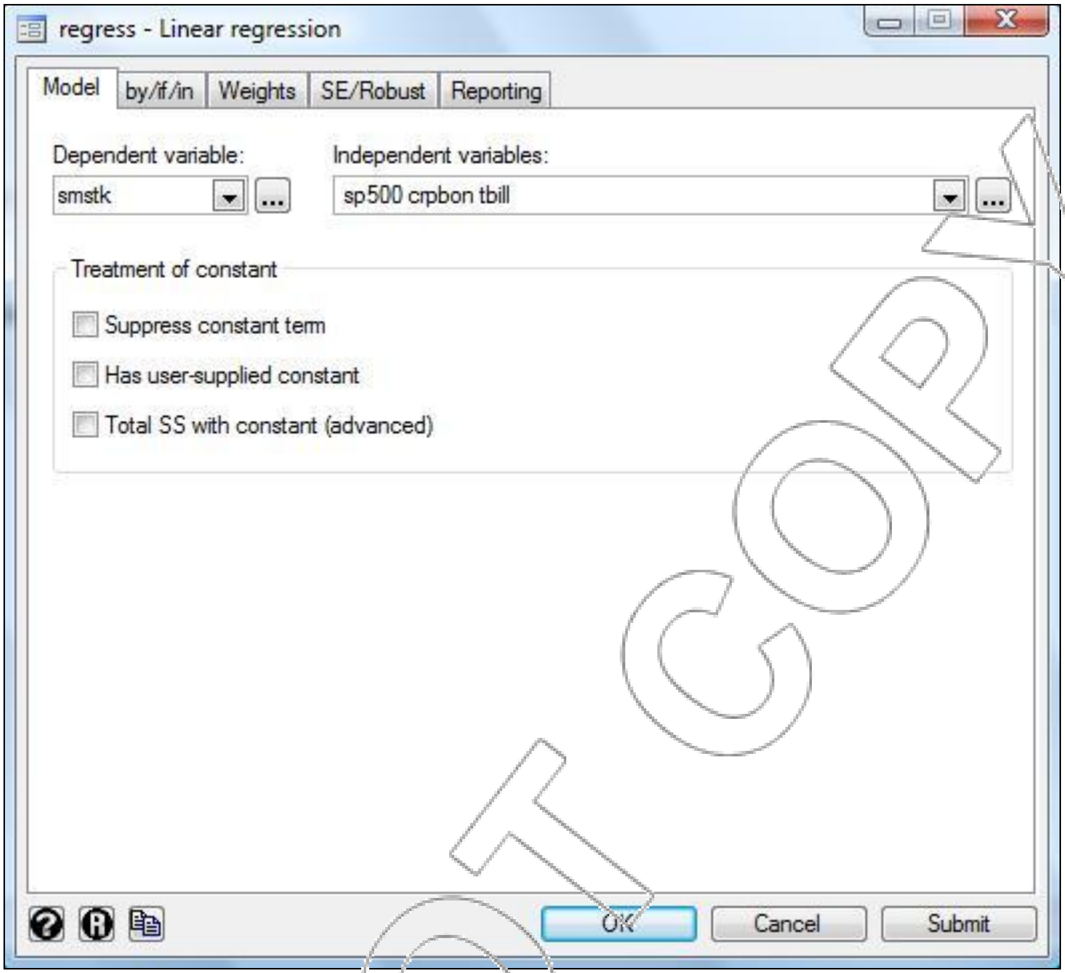
Specify your variables and click **OK**, and Stata will return the test statistic as well as the p-values corresponding to the alternative hypotheses that the difference in population means is less than, not equal to, or greater than 0.

Regression

In this section, we will use the **capm.dta** data.

The command you will probably use most frequently is the **regress** command. You can access the dialog box for this command by clicking **User>Core Statistics>Regression (regress)** (or type **db regress**).[10] Stata's native menu option is **Statistics>Linear models and related>Linear regression**. Clicking on the menus will open the following dialog box:

[10] The direct command is **regress** ***depvar indepvar(s)***.



Dependent variable: smstk

Independent variables: sp500 crpbon tbill

Treatment of constant

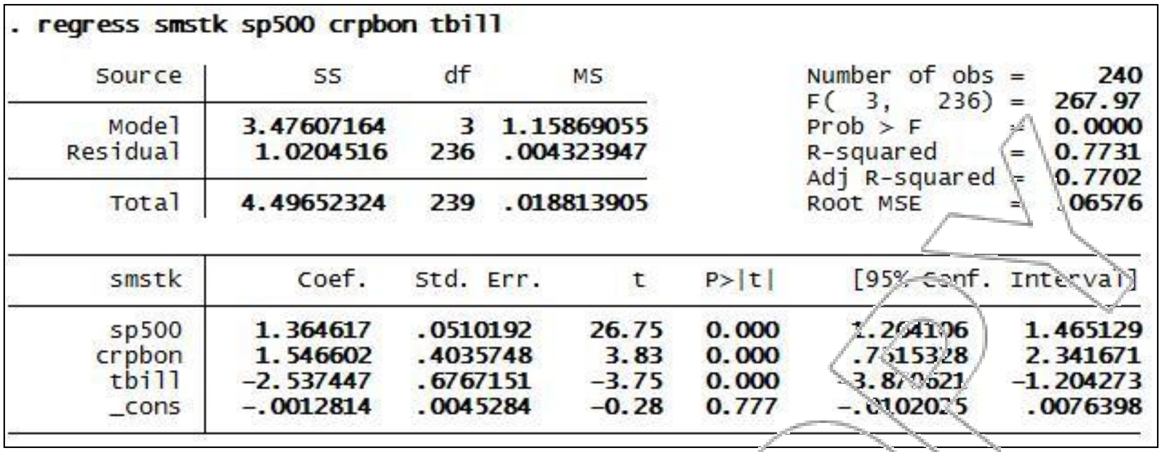
☐ Suppress constant term

☐ Has user-supplied constant

☐ Total SS with constant (advanced)

In this example, we will choose **smstk** as our dependent variable and **sp500**, **crpbon**, and **tbill** as our independent variables.

When you click **OK**, Stata will display the regression output as in Figure A.3.



```
. regress smstk sp500 crpbon tbill

      Source |       SS       df       MS              Number of obs =     240
-------------+------------------------------           F(  3,   236) =  267.97
       Model |  3.47607164     3  1.15869055           Prob > F      =  0.0000
    Residual |   1.0204516   236  .004323947           R-squared     =  0.7731
-------------+------------------------------           Adj R-squared =  0.7702
       Total |  4.49652324   239  .018813905           Root MSE      =  .06576

------------------------------------------------------------------------------
       smstk |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       sp500 |   1.364617   .0510192    26.75   0.000     1.264106    1.465129
      crpbon |   1.546602   .4035748     3.83   0.000     .7515328    2.341671
       tbill |  -2.537447   .6767151    -3.75   0.000    -3.870621   -1.204273
       _cons |  -.0012814   .0045284    -0.28   0.777    -.0102025    .0076398
------------------------------------------------------------------------------
```

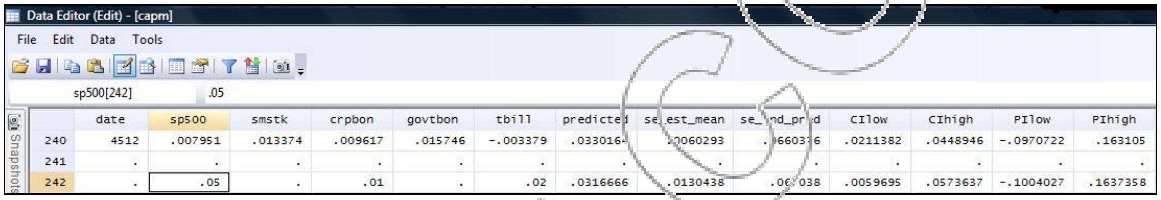
Figure A.3: Regression of smstk on sp500, crpbon, and tbill.

From the output above, we can see that our regression equation is.

smstk = -0.0012814+1.364617*sp500+1.546602*crpbon-2.537447*tbill. Stata lists the standard errors, t-ratios, p-values, and 95 % confidence intervals for each coefficient in the **Std. Err.**, **t**, **P>|t|**, and **95% Conf. Interval** columns, respectively. Under the **SS** column, you can find explained sum of squares, residual sum of squares, and total sum of squares in rows **Model**, **Residual**, and **Total**, respectively. The degrees of freedom of the error term is listed in the **Residual** row and the **df** column. The number of observations, F-ratio, p-value (Prob > F), R^2, adjusted R^2, and the standard error of the regression (Root MSE) are listed in the top right corner. The p-value (Prob > F) listed just above the R^2 in the regression output is for the hypothesis test with the null hypothesis that the coefficients for all the variables are equal to zero. The p-value of zero says we can reject the null hypothesis with high confidence, and thus have strong evidence that at least one of the independent variables is related to the dependent variable.

To have Stata calculate the beta-weights for each coefficient, you can click the **Reporting** tab in the **regress** dialog box and check the box next to “Standardized beta coefficients.” You can alternatively type the direct command **regress *depvar indepvars*, beta**.

To make predictions using your most recently performed regression, first open the Data Editor. Suppose we want the predicted value for **smstk** where **sp500** = 0.05, **crpbon** = 0.01 and **tbill** = 0.02. Enter these numbers into an empty row in the cells corresponding to each variable (we leave a blank row above our entry to remind ourselves where the original data ends; in this case we will enter our new set of values in row 242). Minimize or exit the Data Editor. Next, click **User>Core Statistics>Prediction, using most recent regression** or type **db confint**. Click **OK**, and you will obtain the following output:



sp500[242] .05

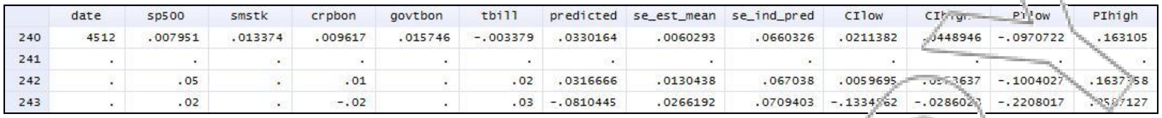
	date	sp500	smstk	crpbon	govtbon	tbill	predicted	se_est_mean	se_ind_pred	CIlow	CIhigh	PIlow	PIhigh
240	4512	.007951	.013374	.009617	.015746	-.003379	.0330164	.0060293	.0660336	.0211382	.0448946	-.0970722	.163105
241	.	.	.	.	.	.	.	.	.	.	.	.	.
242	.	.05	.	.01	.	.02	.0316666	.0130438	.067038	.0059695	.0573637	-.1004027	.1637358

As you can see, Stata has generated new variables corresponding to fitted or predicted values (**predicted**), the standard error of the estimated mean (**se_est_mean**), the standard error of prediction (**se_ind_pred**), as well as 95% confidence and prediction intervals (**CIlow/CIhigh** and **PIlow/PIhigh**, respectively).

To change the confidence level for these intervals, open the **confint** dialog box again and type the confidence level you want in the "Confidence level in %" field. Click **OK**, and Stata will regenerate the variables listed in the previous paragraph using the new confidence level.

To do predictions for more than one set of values, simply enter each set of values for the independent variables in a separate row in the Data Editor. Suppose we want to make predictions for **sp500** = 0.05, **crpbon** = 0.01, and **tbill** = 0.02, as well as **sp500** = 0.02, **crpbon** = -0.02, and **tbill** = 0.03. After you have entered these values in the Data Editor and clicked **User>Core**

Statistics>Prediction, using most recent regression (confint), the Data Browser should look like this:



	date	sp500	smstk	crpbon	govtbon	tbill	predicted	se_est_mean	se_ind_pred	CIlow	CIhigh	PIlow	PIhigh
240	4512	.007951	.013374	.009617	.015746	-.003379	.0330164	.0060293	.0660326	.0211382	.0448946	-.0970722	.163105
241	.	.	.	.	.	.	.	.	.	.	.	.	.
242	.	.05	.	.01	.	.02	.0316666	.0130438	.067038	.0059695	.0573637	-.1004027	.1637358
243	.	.02	.	-.02	.	.03	-.0810445	.0266192	.0709403	-.1334862	-.0286028	-.2208017	.0587127

If you want to generate only predicted values, only the standard error of the estimated mean, or only the standard error of prediction after running a regression, you can click **Statistics>Postestimation>Prediction, residuals, etc**. or type **db predict**. In the “New variable name” field, type in the name for which you want your predicted values or standard errors to be displayed as, and choose the appropriate variable from the “Produce” list:

a. To generate predicted values, choose “Linear prediction (xb).”
b. To generate the standard error of the estimated mean, choose “Standard error of the prediction.”
c. To generate the standard error of prediction, choose “Standard error of the forecast.”

The corresponding direct commands are:

a. **predict *newvar*, xb**
b. **predict *newvar*, stdp**
c. **predict *newvar*, stdf**

Note that Stata’s native **predict** command does not automatically generate the confidence and prediction intervals for fitted values. Therefore, it is generally more convenient to use the **prediction (confint)** command from the Core Statistics custom menu instead.

After performing a regression, you can use some other advanced options by clicking **User>Core Statistics>Model Analysis, using most recent regression**. This will expand the submenu from which you can select the respective commands that will calculate variance inflation factors for the coefficients, display the test statistic and p-value for the Breusch-Pagan heteroskedasticity test, plot residuals against predicted values, identify outliers and high leverage points, and calculate the Durbin-Watson d-statistic for detecting autocorrelation.[11] The corresponding native menu options in Stata and the direct commands for each of the options in the Model Analysis submenu are the following:

i) **Variance Inflation Factors (vif)** (or type **db vif**)

- Stata menu: **Statistics>Linear models and related>Regression diagnostics>Specification tests, etc.** (or type **db estat**) → Variance inflation factors for the independent variables (vif)
- Direct command: **vif**

ii) **Breusch-Pagan heteroskedasticity test (hettest)** (or type **db hettest**)

- Stata menu: **Statistics>Linear models and related>Regression diagnostics>Specification tests, etc.** (or type **db estat**) → Tests for heteroskedasticity (hettest)
- Direct command: **hettest**

iii) **Plot residuals vs predicted values (rvfplot)** (or type **db rvfplot**)

- Stata menu: **Graphics>Regression diagnostic plots>Residual-versus-fitted**
- Direct command: **rvfplot**

iv) **Residuals, outliers and influential observations (inflobs)** (or type **db inflobs**)

[11] The Jarque-Bera non-normality test is also included in the Model Analysis submenu, although we will not be using this command in this text.

- Stata menu: **Statistics>Postestimation>Predictions, residuals, etc.** (or type **db predict**) →You can have Stata generate residuals, studentized residuals, Cook's distance, and leverage using this dialog box.
- Direct command: **inflobs**

v) **Default Durbin-Watson Statistic (ddw)** (or type **db ddw**)

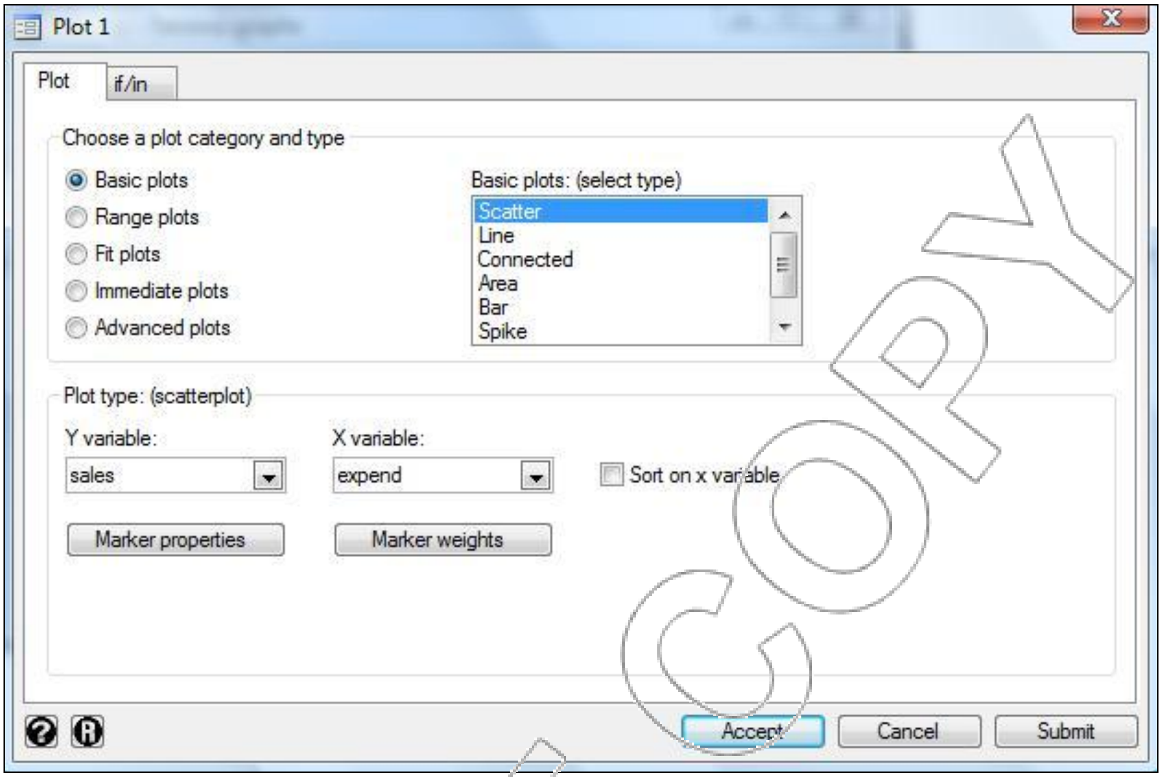
- Stata menu: **Statistics>Linear models and related>Regression diagnostics>Specification tests, etc.** (or type **db estat**) → Durbin-Watson d statistic (dwatson - time series only). Note that you need to declare a time index variable prior to using this command. See the **Other Stata Commands>Declaring a Time Index Variable** section for instruction on declaring time index variables.
- Direct command: **ddw**

Graphs

In this section, we will use the **adsales.dta** data. Load this file into Stata by clicking **User>Load Data…>Stata Data Set (use)** or type **db use**.

To plot one variable in your data against another, such as Y vs. X, click **User>Core Statistics>Bivariate Statistics>Bivariate Plots (twoway)** or type **db twoway**.[12] Click **Create…**, choose **Basic plots → Scatter**, and choose the corresponding variables from the Y/X variable drop-down lists. For example, to plot **sales** against **expend**, you should have a dialog box that looks like this:

[12] Stata's native menu option is **Graphics>Twoway graph (scatter, line, etc.)**.



Choose a plot category and type

- ● Basic plots
- ○ Range plots
- ○ Fit plots
- ○ Immediate plots
- ○ Advanced plots

Basic plots: (select type)

Scatter
Line
Connected
Area
Bar
Spike

Plot type: (scatterplot)

Y variable: sales

X variable: expend

□ Sort on x variable

Marker properties

Marker weights

If you want the regression line to appear on your graph, first click **Accept** to close the **Plot 1** dialog box. Next, click **Create…** again and select **Fit plots → Linear prediction**. Choose **sales** and **expend** as your Y and X variables, respectively, and click **Accept → OK**. You should obtain a scatterplot as well as the regression line of **sales** versus **expend** as shown in Figure A.5.

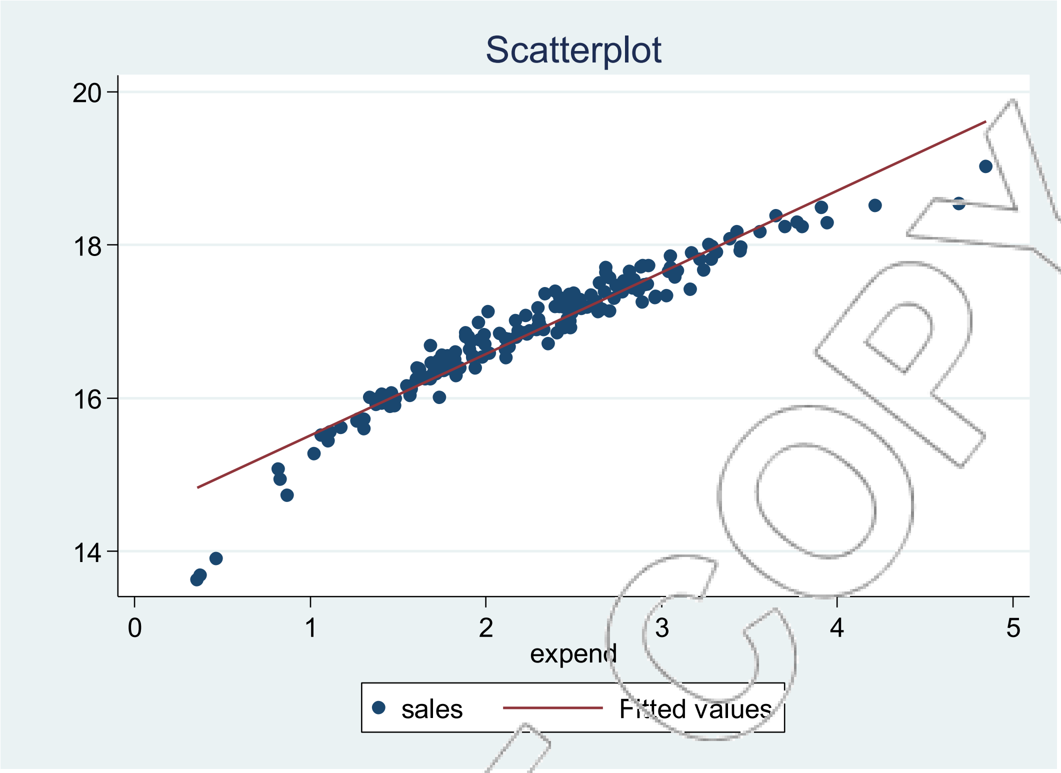
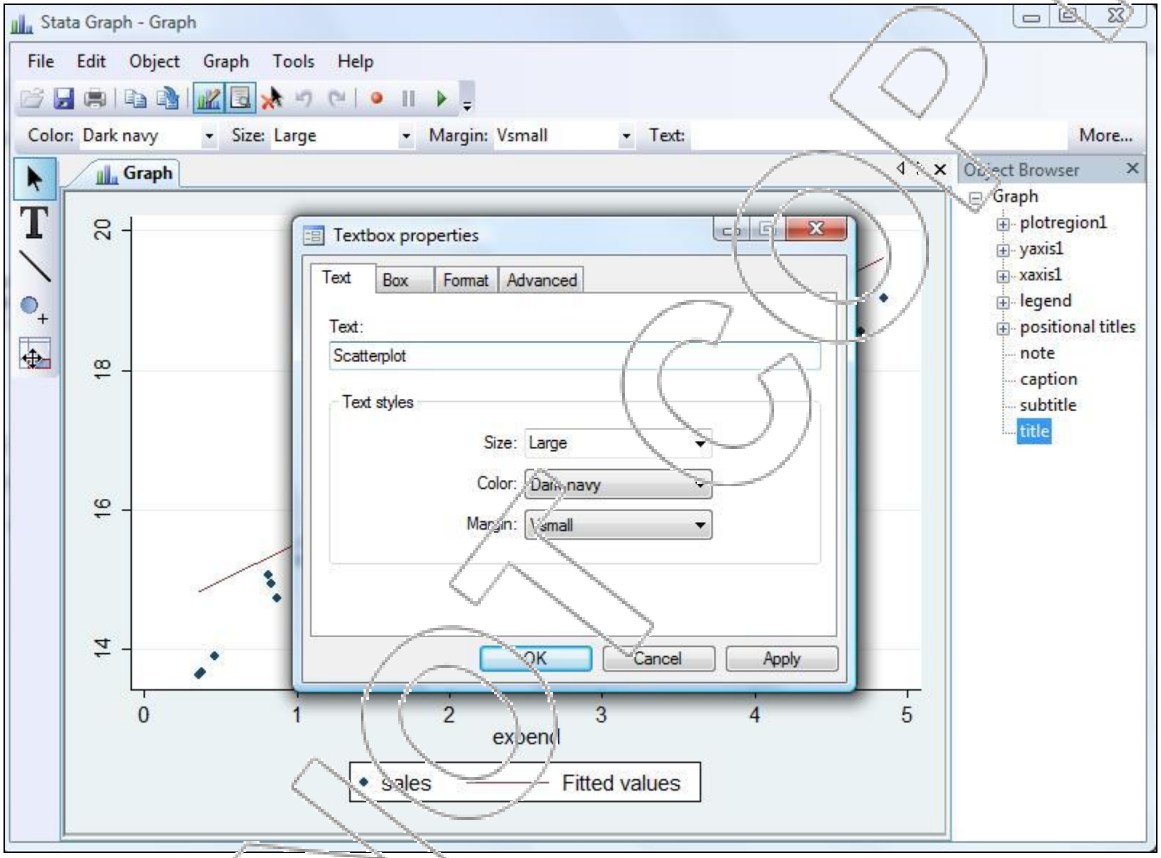


Figure A.5: Scatterplot of sales vs. expend.

To save this graph, you can click **File>Save** or right-click on the graph and select **Save As…**. Doing so saves your graph as a **.gph** file by default, which can be opened only in Stata. To insert a graph into a different file or program, you can right-click on the graph, select **Copy**, and paste that graph into the desired location.

You may have noticed that when you generated the scatterplot and regression line for **sales** versus **expend** by following the instructions above, your graph does not have a title or a y-axis label as shown in Figure A.5. You can easily add these elements as well as make various other adjustments to your graph by using Stata's Graph Editor. For example, to add the title "Scatterplot" to the graph in Figure A.5, click **File>Start Graph Editor** from the Stata Graph window.[13] In the Object Browser window, double click **title** under Graph>positional titles and type “Scatterplot” in the “Text” field. Click **OK**, and your scatterplot will now have an appropriate title:



Color: Dark navy Size: Large Margin: Vsmall Text:

Text: Scatterplot

Size: Large

Color: Dark navy

Margin: Vsmall

Similarly, double click **title** under Graph>yaxis1from the Object Browser and type “sales” to label the y-axis accordingly.

To edit the y-axis (x-axis), right-click on yaxis1 (xaxis1) from the Object Browser and select **Axis Properties**. You can adjust various aspects of the axes such as scaling, fonts, and label orientation.

[13] You can also right-click on the graph and select “Start Graph Editor.”

Note that instead of editing a graph after it has been generated, you can specify graph properties in advance via the optional tabs in the **twoway** dialog box. For more information on editing graphs, you can refer to Stata's accompanying manual or type **help graph editor** into the Command box.

In general, it is easier to use dialog boxes instead of direct commands to generate graphs in Stata because of the various graph options available. Nevertheless, you can use these following commands to generate common graphs:

- Scatterplot: **twoway scatter *varY varX***
- Connected graph: **twoway connected *varY varX***
- Graph of regression line: **twoway lfit *varY varX***
- Graphing regression line on top of a scatterplot: **twoway** (**scatter *varY varX***) (**lfit *varY varX***)

In evaluating a regression, the graph of residuals versus predicted (or fitted) values will often be useful. Here is how to generate such a graph for a regression of **expend** against **sales**. First, run a regression where **expend** is the dependent variable and **sales** is the independent variable. Then, click **User>Core Statistics>Model Analysis, using most recent regression>Plot residuals vs predicted values (rvfplot)** (or type **db rvfplot**).[14] Click **OK** in the ensuing dialog box, and you will obtain the graph shown in Figure A.6 of residuals against the fitted values.

[14] Stata's native menu option is **Graphics>Regression diagnostic plots>Residual-versus-fitted**.

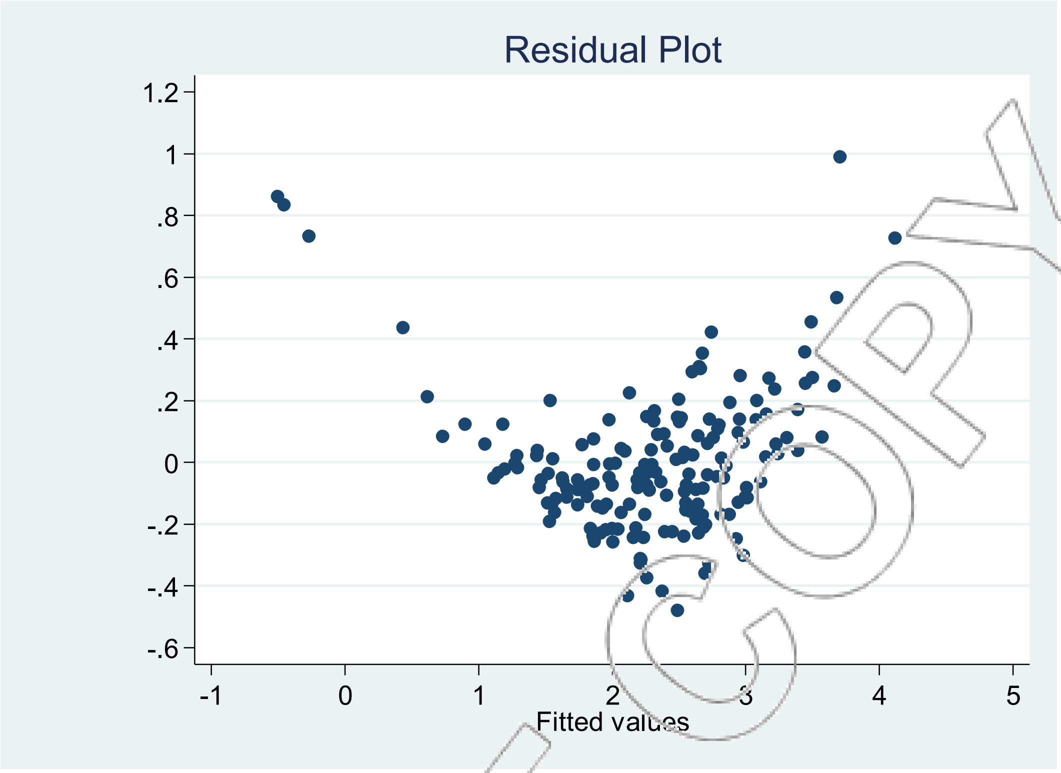
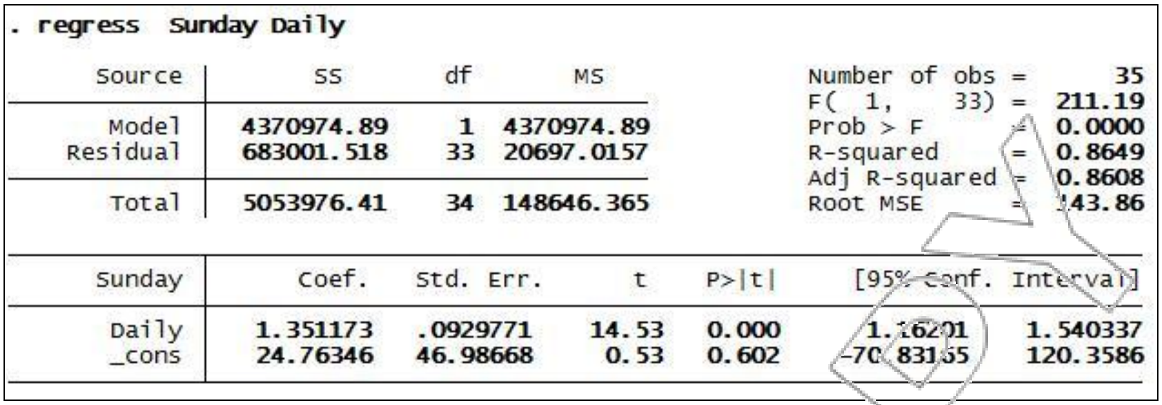


Figure A.6: Plot of residuals vs. predicted values from regression of expend on sales.

Alternatively, you may type the direct command **rvfplot** after running a regression.

Getting P-values

In this section, we will use the **newspapers.dta** data. A regression of **Sunday** against **Daily** generates the output in Figure A.7.



```
. regress  Sunday Daily

      Source |       SS       df       MS              Number of obs =      35
-------------+------------------------------           F(  1,    33) =  211.19
       Model |  4370974.89     1  4370974.89           Prob > F      =  0.0000
    Residual |  683001.518    33  20697.0157           R-squared     =  0.8649
-------------+------------------------------           Adj R-squared =  0.8608
       Total |  5053976.41    34  148646.365           Root MSE      =  143.86

------------------------------------------------------------------------------
      Sunday |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Daily |   1.351173   .0929771    14.53   0.000      1.16201    1.540337
       _cons |   24.76346   46.98668     0.53   0.602    -70.83155    120.3586
------------------------------------------------------------------------------
```

Figure A.7: Regression of Sunday on Daily.

The p-value of 0.000 corresponding to **Daily** in Figure A.7 is for one particular hypothesis test, where the null hypothesis is that β_1, the coefficient of **Daily**, is equal to zero. This p-value says we can reject the null with high confidence—we can be (virtually) 100% confident β_1 is not zero. If we wanted to test some other null hypothesis—for example, $\beta_1 = 1.1$—we would have to do the test manually. The t-statistic for this test is the following:

$$\frac{1.351173-1.1}{0.0929771} = 2.7015$$

Now we can use Stata's **ttail** function to look up the p-value corresponding to this value of t. The full syntax for this function is **display ttail(n, t)**, where Stata will compute the area to the right of **t** under a t-distribution with **n** degrees of freedom. In this example, **n** equals the residual degrees of freedom (=33), and **t** is our t-statistic (=2.7015). Since we are talking about the probability associated with a two-tailed test, we need to multiply the value ttail(33, 2.7015) by 2. Type **display 2*ttail(33, 2.7015)** into the Command box, and you should get the value 0.0108128. Thus, the p-value for the test is 0.0108128; that is, if the coefficient on **Daily** were 1.1, there would only be a 1.081% chance of obtaining a coefficient as far away from 1.1 as 1.351173 because of randomness in the data. We would reject the null hypothesis at any confidence level up to about 99% (or any significance level down to about 1%).

We can also use Stata's **invttail** function instead of a table to find critical values of t. The full syntax for this function is **display invttail(n, p)**, where Stata will calculate the value *x* for which the probability of falling to the right of that value is **p** under a t-distribution with **n** degrees of freedom. To find the t-statistic corresponding to $\alpha = .10$ for our two-tailed test, you can type **display invttail(33, 0.05)** into the Command box (remember that p = .10/2 = 0.05 since we are interested in a two-tailed test). The result tells us the t-statistic is 1.6923603. So, we would reject the null with $\alpha = .10$ if we obtained a t-statistic greater than 1.6923603 or less than -1.6923603 (which we did). This additionally tells us that for a one-sided test with a 'greater than' alternative, we would reject the null with $\alpha = .05$ if we obtained a t-statistic greater than 1.6923603, and for a one-sided test with a 'less-than' alternative, we would reject the null with $\alpha = .05$ if we obtained a t-statistic less than -1.6923603.

We can also use Stata's **normal(z)** function in place of a z-table. The full syntax for this function is **display normal(z)**, where Stata will calculate the area to the left of **z** under the standard normal distribution. Suppose we want to look up the p-value corresponding to a test statistic of z=2.7 for a one-sided test with a 'less-than' alternative. Type **display normal(2.7)** into the Command box, and you should get 0.99653303 (=P(Z<2.7)).

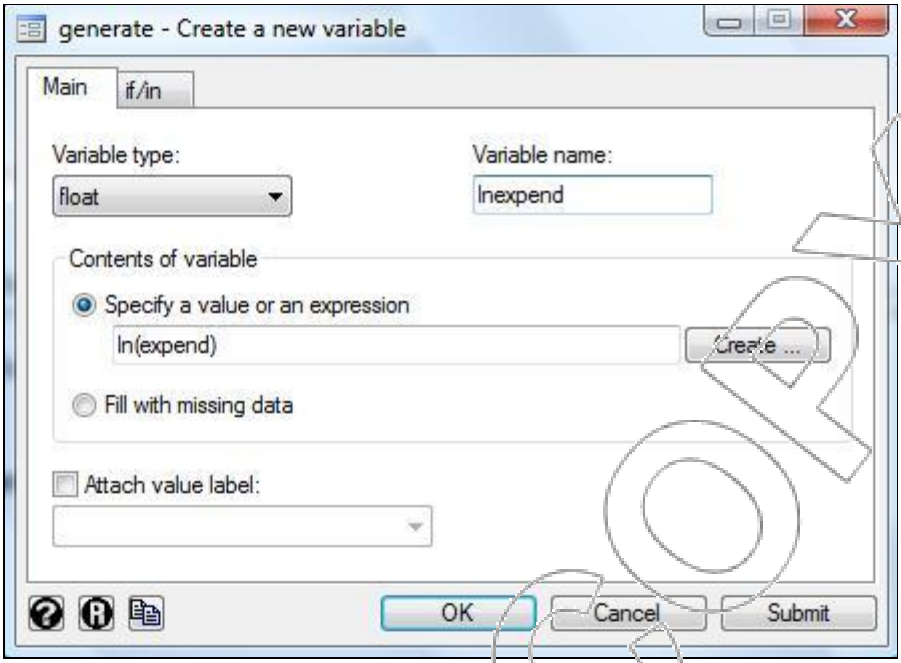
Suppose we wanted to find the z-statistic corresponding to $\alpha = .10$ for a two-tailed test. We can do this using Stata's **invnormal(p)** function. The full syntax for this function is **display invnormal(p)**, where Stata will return the value *x* for which the probability of falling to the left of that value under the standard normal distribution is **p**. For this example, we want the number *x* such that there is a 5% (i.e., α/2 %) chance of being greater than *x*, or, equivalently, a 95% (p=0.95) chance of being less than *x*. Type **display invnormal(0.95)** in the Command box, and the result tells us the appropriate z-statistic is 1.6448536.

Creating New Variables

Sometimes, you will need to make a new variable out of the ones given in a file. For example, you may want to use the logarithm of a variable as a predictor or response. As an example, create a new column, which includes the logarithm of the variable **expend**. To do this, first open the **adsales.dta** data. Next, click **User>Manipulate Variables and Obs>Generate New Variable (generate)** or type **db generate**.[15] Type the name you want to give to the new variable, say **lnexpend**, into the "New variable name" field. Type **ln(expend)** into the "Contents of new variable: Specify a value or an expression" field.[16] You should have a dialog box that looks like this:

[15] Stata's native menu option is **Data>Create or change data>Create new variable**.

[16] Alternatively, in the **generate** dialog box you may click **Create…** and select **Mathematical>ln()**. You need to type **expend** in place of **x** inside the ln() expression.



generate - Create a new variable

Variable type: float

Variable name: lnexpend

Contents of variable

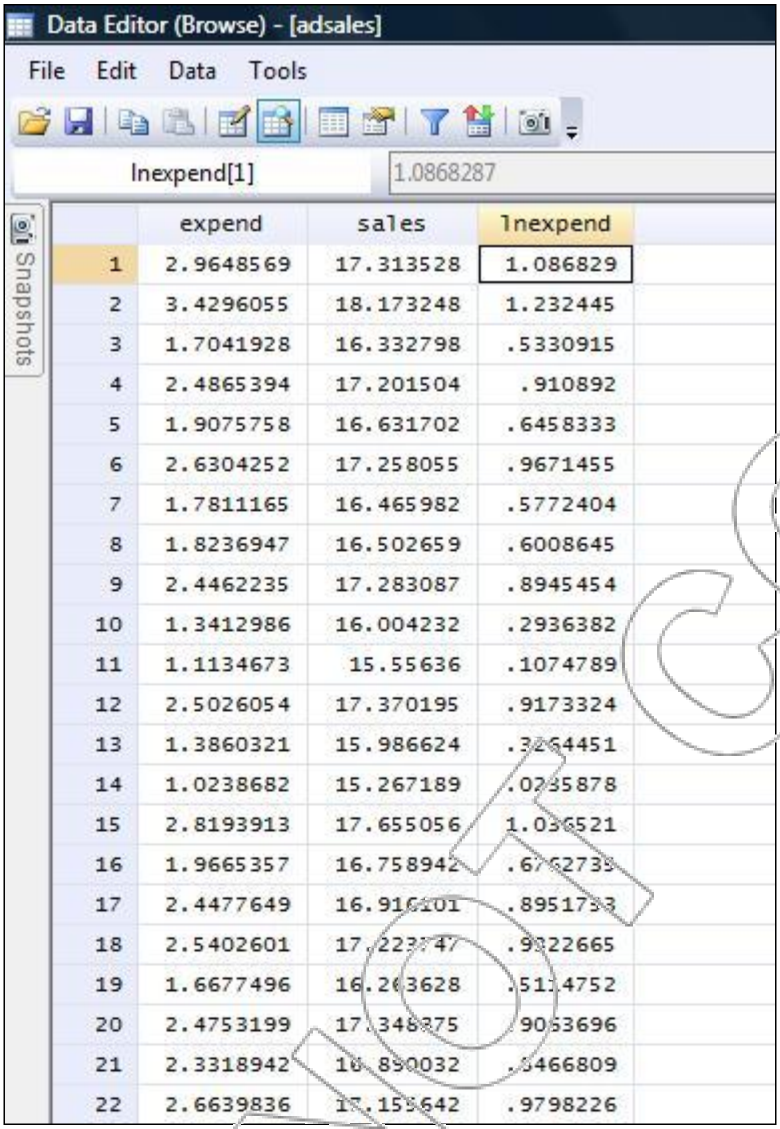
● Specify a value or an expression

ln(expend)

○ Fill with missing data

□ Attach value label:

Click **OK** and open the Data Browser. Your datasheet will look like this:



lnexpend[1] 1.0868287

	expend	sales	lnexpend
1	2.9648569	17.313528	1.086829
2	3.4296055	18.173248	1.232445
3	1.7041928	16.332798	.5330915
4	2.4865394	17.201504	.910892
5	1.9075758	16.631702	.6458333
6	2.6304252	17.258055	.9671455
7	1.7811165	16.465982	.5772404
8	1.8236947	16.502659	.6008645
9	2.4462235	17.283087	.8945454
10	1.3412986	16.004232	.2936382
11	1.1134673	15.55636	.1074789
12	2.5026054	17.370195	.9173324
13	1.3860321	15.986624	.3264451
14	1.0238682	15.267189	.0235878
15	2.8193913	17.655056	1.036521
16	1.9665357	16.758942	.6762735
17	2.4477649	16.916101	.8951733
18	2.5402601	17.223747	.9322665
19	1.6677496	16.263628	.5114752
20	2.4753199	17.348875	.9063696
21	2.3318942	16.890032	.8466809
22	2.6639836	17.155642	.9798226

Now we are done. We created a new variable called **lnexpend**. Each observation in **lnexpend** is the logarithm of the corresponding observation in **expend**.

Note that you can also open the **generate** dialog box within the Data Editor by clicking **Data>Create or change data>Create new variable**. You can see new variables generated live when using this method.

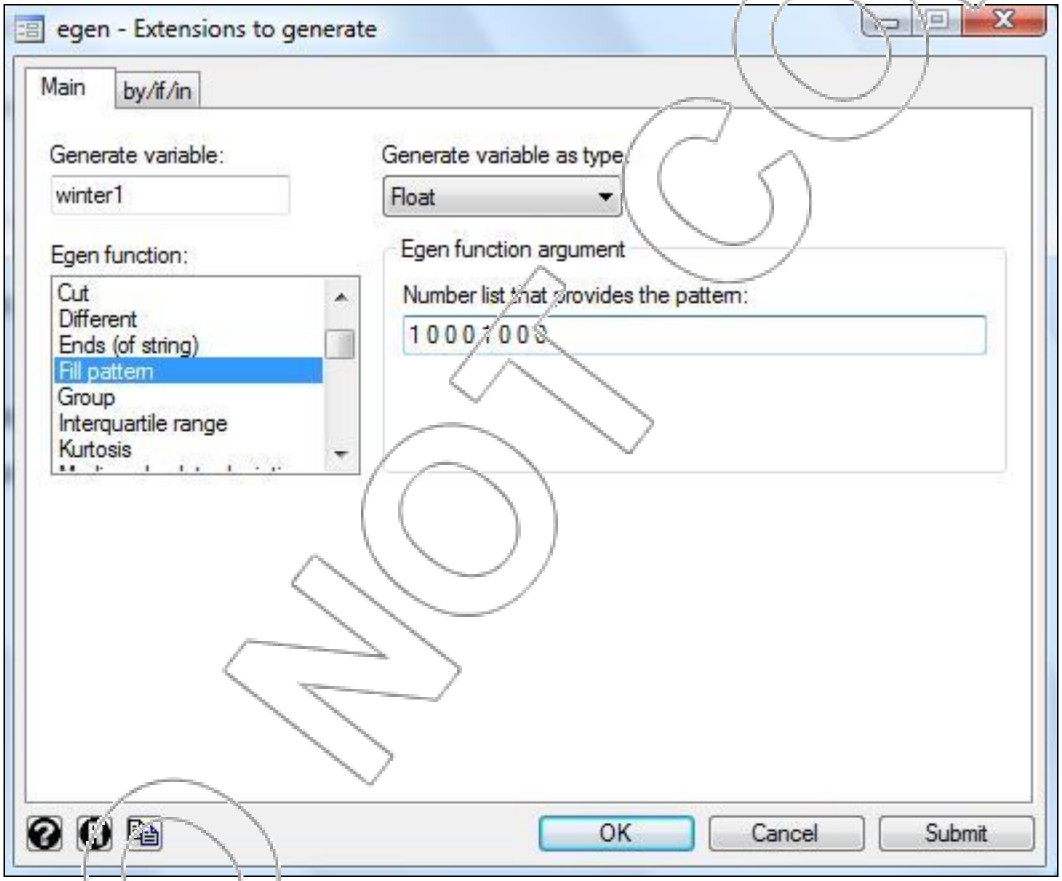
Many other functions are available in the **Expression builder** dialog box (accessible via the **Create…** button in the **generate** dialog box) that you can use to manipulate data. For more information on data generating options, you can type **help generate** into the Command box or refer to Stata's accompanying manual on data management.

Another type of variable we may want to create using Stata is a seasonal dummy variable. In the **soda.dta** dataset, we have the dummy variables **winter**, **spring**, and **summer**. **Winter**, for example, is a column with the following sequence of numbers:

1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0

There is a one for each row of data that corresponds to a winter quarter, and a zero for any other quarter. One way to construct a variable like this is to open the Data Editor, type a **1** into the first cell of an empty column, and type three zeroes into the second, third, and fourth cells. Then, copy these four cells and paste them by choosing the appropriate cells as a destination. In the soda example, you need to paste this pattern three more times. Stata automatically names a new variable "**var#**" when you initially enter data manually into a new column. To rename your variable, right-click on the variable name at the top of the column, and click **Variable Properties…**. Type in the name that you want and click **Apply**, and your new variable will be renamed appropriately. The direct command for renaming a variable is **rename *oldvar newvar***.

Manually entering data with repeated patterns can be very tedious, especially when you have a very large dataset. Fortunately, you can use the **fill()** function of the **egen** command to generate a variable with repeating patterns easily. For example, suppose we want to generate an additional column of data in the **soda.dta** dataset, say, **winter1**, that is identical to the **winter** variable. To do this using the **fill()** function, click **User>Manipulate Variables and Obs>Extended Generate New Variables (egen)** or type **db egen**.[17] Type **winter1** in the "Generate variable" field, select **Fill pattern** from list of egen functions, and enter **1 0 0 0 1 0 0 0** in the "Number list that provides the pattern" field (you should generally enter a pattern twice so that Stata understands exactly what pattern you would like it to repeat).[18] You should have a dialog box that looks like this:



Generate variable: winter1

Generate variable as type: Float

Egen function: Fill pattern

Number list that provides the pattern: 1 0 0 0 1 0 0 0

Click **OK** and examine the Data Browser. You will see that Stata has generated the variable **winter1** with the sequence 1 0 0 0 repeated four times.

[17] The native menu option in Stata is **Data>Create or change data>Create new variable (extended)**.

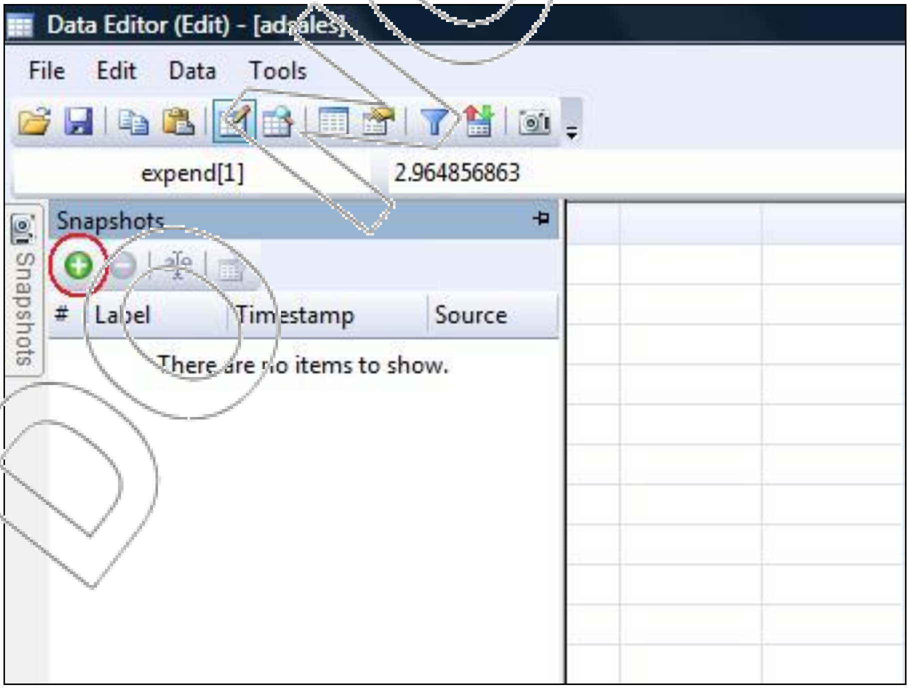
[18] Alternatively, you can directly type the command **egen winter=fill(1 0 0 0 1 0 0 0)**.

Other Stata Commands

Keeping Track of Edited Data

The **snapshot** command in Stata is very useful in recording the changes that you have made to your dataset. Every time you create a snapshot, Stata will save a copy of your dataset up to that moment. Therefore, if you make any editing error or simply want to restore your dataset to an earlier state, you can select the appropriate snapshot that you want to return to.

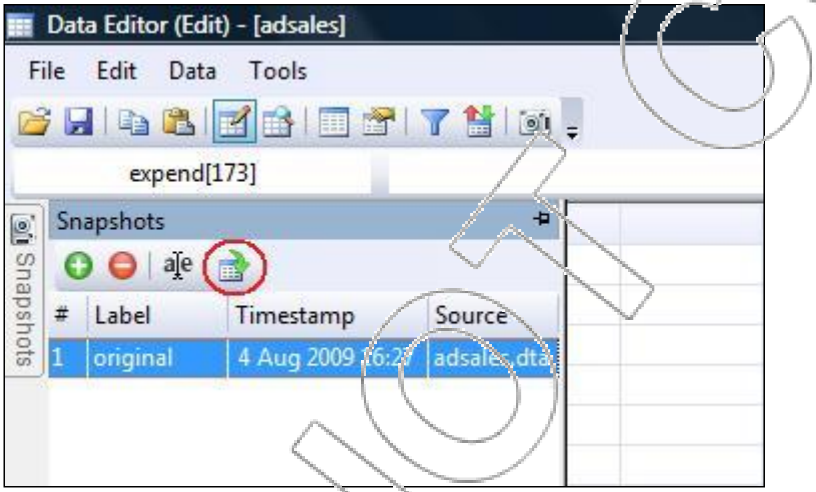
For example, suppose we want to edit the **adsales.dta** data. The original dataset contains 172 observations, and we want to add the 173rd observation where expend=2.2 and sales=16.79 (for illustrative purpose only). Open the Data Editor. Before making any changes, you can create a snapshot of the original dataset by clicking **Tools>Snapshots…** or clicking the **Snapshots** tab. This will expand the Snapshots window on the left of the Data Editor. Click on the Add button (shown below):



expend[1] 2.964856863

There are no items to show.

Enter a name, say, **original**, to remind ourselves what the data snapshot contains. Click **OK**, and you can see in the Snapshots list that Stata has created the first snapshot of your data. Now we can proceed to enter new values in the **adsales** dataset. Suppose, however, that we accidentally entered 2.2 in cell **expend[172]** instead of **expend[173]**. The original value in cell **expend[172]**, 2.507401228, is now lost, and we want to rectify this mistake. To do this, click on the **Snapshots** tab again. Select the snapshot that you want to restore to (**original** in this case) and click on the Restore button as shown:



#	Label	Timestamp	Source
1	original	4 Aug 2009 16:2	adsales.dta

Click **Yes**, and Stata will restore our data back to its original state.

The direct command for creating a snapshot is **snapshot save, label("*snapshotname*")**; the direct command for restoring to an earlier snapshot is **snapshot restore snapshot#**, where **snapshot#** corresponds to the number under the **#** column in the Snapshots list.

The by/if/in Option

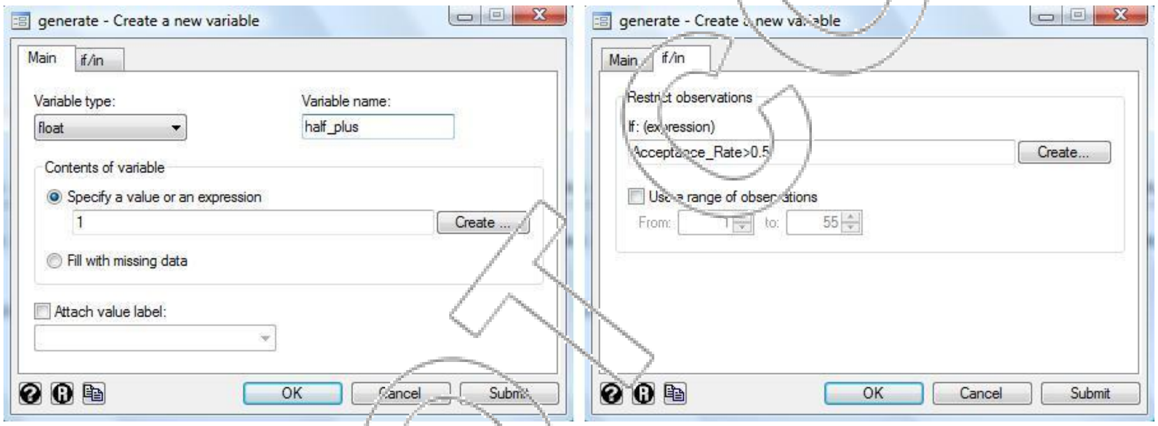
The **by**, **if**, and **in** options are useful for specifying particular portions of data that you want to use. Specifically, the **by *varlist*** option repeats a command for groups of observations defined as having the same values for the variables in ***varlist***. The **if *exp*** option specifies that a command is carried out only for observations satisfying the expression in ***exp***. The **in** option specifies a range of data for which you want to carry out a command.

As an example, consider the California Strawberries case from Section 5.2, where we want to run the regression of Time versus Boxes separately for the Monterey and the Bakersfield systems. Open the **california.dta** dataset, which contains a dummy variable **Plant** that equals 0 if the data come from the Monterey plant and 1 if the data come from the Bakersfield plant. We can utilize the **Plant** variable and the **by/if/in** options to run the separate regressions in three different ways. The corresponding direct commands are the following:

1. Using the **by** option:
    a. **by Plant, sort: regress Time Boxes**
2. Using the **if** option:
    a. **regress Time Boxes if Plant==0**
    b. **regress Time Boxes if Plant==1**
3. Using the **in** option:
    a. **regress Time Boxes in 1/15**
    b. **regress Time Boxes in 16/30**

You should try these three sets of commands and verify that they produce the same regression output. Note that the **by** option sorts the data by the value of **Plant** before doing the regression. It doesn't matter in this example (because the data is already sorted in this way), but more generally, you should be careful not to save the data in its sorted form if you wish to maintain the original observation order.

The **if *exp*** option is also frequently used in generating or manipulating variables. For example, in Case Exercise 4 of Chapter 1, we wanted to create a new variable called **half_plus** that equals 1 if Acceptance_Rate is greater than 50 percent and equals 0 otherwise. To do this, you can click **User>Manipulate Variables and Obs>Generate New Variable (generate)** or type **db generate**. Type **half_plus** into the "Variable name" field, and type **1** into the "Specify a value or an expression" field. Switch to the **if/in** tab and type **Acceptance_Rate>0.5** into the "If: (expression)" field.[19] You should have a dialog box that looks like this.



generate - Create a new variable

Main tab: Variable type: float; Variable name: half_plus

Contents of variable

● Specify a value or an expression: 1

○ Fill with missing data

□ Attach value label:

generate - Create a new variable

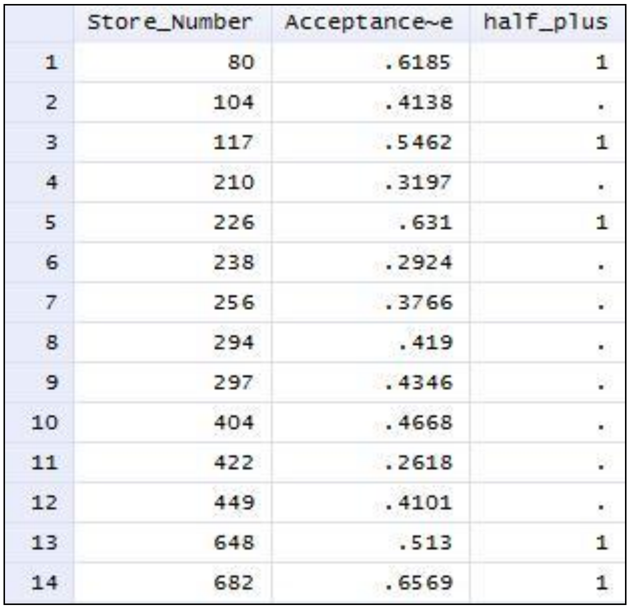
if/in tab: Restrict observations

If: (expression): Acceptance_Rate>0.5

□ Use a range of observations From: 1 to: 55

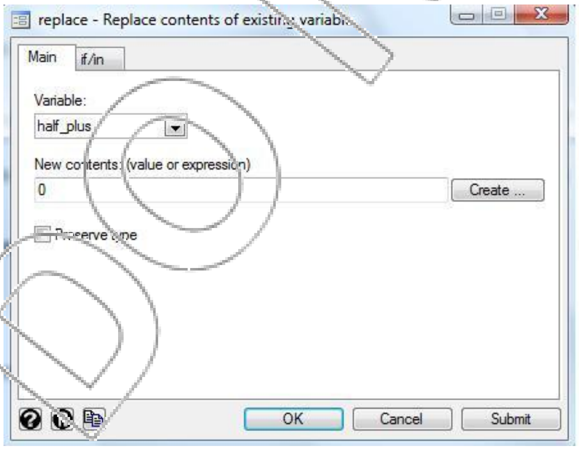
Click **OK** and open the Data Browser. You can see below that **half_plus** has a value of 1 for all observations where **Acceptance_Rate** is greater than 0.5, or 50 percent:

[19] The direct command is **generate half_plus=1 if Acceptance_Rate>0.5**.



	Store_Number	Acceptance~e	half_plus
1	80	.6185	1
2	104	.4138	.
3	117	.5462	1
4	210	.3197	.
5	226	.631	1
6	238	.2924	.
7	256	.3766	.
8	294	.419	.
9	297	.4346	.
10	404	.4668	.
11	422	.2618	.
12	449	.4101	.
13	648	.513	1
14	682	.6569	1

For any observation where Acceptance_Rate is less than or equal to 0.5, Stata has left a corresponding blank cell in the **half_plus** column. To replace the empty cells with 0, you can click **User>Manipulate Variables and Obs>Replace/Change Existing Variables (replace)** or type **db replace**. Select **half_plus** in the "Variable" field, and enter **0** in the "New contents: (value or expression)" field. Switch to the **if/in** tab, and type **half_plus==.** in the "If: (expression)" field.[20] You should have a dialog box that looks like this:

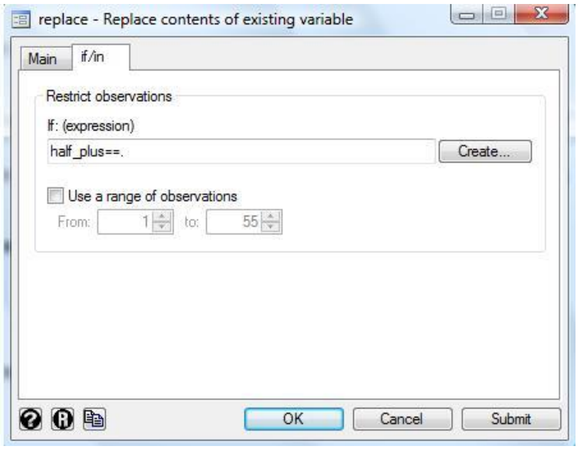


replace - Replace contents of existing variable

Variable: half_plus

New contents: (value or expression) 0

☐ Preserve type



replace - Replace contents of existing variable

Restrict observations

If: (expression) half_plus==.

☐ Use a range of observations

From: 1 to: 55

[20] The direct command is **replace half_plus=0 if half_plus==.**.

Click **OK** and look at the Data Browser again. You should see that all previous empty cells in the **half_plus** column have now been replaced with 0's instead.

There are many other expressions that you can use with the **if** option to automate the task of data analysis and/or data manupulation. You can explore them by typing **help if** into the Command box or by referring to Stata's pdf manuals.

Declaring a Time Index Variable

When analyzing time series data in Stata, you first need to designate or generate a time index variable by using the **tsset** command. If you want to declare an existing variable as a time index, you can click **Statistics>Time series>Setup and utilities>Declare dataset to be time-series data** or type **db tsset**, and select the desired variable from the "Time variable" field. The direct command is **tsset *varname***.

An easy way to generate a generic time index variable is by first typing the command **generate *newvar*=[_n]** where ***newvar*** is whatever name you want to give to the variable. This command generates a new variable with values corresponding to the observation numbers of your dataset. Then, declare ***newvar*** as a time index by using either the **tsset** dialog box or the direct command **tsset *newvar***.

To stop designating a variable as a time index, you can click the "Clear time-series settings" button in the **tsset** dialog box or type the direct command **tsset, clear**.

Doing Calculations in Stata

You can use Stata's **display** command as a hand calculator. For example, to calculate ln(2)/5, you can type **display ln(2)/5** into the Command box and get 0.13862944. The abbreviation **di** can also be used in place of **display**.

Everything else

Stata is capable of many tasks not discussed here. As you work through the problems in this book, you will become more familiar with the program and a few of its many capabilities. To learn more about a particular command, you can type **help *commandname*** in the Command box. The Stata User's Guide (in the pdf manual that comes with Stata) also provides a comprehensive description of its commands. The Stata FAQ website (http://www.stata.com/support/faqs/ or click **Help>Stata Web Site>Frequently Asked Questions**) and the Stata listserver (http://www.stata.com/statalist/) are also good online sources for technical and/or statistical questions.

Prediction Intervals

What is a prediction interval?

A **prediction interval** is a confidence interval for a particular observation, rather than for the population mean, μ. In Chapter 1, you learned the formulas for confidence intervals for μ. The formulas for prediction intervals differ in two important ways from those formulas:

1. We can only calculate prediction intervals easily if we assume that the population is normally distributed.
2. For prediction intervals, we need to take into account the variance of an individual observation (the population variance) as well as the variance of $\overline{X}$. For confidence intervals concerning μ, it was only necessary to consider the variance of $\overline{X}$.

How do we calculate a (1-α)*100% prediction interval?

Assume our sample of size n is i.i.d. and is drawn from a normally distributed population.

1. If we know the population standard deviation, σ, the P.I. is the following:

$$\overline{X} \pm z_{\alpha/2}\sigma\sqrt{\frac{1}{n}+1}$$

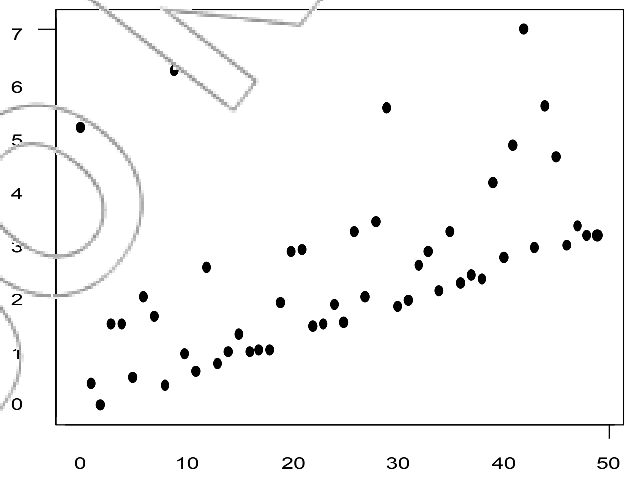
2. If the population standard deviation is not known, then the P.I. is the following:

$$\overline{X} \pm t_{\alpha/2;n-1}s\sqrt{\frac{1}{n}+1}$$

Correlation

Usually, the value of a random variable conveys some information regarding the value of another random variable. For example, if you know the height of someone, this gives you some idea about this person's weight. Typically, a taller person is heavier than a shorter person. This is not always the case, but it is fair to say that height and weight are positively correlated. Examples of positively correlated random variables abound, such as sales and advertising expenditures, the price of a Coke and the price of a Pepsi, inflation and the increase in the money supply, education and wages. In all these examples, the random variables are positively correlated because the probability of a high realization of one random variable is higher when the realization of the other random variable is high than when the realization of the other random variable is low.

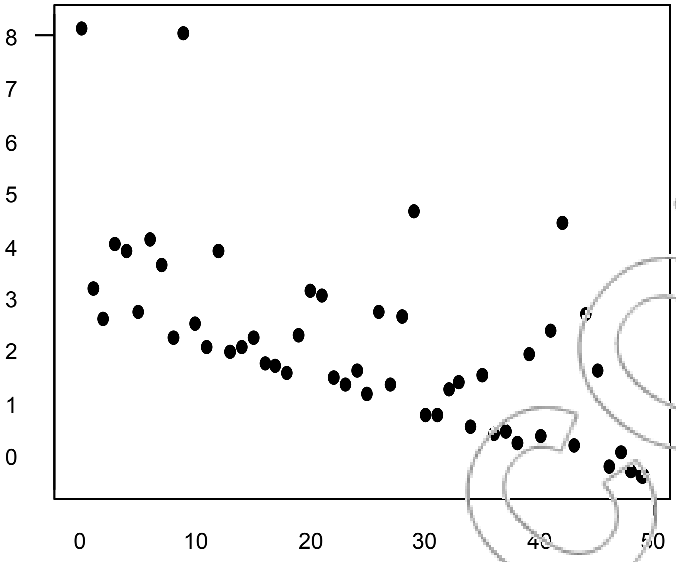
A plot of two positively correlated variables may look like this:



An extreme case of positively correlated variables is the case of two variables perfectly and positively correlated. In this extreme case, one variable is a positive linear transformation of the other, such as the price of a hamburger measured in cents and the price of a hamburger measured in dollars. One random variable is the other multiplied by 100.

Analogously, two random variables are negatively correlated if one is likely to be above average when the realization of the other random variable is low and below average when the realization of the other random variable is high. Examples of negatively correlated random variables also abound: inflation and contraction in the money supply, wages and poverty, and health and cigar consumption.

A plot of negatively correlated random variable may look like this.



An extreme case of negatively correlated variables is the case of two variables perfectly and negatively correlated. In this extreme case, one variable is a negative linear transformation of the other.

Two random variables are independent if the realization of one random variable does not affect the probability distribution of the other random variable. A typical example of two independent random variables is given by tossing two different coins. Two independent random variables are not correlated.

The sample correlation coefficient of two variables x and y is obtained by dividing the sample covariance by the product of the sample standard deviation of x and the sample standard deviation of y:

$$r_{xy} = s_{xy}/(s_x s_y)$$

The components are as follows:

r_{xy} = sample correlation coefficient

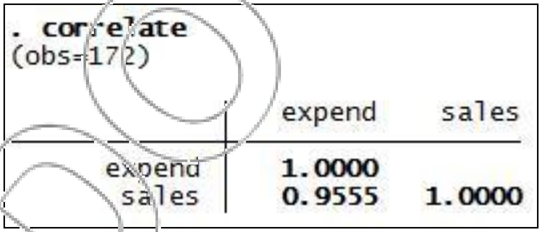
s_{xy} = sample covariance

s_x = sample standard deviation of x

s_y = sample standard deviation of y

The correlation coefficient of two variables is always between -1 and 1. If it is -1, the two variables are perfectly negatively correlated. If it is 1, the two variables are perfectly positively correlated.

Using Stata, you can find the correlation coefficients between all possible pairs of variables in your dataset. To do this, click **User>Core Statistics>Bivariate Statistics>Correlations (correlate)** or type **db correlate**. For example, using the **adsales.xls** data, we produce the following output:



```
. correlate
(obs=172)

             |   expend    sales
-------------+------------------
      expend |   1.0000
       sales |   0.9555   1.0000
```

Here, 0.9555 is the correlation between expend and sales.

If your dataset contains more than two variables, Stata will return a table giving the correlation between any pair. If any of the variables are non-numeric, **correlate** will return an error. To avoid this, you can specify in the **correlate** dialog box exactly which variables you would like to see the correlations among.

Properties Of Logarithms

In this section, we outline some of the mathematical properties of logarithms, logs from here on, we will need to use in this text. In this book (as in most real-world applications), we will use only **natural logs**. Natural logs are called “natural” because they use the natural number **e = 2.71...**. We will use the notation **ln** for natural logs. Other common notations are loge or log though the latter more often refers to a different kind of logarithm, i.e., log base 10.

Definition: the natural logarithm of a number x is the number y that satisfies: $e^y = x$.

So, y=ln x means y is the power you have to raise e to in order to get x. It’s okay if the log of something is negative. It means you need to raise e to a negative number to get that value. On the other hand, there is no number you can raise e to and get -1; ln -1 is not defined. In fact, ln x is not defined for any negative x.

Fractional values for the log are possible:

$$\ln \sqrt{e} = 1/2 \text{ since } e^{1/2} = \sqrt{e}\,.$$

Negative fractions are allowed as well. ln x = -0.5 means that x is the -1/2 power of e or 1 over the square root of e. One general rule is that as x goes up, ln x goes up as well, but not nearly as fast as x does. In fact, as x goes up geometrically, ln x goes up linearly.

Raising something to a power ‘undoes’ the log as in this example:

$$e^{\ln x} = x, \text{ e.g., } e^{\ln 4} = 4.$$

The same holds in the opposite order as well:

$$\ln e^x = x, \text{ e.g., } \ln e^2 = 2.$$

SUMMARY OF PROPERTIES OF LOGS

There are a handful of properties of logs that get used a lot in general and in this book in particular. Here are some of the most important ones.

Property 1: Exponentiation and logs are inverses in that they undo each other. In particular, for any positive number x, the following is true:

$$e^{\ln (x)} = x \text{ and } \ln (e^x) = x$$

Example: $e^{\ln e} = e^1 = e$ and $\ln (e^1) = \ln (e) = 1$.

Property 2: Logs of products are sums:

$$\ln (x*y) = \ln (x) + \ln (y)$$

This is true because you can add exponents in products as in this example.

$$\ln (e^2\, e) = \ln (e^3) = 3 = 2+1 = \ln (e^2) + \ln (e)$$

Property 3: Logs of powers are products:

$$\ln (x^y) = y \ln (x)$$

This is the same as property 2 above when you multiply the same thing together y times as in this example:

$$\ln (e^3) = \ln (e*e*e) = \ln e + \ln e + \ln e = 3 \ln (e)$$