Linear Regression Analysis:
Assumptions and Applications

John P. Hoffmann
Department of Sociology
Brigham Young University
Contents

Preface iii

1 A Review of Some Elementary Statistical Concepts 1

2 Simple Linear Regression Analysis 21

3 Multiple Linear Regression Analysis 43

4 The ANOVA Table and Goodness-of-Fit Statistics 59

5 Comparing Linear Regression Models 69

6 Dummy Variables in Linear Regression Models 77

7 Specification Errors in Linear Regression Models 93

8 Measurement Errors in Linear Regression Models 111

9 Collinearity and Multicollinearity 127

10 Nonlinear Associations and Interaction Terms 143

11 Heteroscedasticity and Autocorrelation 175

12 Influential Observations: Leverage Points and Outliers 203

13 A Brief Introduction to Logistic Regression 217

14 Conclusions 233

Appendix A: Data Management 239

Appendix B: Formulas 251
Preface

Rarely does a day go by that we are not exposed to data of one sort or another. We find reports of studies in newspapers, magazines, and on-line that show graphs and charts of data collected from people, animals, or even abstract entities such as cities, counties, or countries. Life expectancies, crime rates, pollution levels, unemployment rates, satisfaction surveys, election results, and numerous other phenomena are presented with overwhelming frequency and in painful detail. Understanding statistics – or at least being able to talk intelligently about percentages, means, margins for error, and the like – has become almost mandatory for a well-educated person. Yet, few people understand enough about statistics to fully grasp not only the strengths, but also the weaknesses, of the way data are collected and analyzed. What does it mean to say that the life expectancy in the United States is 75.8? Should we trust exit polls that claim that Barker will win the election over Hanks by five percent (with a margin of error of ± two percentage points)? When someone claims that “taking calcium supplements is not associated with a significantly lower risk of bone fractures in elderly women” what are they actually saying? Is it really meaningful for a sociologist to say that each one-year increase in a person’s education is associated with a five point increase in life satisfaction? These questions, as well as many others, are common in the world of contemporary statistical analysis.

For the budding social scientist, whether sociologist, psychologist, political scientist, or economist, it is almost impossible to avoid sophisticated quantitative analyses that move well beyond simple statistics such as percentages, means, t-tests, and standard deviations. A large percentage of the studies found in journals devoted to these disciplines use statistical models that are designed to predict (or explain) the occurrence of one variable with information about another variable. The most common type of model designed to predict a variable is the linear regression model. Numerous books and articles explain how to conduct what is usually termed a linear regression analysis or estimate a linear regression model. Students are usually exposed to this model in a second course on applied statistics. As we shall learn, the reason for its popularity is because (1) it is relatively easy to
understand; (2) statistical software designed to estimate it is widely available; and
(3) it is a flexible and powerful method for conducting important types of analyses.

Nonetheless, even though linear regression models are widely used and offer an
important set of tools for understanding the association between two or more
variables, they are often misused by those who fail to take into account the many
assumptions analysts are forced to make. For instance, a linear regression model
does not provide suitable results when two or more of the independent variables (if
you don’t know this term, you will; just keep reading) have high correlations (see
Chapter 9). Similarly, if one of the data points is substantially different from all the
others, then the linear regression model is forced to compensate, often with poorly
fitting results (see Chapter 12). One of the goals of this presentation, therefore, is
to provide a relatively painless overview of the assumptions we make when using
linear regression models.

The main purpose of this presentation, though, is to show the reader how to
use linear regression models in studies that include quantitative data. Specific
objectives include discussing why linear regression models are used, what they tell
us about the relationships between two or more variables, what the assumptions of
the model are and how to determine if they are satisfied, what to do when the
assumptions are not satisfied (sometimes we may still use linear regression by
making some simple modifications to the model), and what to do when the
dependent variable is measured using only two categories. As we shall learn, linear
regression models are not designed for two-category dependent variables, so we’ll
learn about another model, known as a logistic regression model, which is designed
for these types of variables (see Chapter 13).

I have been teaching these methods for several years and have seen many
students struggle and more succeed. Thus, I know how important it is that
students are familiar with some standard statistical concepts before beginning to
learn about linear regression analysis. Hence, the first chapter is designed as a quick
and dirty review of elementary statistics. For the reader who has not been exposed
to this material (or who has forgotten it), I recommend reviewing a basic statistics
textbook to become familiar with means, medians, standard deviations, standard
errors, z-scores, t-tests, correlations and covariances, and analysis-of-variance
(ANOVA). I also suggest that the reader take some time to learn SPSS, a statistical
software package designed to carry out most of the analyses presented herein. It is a
relatively easy program to master, especially for students who have used some type of spreadsheet software. I find SPSS’s drop-down menus especially intuitive and simple to use. It also may be used through syntax files, which are simple programs of instructions for running analyses.

The chapters follow the typical format for books on linear regression analysis. As mentioned earlier, we shall first review elementary statistics. This is followed by an introductory discussion of the simple linear regression model. Second, we will learn how to interpret the results of the linear regression model. Third, we shall see how to add additional independent variables to the model. This transforms it into a multiple linear regression model. We then learn about goodness-of-fit statistics, model comparison procedures, and dummy independent variables. Fourth, we move into an in-depth discussion of the assumptions of linear regression models. We’ll spend several chapters on exciting (okay, too strong of a word) and mystifying topics such as multicollinearity, heteroscedasticity, autocorrelation, and influential observations. We’ll finish the presentation by learning about the logistic regression model, which, as mentioned earlier, is designed for dependent variables that include only two categories (e.g., “Do you support the death penalty for murder?” 0 = no, 1 = yes).

There is an important issue that I hope readers will consistently consider as they read this text. Statistics has, for better or worse, been maligned by many observers in recent years. Books with titles such as How to Lie with Statistics are popular and can lend an air of disbelief to many studies that use statistical analysis. Researchers and statistics educators are mainly to blame for this disbelief. We often fail to impart two things to students and consumers: (1) a healthy dose of skepticism; and (2) encouraging them to use their imagination and common sense when assessing data and the results of analyses. I hope that readers of the following chapters will be comfortable using their imagination and reasoning skills as they consider this material and as they embark on their own quantitative studies. In fact, I never wish to underemphasize the importance of imagination for the person using statistical techniques. Nor should we suspend our common sense and knowledge of the research literature simply because a set of numbers demonstrates some unusual conclusion. This is not to say that statistical analysis is not valuable or that the results are misleading. Clearly, statistical analysis has led to many important discoveries in medicine and the social sciences, as well as informed policy in a
positive way. The point I wish to impart is that we need a combination of tools –
including statistics, but also our own ideas and reasoning abilities – to help us
understand important social and behavioral issues.

Acknowledgements

I would first of all like to thank the many students, undergraduate and graduate,
who have completed courses with me on this topic or on other statistical topics. I
have learned so much more from them than they will ever learn from me. I have
had the privilege of having as research assistants Scott Baldwin, Colter Mitchell,
and Bryan Johnson. Each has contributed to my courses on statistics in many ways,
almost all positive! Finally, Kristina Beacom and Amanda Cooper helped me put
together the chapters. I am indebted to them for the wonderful assistance they
provided.
A Review of Some Elementary Statistical Concepts

Elementary statistics can be confusing, especially to people who are uncomfortable with numbers. Many of us were introduced to statistics in a pre-algebra or an algebra course. However, your initial introduction probably occurred in elementary school. Do you remember the first time you heard the word *mean* used to indicate the average of a set of numbers? This likely took place in some math sequence you were exposed to in elementary school. How about graphing exercises? Do you recall being given two sets of points and being asked to plot them on graph paper? You were introduced to the $x$-axis and the $y$-axis, or the coordinate axes.

Around the same time, or perhaps a little later, you became familiar with elementary *probability*. This likely took the form of a question such as “What is the probability of a die being thrown and landing on a five?” You first learned that you needed to count the number of possible outcomes (there are six faces on a typical die, so there are six possible outcomes). This was the denominator. Then you counted the particular outcome. This was the numerator. Putting these two counts together, you learned that the probability of the roll coming up as five is $1/6$, or approximately 0.167. This latter value is known as a *proportion*. Proportions – and probabilities – must fall between zero and one. They can easily be transformed into percentages by moving the decimal place over two spaces to the right (or multiplying by one-hundred) and placing a percentage sign next to the number. What does this mean, though? Well, one way to consider it is to say that we expect a five to come up about 16.7% of the time when we roll a typical die numerous times. Of course, you can probably confirm this by rolling the die many, many times. Some statisticians refer to such a view as a *frequentist* interpretation of or approach to statistics.

Probabilities are normally presented using, not surprisingly, the letter $P$. One way to represent the probability of a five from a roll of a die is with $P(5)$. So we may write $P(5) = 0.167$ or $P(5) = 1/6$. You might recall that some statistical tests,
such as $t$-tests (see the description later in the chapter) or ANOVAs, are often accompanied by $p$-values. As we shall learn, $p$-values are a type of probability value used in many statistical tests.

By combining the principles of probability and elementary statistical concepts, we may develop the basic foundations for statistical analysis. In general, there are two types of statistical analyses: Descriptive and inferential. The former set of methods is normally used to describe or summarize one or more variables (recall that the term variables is used to indicate phenomena that can take on more than one value; this contrasts with constants, or phenomena that take on only one value). Some common terms that you are probably familiar with are measures of central tendency and measures of dispersion. We'll see several of these measures a little later. Then there are the many graphical techniques that may be used to “see” the variable. You might recognize techniques such as histograms, stem-and-leaf plots, dot-plots, and box-and-whisker plots.

Inferential statistics are designed to infer or deduce something about a population from a sample. Suppose, for instance, that we are interested in determining who is likely to win the next Presidential election in the United States. We'll assume there are only two candidates from which to choose: Clinton and Bush. Of course, it would be enormously expensive to ask each person who is likely to vote in the next election who they plan to vote for. Therefore, we may take a sample of likely voters and ask them who they plan to vote for. Can we infer anything about the population of voters based on our sample? The answer is that it depends on a number of factors. Did we collect a good sample? Were the people who responded honest? Do people change their minds as the election approaches? We don’t have time to get into the many issues involved in sampling, so we'll have to assume that our sample is a good representation of the population from which it is drawn. Most important for our purposes is this: Inferential statistics include a set of techniques designed to help us answer questions about a population from a sample.

Another way of dividing up statistics is to compare techniques that deal with one variable from those that deal with two or more variables. Most readers of this presentation will likely be familiar with techniques designed for one variable. These include, as we shall see later, most of the descriptive statistical methods. The bulk
of this presentation, at least in later chapters, concerns a technique designed for analyzing two or more variables simultaneously. A key question that motivates us is whether two or more variables are associated in some way. As the values of one variable increase, do the values of the other variable also tend to increase? Or do they decrease? In elementary statistics students are introduced to covariances and correlations, two techniques designed to answer these questions generally. But recall that you are not necessarily saying that one variable somehow changes another variable. Remember the maxim: *Correlation does not equal causation.* We'll try to avoid the term *causation* in this presentation because it involves many thorny philosophical issues (see Judea Pearl (2000), *Causality: Modeling, Reasoning, and Inference*, New York: Cambridge University Press). Nonetheless, one of our main concerns is whether one or more variables is associated with another variable in a systematic way. Determining the characteristics of this association is one of the main goals of the linear regression model that we shall learn about later.

**Measures of Central Tendency & Dispersion**

Now that we have some background information on elementary statistics, let’s learn more about the most important measures, including how they are used and how they are computed. We’ll begin first with measures of central tendency. Suppose we have collected data on a variable such as weight in kilograms. Our intrepid researchers have carefully placed each person in the sample on a scale and recorded their body weights. To simplify things, we’ll assume they rounded the weights to the nearest kilogram. What would be your best guess of the average weight among the sample? It is not always the best, but the most frequent measure is the *arithmetic mean*, which is computed using the following formula:

$$
E[X] = \bar{x} = \frac{\sum x_i}{n}
$$

The term on the left-hand side of the equation is $E[X]$. This is a short-hand way of saying that this is the *expected value* of the variable $X$. It is often used to represent the mean. To be more precise, we might also list this term as $E[\text{weight in kg}]$, but normally, as long as it’s clear that $X = \text{weight in kilograms}$, using $E[X]$ is sufficient. The middle term – read as *x-bar* – may also be familiar as a common symbol for the
mean. The formula for computing the mean is rather simple. We add all the values of the variable and divide this sum by the number of observations. Note that the rather cumbersome symbol that looks like an overgrown $E$ in the right-hand part of the equation is the summation sign; it tells us to add whatever is next to it. The symbol $x_i$ denotes the specific values of the $x$ variable, or the individual weights that we’ve measured. The subscript $i$ indicates each observation. The symbol $n$ represents the sample size. Sometimes the individual observations are represented as $i…n$. If you know that $n = 5$, then you know there are five individual observations in your sample. In statistics, we often use upper-case Roman letters to represent population values and lower case Roman letters to represent sample values. Therefore, when we say $E[X] = \bar{x}$, we are implying that our sample mean estimates the population expected value, or the population mean.

Here’s a simple example: We have a sample of people’s weights that consist of the following set: [84, 75, 80, 69, 90]. The sum of this set is $[84 + 75 + 80 + 69 + 90] = 398$; therefore the mean is $398/5 = 79.6$. Another way of thinking about this mean value is that it represents the center of gravity. Suppose we have a plank of wood that is magically weightless (or of uniform weight across its span). We order the people from lightest to heaviest – trying to space them out proportional to their weights – and ask them to sit on our plank of wood. The mean is the point of balance, or the place at which we would place a fulcrum underneath to balance the people on the plank.

There are some additional things you should know about the mean. First, it is measured in the same units as the observations. If your observations are not all measured in the same unit (e.g., some people’s weights are in kilograms, others in pounds), then the mean cannot be interpreted. Second, the mean provides a good measure of central tendency if the variable is measured continuously and is
normally distributed. What do these terms mean? A variable is measured continuously – or we say the variable is *continuous* – if it can conceivably take on any real number. Of course, we usually cannot be so precise when we measure things, so it is not uncommon to round our measures to whole numbers or integers. We also often measure things using positive numbers only; it makes little sense, for instance, to try to measure a person’s weight using negative numbers.

The other type of variable is known as *discrete* or *categorical*; these variables have a finite number of possible values. For example, we normally use only two categories to measure gender: female and male. Hence, this is a discrete variable.

We say a variable is normally distributed if it follows a bell-shaped curve. Order the values of the variable from lowest to highest and then plot them by their frequencies or the percent of observations that have a particular value (we must assume that there are many values of our variable). We may then view the “shape” of the distribution of the variable. The graph on the next page shows an example of a bell-shaped distribution, usually termed a normal or Gaussian distribution, using a simulated sample of weights. (It is known as Gaussian after the famous German mathematician, Carl Friedrich Gauss, who purportedly discovered it.)

We shall return to means and the normal distribution frequently in this presentation. To give you a hint of what is to come, the linear regression model is designed, in part, to predict means for particular sets of observations in the sample. For instance, if we have information on the heights of our sample members, we may wish to use this information to predict their weights. Our predictions could include predicting the mean weight of people who are 72 centimeters tall. We may use a linear regression model to do this.

But, suppose that our variable does not follow a normal distribution. May we still use the mean to represent the average value? The simple answer is yes, as long as the distribution does not deviate too far from the normal distribution. In many situations in the social and behavioral sciences, though, variables do not have normal distributions. A good example of this involves annual income. When we ask a sample of people about their annual incomes, we usually find that a few people earn a lot more than others. Measures of income typically provide *skewed* distributions, with long right tails. If asked to find a good measure of central tendency for income, there are several solutions available. First, we may take the
natural (Naperian) logarithm of the variable. You might recall from an earlier mathematics course that using the natural logarithm (or the base 10 logarithm) pulls in extreme values. If this is not clear, try taking your calculator and using the LN function with some large and small values (e.g., 10 and 1,500). You will easily see the effect this has on the values of a variable. If you’re lucky, you may find that taking the natural logarithm of a variable with a long right-tail transforms it into a normal distribution. The square root or cube root of a variable may also work to “normalize” a skewed distribution. We’ll see examples of this in Chapter 10.

Second, there are several direct measures of central tendency appropriate for skewed distributions (or other distributions plagued by extreme values such as outliers; see Chapter 12), such as the trimmed mean and the median. The trimmed mean cuts off some percentage of values from the upper and lower ends of the distribution, usually five percent, and uses the remaining values to compute the mean value. The median should be familiar to you. It is the middle value of the distribution. In order to find it, we first order the values of the variable from lowest to highest. Then we choose the middle value if there are an odd number of
observations, or the average of the middle two values if there are an even number of observations. If you are familiar with percentiles (or quartiles or deciles), then you might recall that the median is the 50th percentile of a distribution. The median is known as a robust statistic because it is relatively unaffected by extreme values. As an example, suppose we have two variables, one that follows a normal distribution (or something close to it) and another that has an extreme value:

Variable 1:  [45, 50, 55, 60, 65, 70, 75]
Variable 2:  [46, 51, 54, 59, 66, 71, 375]

Variable 1 has a mean of 60 and a median of 60, so we make the same estimate of its central value regardless of which measure is used. In contrast, Variable 2 has a mean of 103, but a median of 59. Although we might debate the point, I think most people would agree that the median is a better representative of the average value than the mean for Variable 2.

The next issue to address from elementary statistics involves measures of dispersion. As the name implies, these measures consider the spread of the distribution of a variable. Most readers are familiar with the term standard deviation, since it is the most common measure for continuous variables. However, before seeing the formula for the standard deviation, it is useful to consider some other measures of dispersion. The most basic measure is the sum of squares, or $SS[X]$:

$$SS[X] = \sum (x_i - \bar{x})^2$$

This formula first computes deviations from the mean $(x_i - \bar{x})$, squares each one, and adds them up. If you’ve learned about ANOVA models, the sum of squares should be familiar. Perhaps you even recall that there are various forms of the sum of squares. We’ll learn more about these in Chapter 4.

A second measure of dispersion that is likely more familiar to you is the variance, or $Var[X]$. It is often labeled as $s^2$. The formula is
LINEAR REGRESSION ANALYSIS

\[
\text{Var}[X] = s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}
\]

Notice that another way of computing the variance is to take the sum of squares and divide it by the sample size minus one. One of the drawbacks of the variance is that it is measured not in units of the variable, but rather in squared units of the variable. To transform it into the same units as the variable, it is therefore a simple matter to take the square root of the variance. This measure is the standard deviation (it is also denoted by the letter \( s \)):

\[
\text{SD}[X] = s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}
\]

A variable’s distribution, assuming it is normal, is often represented by its mean and standard deviation (or variance). In short-hand form, this is listed as \( x \sim N(\bar{x}, s) \) (the wavy line means “distributed as”). Obviously, a variable that is measured in the same units as another and that shares the same mean is less dispersed if its standard deviation is smaller. Although not often used, another promising measure of dispersion is the coefficient of variation (CV), which is computed by dividing the standard deviation by the mean of the variable \((s/\bar{x})\). It is often multiplied by 100. The CV is valuable because it shows how much a variable varies about its mean.

An important point to remember is that symbols such as \( s \) and \( s^2 \) are used to represent sample statistics. Greek symbols – or, as we have seen up until this point, upper-case Roman letters – are often used to represent population statistics. For example, the symbol for the population mean is the Greek letter \( \mu \) (\( \mu \)), whereas the symbol for the population standard deviation is the Greek letter \( \sigma \) (\( \sigma \)). However, we’ll see when we get into the symbology of linear regression that Greek letters are used to represent other quantities as well.

Another useful measure of dispersion or variability refers not to the variable directly, but rather to its mean. When we compute the variance or the standard deviation, we are concerned with the spread of the distribution of the variable. But
imagine that we take many, many samples from a population and compute a mean for each sample. We would end up with a sample of means from the population rather than simply a sample of observations. We could then compute a mean of these means, or an overall mean, which should reflect pretty accurately – assuming we do a good job of drawing the samples – the actual mean of the population of observations. Nonetheless, these numerous means will also have a distribution. It is possible to plot these means to see if they follow a normal distribution. In fact, an important theorem from mathematical statistics states that the sample means follow a normal distribution even if they come from a non-normally distributed variable in the population (see Chapter 3). This is a very valuable finding because it allows us to make important claims about the linear regression model. We shall learn about these claims in later chapters.

Our concern here is not whether the means are normally distributed, at least not directly. Rather, we need to consider a measure of the dispersion of these means. Statistical theory suggests that a good measure (estimate) of dispersion is the standard error of the mean. It is computed using the sample standard deviation as

\[
\text{se(mean)} = \frac{s}{\sqrt{n}}
\]

Standard errors are very useful in linear regression analysis. We shall see later that there is another type of standard error – known as the standard error of the slope coefficient – which we use to make inferences about the regression model.

**Standardizing a Variable**

One of the difficult issues when we are presented with different continuous variables is that they are rarely measured in the same units. Education is measured in years, income is measured in dollars, body weight is measured in pounds or kilograms, and food intake is measured in kilocalories. It is convenient to have a way to adjust variables so their measurement units are similar. You might recall that this is one of the purposes of z-scores. Assuming we have a normally
distributed set of variables, we may transform them into \( z \)-scores so they are comparable. A \( z \)-score transformation uses the following formula:

\[
    z - \text{score} = \frac{(x_i - \overline{x})}{s}
\]

Each observation of a variable is put through this formula to yield \( z \)-scores, or what are commonly known as **standardized values**. The unit of measurement for \( z \)-scores is standard deviations. The mean of a set of \( z \)-scores is zero, whereas its standard deviation is one.

You may remember that \( z \)-scores are used to determine what percentage of a distribution falls a certain distance from the mean. For example, 95% of the observations from a normal distribution fall within 1.96 standard deviations of the mean. This translates into 95% of the observations using standardized values falling within 1.96 \( z \)-scores of the mean. With a slight modification, this phenomenon is helpful when we wish to make inferences from the results of the linear regression model to the population. The plot of \( z \)-scores from a normally distributed variable is known as the **standard normal distribution**. As mentioned earlier, one of the principal advantages of computing \( z \)-scores is that they provide a tool for comparing variables that are measured in different units; this will come in handy as we learn about linear regression models. Of course, we must be intuitively familiar with standard deviations to be able to make these comparisons.

**Covariance and Correlation**

Our next task involves moving from one variable to two variables. An important use of statistics is to consider the association or relationship between two variables. As mentioned earlier, an interesting question we might ask is whether two variables shift or change together. To give an obvious and not-very-interesting example, is it fair to say that height and weight shift together? Are taller people, on average, heavier than shorter people? The answer, again on average, is most certainly yes. In statistical language, we say that height and weight **covary** or are **correlated**. The two measures most commonly used to assess the association between two continuous variables are, not surprisingly, the **covariance** and the **correlation**. To be precise, the
correlation used most often is the Pearson’s product-moment correlation (there are actually many types of correlations; the type attributed to the statistician Karl Pearson is the most common).

A covariance is a measure of the joint variation of two continuous variables. In less technical terms, we claim that two variables covary when there is a high probability that large values of one are accompanied by large or small values of the other. For instance, height and weight covary because large values of one tend to accompany large values of the other in a population or in most samples. This is not a perfect association because there is clearly substantial variation in heights and weights among people. The equation for the covariance is

\[ \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} \]

The equation computes deviations from the means of both variables, multiplies them, adds up these products for each observation, and then divides this sum by the sample size minus one. Don’t forget that this implies the x’s and y’s come from the same unit, whether it is a person, place, or thing.

One of the problems with the covariance is that it depends on the measurement scheme of both variables. It would be helpful to have a measure of association that did not depend on these measurement units, but rather offered a way to compare various associations of different sets of variables. The correlation coefficient accomplishes this task. Among the several formulas we might use to compute the correlation, the following equations are perhaps the most intuitive:

\[ \text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}[x] \times \text{Var}[y]}} \]

\[ \text{Corr}(x, y) = \frac{\sum (z_x)(z_y)}{n-1} \]

The first equation shows that the correlation is simply the covariance divided by a joint measure of variability: The variances of each variable multiplied, with the
square root of this quantity representing what we might call the joint or pooled standard deviation. The second equation shows the relationship between z-scores and correlations. We might even say, without violating too many tenets of the statistical community, that the correlation is a standardized measure of association.

A couple of interesting properties of correlations are (1) they always fall between –1 and +1, with positive numbers indicating a positive association and negative numbers indicating a negative association (as one variable increases the other tends to decrease). A correlation of zero means there is no statistical association, at least not one that can be measured assuming a straight line association, between the two variables. (2) The correlation is unchanged if we add a constant to the values of the variables or if we multiple the values by some constant number. However, these constants must have the same sign, negative or positive.

As mentioned earlier, there are several other types of correlations (or what we refer to generally as measures of association) in addition to Pearson’s. For instance, a Spearman’s correlation is based on the ranks of the values of variables, rather than the actual values. Similar to the median when compared to the mean, it is less sensitive to extreme values. There are also various measures of association designed for categorical variables, such as gamma, Cramer’s V, lambda, eta, and odds ratios. Odds ratios, in particular, are popular for estimating the association between two binary (two-category) variables. We shall learn more about odds ratios in Chapter 13.

**Comparing Means from Two Samples**

Another important topic that we’ll discuss before moving into linear regression analysis involves comparing means from two distributions. Of course, we may compare many statistics from distributions, including standard deviations, correlations, and standard errors, but an important issue in applied statistics is determining whether the mean from one sample (or subsample) is different from the mean of another sample (or subsample). For example, I may wish to know whether the mean income of adults from Salt Lake City, UT, is higher than the mean income of adults from St. George, UT. If I have good samples of adults from these two cities, I can consider a couple of things. First, I can take the difference
between the means. Let’s say that the average annual income among a Salt Lake City sample is $35,000 and the average annual income among a St. George sample is $32,500. It appears as though the average income in Salt Lake City is higher. But we must consider something else: We have two samples, so we must consider the possibility of sampling error. Our samples likely have different means than the true population mean, so we should take this into account. A t-test is designed to take these issues into account by, first, taking the difference between the two means and, second, by considering the sampling error with what is known as the pooled standard deviation. This provides an estimate of the overall variability in the means.

The name t-test is used because the t-value that results from the test follows what is termed a Student’s t distribution. This distribution looks a lot like the normal distribution; in fact, it is almost indistinguishable when the sample size is greater than 50. At smaller sample sizes the t-distribution has fatter tails and is a bit flatter in the middle than the normal distribution.

As mentioned earlier, the t-test has two components: The difference between the means and an estimate of the pooled standard deviation. The following equation shows the form the t-test takes:

\[ t = \frac{\bar{x} - \bar{y}}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \]

where \( s_p \) = \( \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2) - 2}} \)

The \( s_p \) in the above equations is the pooled standard deviation. The \( n \)'s are the sample sizes and the \( s^2 \)'s are the variances for the two groups. A key assumption that this type of t-test makes is that the variances are equal for the two groups represented by the means. Some researchers are uncomfortable making such an assumption, so they use the following test, which is known as Welch’s t-test:

\[ t' = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\text{Var}[x]}{n_1} + \frac{\text{Var}[y]}{n_2}}} \]
Unfortunately, we must use special tables if we wish to compute this value by hand and determine the probability that there is a difference between the two means. Fortunately, though, many statistical software packages provide both types of mean comparison tests, along with another test that is designed to show whether we should use the standard $t$-test or Welch’s $t$-test.

An important assumption that we are forced to make if we wish to use these mean-comparison procedures is that the variables follow a normal distribution. The $t$-test, for example, does not provide accurate results if the variable from either sample does not follow a normal distribution. There are other tests, such as those designed to compare ranks or medians (e.g., Wilcoxon-Mann-Whitney test), which are appropriate for non-normal variables.

There are many other types of comparisons that we might wish to make. Clearly, comparing two means does not exhaust our interest. Suppose we wish to compare three means, four means, or even ten means. We might, for instance, have samples of incomes from adults in Salt Lake City, UT, St. George, UT, Reno, NV, Carson City, NV, and Boise, ID. We may use ANOVA procedures to compare means that are drawn from multiple samples. Using multiple comparison procedures, we may also determine if one of the means is significantly different from one of the others. Books that describe ANOVA techniques provide an overview of these procedures. As we shall learn in subsequent chapters, we may also use linear regression analysis to compute and compare means for different groups that are part of a larger sample.

**Samples, Inferences, Significance Tests, and Confidence Intervals**

The last substantive topic to cover involves the core of inferential statistics: How do we know that what we found actually reflects what is occurring in the population? The cynical – but perhaps most honest – answer is that we never know if what we found says anything accurate about the population. After all, statistics has been called the science of uncertainty for good reason. We can only offer degrees of confidence that our results reflect characteristics of the population. But what do we mean by a *population*? Populations may be divided into *target* populations and *study* populations. Target populations are the group about which you wish to learn something. Since this might include a group in the future ("I
wish to know the risk of heart attacks at 60-years-old among people who are now 40-years-old”), we typically find a population that closely resembles the target population; this is the study population. There are clearly many types of populations. For instance, we might be interested in the population of sea lions off the coast of San Diego in July 2005; the population of poodles in New York City; or the population of voters in Massachusetts during the 2006 election. Yet quite a few people, when they hear the term population used in statistics, assume it means the U.S. population, the world’s population, or some other extremely large group.

A sample is a set of items selected from the population. There are many books and articles on the various types of samples that one might draw from a population. The most commonly described in elementary statistics is the simple random sample. This means that each item in or member of the population has an equal chance of being in the sample. There are also clustered samples, stratified samples, and many other types. Most of the classic theoretical work that has gone into inferential statistics is based on the simple random sample.

It should be obvious by now that one of the valuable results of statistics is the ability to say something about a population from a sample. But recall the lesson we learned when discussing the standard error of the mean: We usually take only one sample from a population, but we could conceivably draw many. Therefore, any sample statistic we compute or test we run must consider the uncertainty involved in sampling. The solution to this dilemma of uncertainty has been the frequent use of standard errors for test statistics, including the mean, the standard deviation, correlations, and medians. As we will see in Chapter 2, there is also a standard error for slope coefficients in linear regression models.

These standard errors may be thought of as quantitative estimates of the uncertainty involved in test statistics. They are normally used in one of two ways. First, recall from elementary statistics that when we use a t-test, we compare the t-value to a table of p-values. All else being equal, a larger t-value equates to a smaller p-value. This approach is known generally as significance testing because we wish to determine if our results are significantly different from some other possible result. It is important to note, though, that the term significant does not mean important. Rather, it was originally used to mean that the results signified or showed something (see David Salsburg (2001), The Lady Tasting Tea, New York: Owl
We often confuse or mislead when we claim that a significance test demonstrates that we have found something special.

Showing where $p$-values come from is beyond the scope of this presentation. It is perhaps simpler to provide a general interpretation. Suppose we find through a $t$-test the following when comparing the mean income levels in Salt Lake City ($n = 50$) and St. George, UT ($n = 50$):

$$t = \frac{35,000 - 32,500}{6,250 \times \sqrt{\frac{1}{50} + \frac{1}{50}}} = \frac{2,500}{1,250} = 2.0$$

If we look up a table of $t$-values (available on-line or in most statistics textbooks), we find, using a sample size of 100, that a $t$-value of 2.0 corresponds to a $p$-value of approximately 0.975 (the precise number is 0.9759). This leaves approximately 0.025 of the area under the curve that represents the $t$-distribution. One way to interpret this $p$-value is with the following long-winded statement:

Assuming we took many, many samples from the population of adults in Salt Lake City and St. George, and there was actually no difference in mean income in this population, we would expect to find a difference of $2,500 or something larger (in Salt Lake City) only 2.5 times, on average, out of every 100 samples we drew.

If you remember using null and alternative hypotheses, such a statement may sound familiar. In fact, we can translate the above inquiry into the following hypotheses:

$$H_0: \text{ Mean Income, Salt Lake City } = \text{ Mean Income, St. George}$$
$$H_a: \text{ Mean Income, Salt Lake City } > \text{ Mean Income, St. George}$$

Astute readers may notice that we have set up a one-tailed significance test. A two-tailed significance test suggests a difference about five percent of the time, or only five times out of every one-hundred samples. We’ll return to the interpretation of $p$-values in Chapter 2.
The second way that standard errors are used is to compute confidence intervals (CIs). There are many applied statisticians who prefer confidence intervals because they provide a range of values within which some measure is likely to fall. Here we contrast a point estimate and an interval estimate. Means and correlations are examples of point estimates: They are single numbers computed from the sample that estimate population values. An interval estimate provides a range of values that (presumably) contains the actual population value. A confidence interval offers a range of possible or plausible values. Those who prefer confidence intervals argue that they provide a better representation of the uncertainty inherent in statistical analyses.

The general formula for a confidence interval is

\[
\text{Point estimate} \pm \left[ \text{(confidence level)} \times \text{(standard error)} \right]
\]

The confidence level represents the percentage of the time, based on a \( z \)-statistic or a \( t \)-statistic, you wish to be able to say that the interval includes the point estimate. For example, assume we’ve collected data on violent crime rates from a representative sample of 100 cities in the United States. We wish to estimate a suitable range of values for the mean violent crime rate in the population. Our sample yields a mean of 104.9 with a standard deviation of 23.1. The 95% confidence interval is computed as

\[
95\% \, \text{CI} = 104.9 \pm \left[ 1.96 \times \left( \frac{23.1}{\sqrt{100}} \right) \right] = \{100.4, 109.4\}
\]

The value of 1.96 for the confidence level comes from a table of standard normal values, or \( z \)-values. It corresponds to a \( p \)-value of 0.05 (two-tailed test). The standard error formula was presented earlier in this chapter.

How do we interpret the interval of 100.4 – 109.4? There are two ways that are generally used:

1. Given a sample mean of 104.9, we are 95% confident that the population mean of violent crime rates falls in the interval of 100.4 and 109.4.
LINEAR REGRESSION ANALYSIS

(2) If we were to draw many samples from the population of cities in the U.S., and we claimed that the population mean fell within the interval of 100.4 – 109.4, we would accurate about 95% of the time.

As we shall see in subsequent chapters it is also possible to construct confidence intervals for point estimates from a linear regression model.

An Example Using SPSS

The file nations.sav is an SPSS data set that contains some old (from the early 1970s) economic data from seven nations. The variables include Public Expenditures (expend), Economic Openness (econopen), and Percent Labor Union (perlabor). We’ll use SPSS to show some of the test statistics we’ve discussed in this chapter. After opening the data set, find the following drop-down menu: Analyze-Descriptive Statistics-Descriptives. Next, move the variable Public Expenditures into the variable box. Click Options (bottom right corner) and place a check mark in the boxes labeled Mean, Std. deviation, Variance, and S.E. mean. Then click Continue. After it returns you to the Descriptive Statistics window, click OK. (If you wish to see how SPSS uses syntax files, try clicking Paste and it will bring up a new screen with the coding.) SPSS’s output screen should show the following table:

<table>
<thead>
<tr>
<th>Statistics</th>
<th>N</th>
<th>Mean</th>
<th>Std.</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistic</td>
<td>Statistic</td>
<td>Std. Error</td>
<td>Statistic</td>
<td>Statistic</td>
</tr>
<tr>
<td>Public Expenditures</td>
<td>7</td>
<td>43.6143</td>
<td>2.62981</td>
<td>6.95783</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As an exercise, see if you can compute the standard error of the mean from the mean and the standard deviation (abbreviated Std.) from the variance listed in the table. Then, return to SPSS and use Analyze-Correlate-Bivariate to estimate the correlation between Public Expenditures and Economic Openness (make sure the Pearson box is checked). Can you figure out how to estimate the covariance between these two variables? (Hint: use Options on the Bivariate Correlations screen.) You should find that the correlation is 0.748 and the covariance is 37.065.
If you are interested in confidence intervals for the mean, try using *Analyze-Descriptive Statistics-Explore* and place Public Expenditures in the *Dependent List* box. You should obtain 95% CIs of 37.18 and 50.05. You may also wish to practice using *t*-tests in SPSS.
2 Simple Linear Regression

Analysis

As mentioned in the Preface, a general goal of regression analysis is to estimate the association between one or more independent variables and a single dependent variable. As the names imply, a dependent variable is presumed to depend or be systematically predicted by the independent variables. But the independent variables are thought to independently affect the dependent variable. We shall see in later chapters that using the term independent can be misleading because these variables may, if the model is set up correctly, depend on one another. Perhaps this is why many researchers prefer to use the names predictor and outcome variables, explanatory and response variables, exogenous and endogenous variables, or some other terms. The outcome, response, or endogenous variable is synonymous with the dependent variable.

In general, the goals of a regression analysis are to predict or explain differences in values of the dependent variable with information about the values of the independent variables. We are primarily interested in the following issues:

1. The form of the relationship among the dependent and independent variables, or what the equation that represents the relationship looks like.
2. The direction and strength of the relationships. As we shall learn, these are based on the valence and size of the slope coefficients.
3. Which independent variables are important and which are not. This issue is based on comparing the size of the slope coefficients (which has some problems) and on the $p$-values (or confidence intervals).
4. Predicting a value or set of values of the dependent variable for a given set of values of the independent variables.

We'll begin by considering an elementary situation: The simple linear regression model, or what some people refer to as bivariate regression. The term simple is used not to indicate that our research question is uninteresting or crude, but rather that we have only one independent and one dependent variable. But,
you might ask, if there are only two variables why not use the Pearson’s correlation coefficient to estimate the relationship? This is certainly a reasonable step; we can learn something about the relationship between two variables by considering the correlation (assuming they are continuous and normally distributed). However, the difference here is primarily conceptual: We think, perhaps because of a particular theory or model, that one variable influences another, or we have control over one variable and therefore wish to see its effect on another (e.g., we control the amount of fertilizer and see how a particular species of tomato plant grows).

The dependent variable is usually labeled as Y and the independent variable is labeled X (for now, let’s assume we’re interested in population values). To help you understand the regression model, recall your days in pre-algebra or algebra class when you were asked to plot points on two axes. You probably labeled the vertical axis as Y and the horizontal axis as X. You then used an equation such as $Y = mX + b$ to either represent the systematic way the points fell on the graph, or to decide where to put points on the graph. The equation provided a kind of map that told you where to place objects. The $m$ represented the slope and the $b$ represented the intercept, or the point at which the line crossed the Y axis. If you can recall this exercise in graphing, then simple linear regression is easy to understand. Consider the following graph:
What is its slope and intercept? Since we don’t know the units of measurement, we cannot tell. However, we do know that the slope is a positive number and the intercept is a positive number (notice the line crosses the Y axis at a positive value of Y). We also know that it has a very high positive correlation because the points are very close to the line. Let’s say it is represented by the following equation:

\[ Y = 2X + 1.5 \]

This suggests that as the X values increase by one unit the Y values increase by two units. Perhaps you recall a definition of the slope as \( \frac{\text{rise}}{\text{run}} \); which is a shorthand way of saying that as the points rise a certain number of units along the Y axis they also run a certain number of units along the X axis.

Unfortunately, real data in almost any research application never follow a straight line relationship. Nonetheless, in a simple linear regression analysis we may represent the association between the variables using a straight line. Whether a straight line is an accurate representation is a question we should always ask ourselves. Consider the graph on the next page. A similar graph may be computed in SPSS with the nations.sav data by using Graphs-Scatterplot-Simple and placing Public Expenditures on the Y axis and Percent Labor Union on the X axis.

The association between the two variables shown in the scatterplot is positive, but notice that the points do not fall on the line that represents the slope. This shows a statistical relationship, which is probabilistic, whereas the previous graph shows a mathematical relationship, which is deterministic. Therefore, we now must say that, on average, public expenditures increase as percent labor union increases. The next question to ask is “On average, how many units increase in Labor Union are associated with how many units increase in Public Expenditures?” This is a fundamental question that we try to answer with a linear regression model.
In a linear regression model, rather than \( Y = mX + b \), we use the following form of the equation:

\[
Y_i = \alpha + \beta_i X_i
\]

The Greek letter alpha represents the intercept and the Greek letter beta represents the slope. These two terms are sometimes labeled *parameters*; however, in statistical analysis parameters are usually defined as properties of a population, in contrast to properties of samples, which are labeled *statistics*. For the real stickler, perhaps we should not even use Greek letters in the equation since, technically speaking, they imply parameters and thus apply to the population. Hence, an alternative representation of the linear regression equation is

\[
y_i = a + b_i x_i
\]

Here we have used lower-case Roman letters throughout the equation to represent a linear regression model that uses data from a sample. You will also notice the subscripted \( i \); as before, this represents the fact that we have
individual observations from the sample. Since another common approach is to preserve Greek symbols but use lower case Roman letters to represent the variables, here is another way the equation is represented:

\[ y_i = \beta_0 + \beta_1 x_i \]

This equation uses beta to represent the intercept and the slope, with a zero subscript used for the intercept. However, the preference in this presentation is to use alpha to indicate the intercept (i.e., \( y_i = \alpha + \beta_1 x_i \)).

Another way of thinking about the slope, whether represented by the Greek letter beta or the letter b, is that it represents the change in \( y \) over the change in \( x \). You may see this referred to as \( \frac{\Delta y}{\Delta x} \), in which the Greek letter delta is used to denote change. This is just another way of saying rise over run. The intercept may also be defined as the predicted value of \( y \) when the value of \( x \) is zero.

Perhaps you’ve noticed a problem with the equations we’ve written so far. A hint is that we used the term probabilistic to describe the statistical relationship. If you’re not sure what this all means, revisit the scatterplot between Public Expenditures and Percent Labor Union. Notice that the points clearly do not fall on the straight line. Take almost any two variables used in the social and behavioral sciences and they will fail to line up in a scatterplot. This presents a problem for those who want their models exact. Our equations thus far call for exactness, which misrepresents the actual relationship between the variables. Our straight line is, at best, an approximation of the actual linear relationship between \( x \) and \( y \). Given a small data set, it is not difficult to draw a straight line that tries to match this relationship. Nonetheless, we must revise our linear regression equation to consider the uncertainty of the straight line. Statisticians have therefore introduced what is known as the error term into the model:

\[ y_i = \alpha + \beta_1 x_i + \epsilon_i \]

The Greek letter epsilon represents the uncertainty in predicting the dependent variable with the independent variable. Another way of thinking about the error
term is that it represents how far away the individual \( y \) values are from the true mean value of \( Y \) for given values of \( x \). Now, this last sentence is loaded with assumptions. We use, for instance, the term \textit{true mean value} of \( Y \). What are we trying to say? Well, think about our sample. If we have done a good job, our sample observations should represent some group from the population. For instance, if we’ve sampled from adults in the U.S., some of our sample members should represent 25-30-year-olds. We assume that these sample members are a good representation of other 25-30-year-olds. Hence, their values on a characteristic such as education should provide a good estimate of mean education among those in this age group in the population. Suppose we wish to use parental education (\( x \)) to predict education (\( y \)) among this age group. Using the linear regression model, we assume that the error terms represent the distance from the points on the graph to the actual mean of \( Y \), or the population mean.

Unfortunately, we usually cannot test – at least not fully – the various assumptions we make when using the linear regression model. Since we normally do not have access to the population parameters, we can only roughly test properties of the various assumptions we make when using linear regression. These assumptions include the following:

1. For any value of \( X \), \( Y \) is normally distributed and the variance of \( Y \) is the same for all possible values of \( X \). (Notice that we use upper-case letters to show that we are referring to the population.)

2. The mean value of \( Y \) is a straight line function of \( X \). In other words, \( Y \) and \( X \) have a linear relationship.

3. The \( Y \) values are statistically independent of one another. Another way of saying this is that the observations are independent. For example, using the earlier example, we assume that our measures of public expenditures across the seven nations are independent. We usually assume that simple random sampling guarantees independence. (However, we should ask ourselves: Are the economic conditions of these nations likely to be independent?)
4. The error term is a random variable that has a mean equal to zero in the population and constant variance (this is implied by (1)). Symbolically, this is shown as $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. As mentioned in Chapter 1, the wavy line means “distributed as.”

In order to test these assumptions, we should have lots of $x$ values and lots of $y$ values. For example, if we collect data on dozens of countries with, say, 20% labor unions, then we expect the public expenditure values to be normally distributed within this value of the variable labor union. And we expect the errors in predicting $Y$ to have a mean of zero.

This graph provides one way to visualize what is meant by some of these assumptions. We have sets of observations at 20%, 40%, and 60% values of labor union. We assume that the mean public expenditure values for these three subsamples are a good representation of the actual means in the population. We also assume that public expenditures at these three values of the independent variable are normally distributed, that the variances of each distribution are identical, and that the errors we make when we estimate the value of public expenditures have a mean of zero (e.g., the underestimation and overestimation that we make cancel out).
An Example of a Linear Regression Model Using SPSS

Perhaps you are confused at this point, although I hope not. Maybe an example using some data would be helpful. The SPSS data set usdata.sav includes a number of variables from states in the U.S. These data were collected in 1995 and include variables that measure crime rates, suicide rates, and various economic factors. One might argue that these data represent a population, but we’ll treat them as a sample. First, use the scatterplot drop-down menu (Graph-Scatterplot-Simple) to estimate a simple scatterplot with per capita income in $1,000s (perinc) as the independent variable (x) and the number of robberies per 100,000 people (robbrate) as the dependent variable (y). If you double-click the graph and highlight the data points, you can place a regression line in the graph using the icon – labeled “add fit line” – at the top of the window (I’m assuming you’re using SPSS 12.0 or higher). You should see a modestly increasing slope. But notice how the points diverge from the line; only a few are relatively close to it.

Next, we’ll try a simple linear regression model using these variables. In SPSS, a linear regression model is set up using the Analyze-Regression-Linear drop-down menu. The window that opens up has the variables listed in a box on the left and several boxes on the right in which we’ll place the variables of interest. Notice that there is a box labeled Dependent and another labeled Independent(s). We’ll ignore the other boxes, although we’ll learn more about them later. To estimate our model, place robbrate in the Dependent box and perinc in the Independent(s) box. Then click OK. The output screen that opens up should include the table on the next page (you may need to scroll down the SPSS output screen to find it).

Can we translate the numbers in this table into our linear regression equation? Yes, we can, but first we need to find the slope and the intercept. The term coefficients is used in a general sense, although the key columns are labeled Unstandardized Coefficients and Standardized Coefficients. We’ll focus for now on the former column since it provides the information we need. SPSS uses an upper case B to indicate the intercept and slope. It also uses the term Constant to label the intercept. Thus, our equation reads:
SIMPLE LINEAR REGRESSION ANALYSIS

\[ \text{Robberies}(y_i) = -165.2 + \{14.459 \times \text{Per Capita Income}(x_i)\} + \varepsilon_i \]

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-165.181</td>
<td>86.847</td>
<td>-1.902</td>
<td>.063</td>
</tr>
<tr>
<td></td>
<td>14.459</td>
<td>3.907</td>
<td>.475</td>
<td>3.701</td>
</tr>
</tbody>
</table>

*a Dependent Variable: Robberies per 100,000*

The \( y \) and the \( x \) are included to remind you that number of robberies is the dependent variable and per capita income is the independent variable. Revisit the scatterplot between these two variables. The slope is positive. In other words, the line tilts up from left to right. Our slope in the above equation is also positive, but provides even more information about the presumed association between the variables. The intercept in the equation is negative. The SPSS scatterplot does not even show the intercept; it is off the chart. But imagine expanding the graph and you can picture the line crossing the \( y \)-axis.

Recall that we said the slope shows the average number of units increase in \( y \) for each one unit increase in \( x \). This provides a way to interpret the slope in this equation:

Each one unit increase in per capita income is associated with a 14.46 unit increase in the number of robberies per 100,000 people.

Of course, we need to specify how these variables are measured to determine what this statement means. For instance, since per capita income is measured in thousands of dollars and the robberies are measured in offenses per 100,000 people we may say:

Each $1,000 increase in per capita income is associated with 14.46 more robberies per 100,000 people.

Similarly, given our understanding of the intercept, we may say:
The expected number of robberies per 100,000 people when per capita income is zero is –165.2.

One of the problems with this interpretation is that (check the data if you’d like) there is no state with a per capita income of zero. It is not a good idea to estimate values from your sample when no observations have this value. Sometimes, such as the situation here, we come up with ridiculous results if we try to make inferences too far outside our set of data values.

Keep in mind that we’ve used the term “associated” in our interpretations. Many people are uncomfortable with some of the terms used to describe the relationship between variables, especially when using regression models in the social and behavioral sciences. Associated is a safe term, one that implies that as one variable changes the other one also tends to change. It should not be seen as a term denoting causation, though; each $1,000 increase in per capita income does not lead to or cause an increase in the number of robberies. I’m sure you can think up several reasons why these two variables might be associated that have nothing to do with one variable causing the other.

In general, we say that the intercept represents the expected value of Y when X equals zero. The term *expected value* should remind you of the mean. The slope represents the expected difference or change in Y that is associated with each one unit change or difference in X. We are inferring from the sample to the population here, although we don’t have to. The two terms that are frequently used when considering the slope are *difference* and *change*. A problem with using the term *change* is that we don’t necessarily observe changes; we (usually) note only differences in the values of variables.

The graph on the next page shows a picture of the shifts that may be inferred from the linear regression model. With a little imagination, it also demonstrates another useful thing about this type of model: We may come up with predictions of the dependent variable. With a graph we simply find that point along the x-axis that interests us, trace vertically to the regression line, and then trace horizontally to the y-axis to find its value. This is known as the *predicted value* from the model. Graphs can be difficult to use to make precise predictions, however, so we may also use the regression equation to find
predicted values. Suppose we wish to predict the value of the robbery rate for states with per capita incomes of $25,000. We may use the equation to do this:

\[
\text{Robberies} = -165.18 + (14.46 \times 25.0) = 196.32 \text{ per 100,000}
\]

Hence, on average, states with mean per capita income values of $25,000 have approximately 196.32 robberies per 100,000 people. We’ve used here the term “on average.” This is another way of implying that for those states that have per capita incomes of $25,000, we estimate that the mean number of robberies is 196.32. In other words, predicted values are predicted means of the dependent variable for particular values of the independent variable.

One of the questions you should ask yourself is why per capita income is positively associated with the number of robberies in a state. Actually, we should ask these types of questions before we estimate a regression model. Perhaps residents of wealthier states offer better targets for robbers. Could higher income states generate more frustration among those who have little in such a way that they tend to lash out through violent crime? As suggested earlier, could there be a third variable – such as migration patterns or the percent of the population that lives in urban areas – that explains the association between the two variables? At this point, we cannot tell, but we should consistently seek reasonable answers to why two or more variables are
associated. And we should be willing to revise our conclusions as new evidence emerges.

Formulas for Slopes and Intercepts

You should be asking at this point: From where do these slopes and intercepts come? Certainly, we can plot the data – by hand if the sample is small enough – and then guess which straight line best represents the association between two variables. It makes sense to try to draw a line that is as close as possible to all the points. Another way of thinking about this is to try to draw a line that, if we calculated an aggregate measure of distance from the points to the line, will have as small a value as possible. This is precisely what the most common technique for fitting the regression line does. Known as ordinary least squares (OLS) or the principle of least squares, this technique is used so frequently that many people refer to linear regression analysis as ordinary least squares analysis or as an OLS regression model. However, there are many types of estimation routines, such as weighted least squares and maximum likelihood, so there are actually a number of ways we might compute the results of a linear regression analysis. Because it is so widely used, though, we’ll focus on OLS in this presentation.

The main goal of OLS is to obtain the minimum value for the following equation:

\[
\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - (a + \beta_i x_i))^2
\]

The quantity, SSE, is the Sum of Squared Errors or the Residual Sum of Squares. There is also a new symbol that we haven’t seen before: \(\hat{y}\) (also called y-hat). This is the symbol for predicted values of \(y\) based on the model that uses the sample data. The goal of OLS is to minimize the SSE, or make it as close to zero as possible. In fact, if the SSE equals zero, then the straight line fits the data points perfectly; they fall right on the line. In addition, if the points fall directly on the line, the Pearson’s correlation coefficient is one or negative one (depending on whether the association is positive or negative). Another new term is residual: It is a common name for a measure of the distance from the regression line to the actual points, so there are actually many residual values
that result from the model. It is typically represented by $\varepsilon$ to indicate that it is another way of showing the errors in estimation \( \text{residual}_i (e_i) = (y_i - \hat{y}_i) \). As we’ll learn in later chapters, the residuals are used for many purposes in linear regression models.

As mentioned earlier, the straight regression line almost never goes through each data point. Rather, there are discrepancies from the line for two reasons: (1) There is simply variation of various types whenever we measure something; and (2) a straight line association is not appropriate. The former situation is a normal part of just about any linear regression analysis. Although we wish to minimize errors, there is usually natural or random variation because of the intricacies and vicissitudes of behaviors, attitudes, and demographic phenomena. The second situation is more serious; if a straight line is not appropriate we should look for a relationship – referred to as nonlinear – that is appropriate. Chapter 10 provides a discussion of what to do when we find a nonlinear relationship among variables.

But how do we come up with an equation that will minimize the SSE and thus allow us to both fit a straight line and estimate the slope and intercept? Those of you who have taken a calculus course may suspect that, since we are concerned with changes in variables, derivatives might provide the answer. You are correct, although through Cramer’s rule and the methods of calculus some simple formulas have been derived. Among the various alternatives, the following least squares equation for the slope has been shown to be the best at minimizing the SSE:

\[
\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
\]

This equation works very well if the assumptions discussed earlier are met. We’ll have a lot more to say about assumptions in later chapters. The numerator and denominator in this equation should look familiar. The numerator is part of the equation for the covariance (see Chapter 1) and the denominator is the formula for the sum of squares of $x$. Here we use the common $x$-bar and $y$-bar symbols for the means; and $\hat{\beta}_1$ to indicate that we are estimating the slope with the
equation. Let’s think for a moment about what happens as the quantities in the equation change. Suppose that the covariance increases. What happens to the slope if the sum of squares does not also change? Clearly, it increases. What happens as the sum of squares of $x$ increases? The slope decreases. So, all else being equal, as the variation of the independent variable increases, the slope decreases. Typically, we wish to have a large positive or negative slope if we hypothesize that there is an association between two variables.

Now that we have the slope, how do we estimate the intercept? The formula for the intercept is the following:

$$\hat{\alpha} = \bar{y} - \left( \hat{\beta}_1 \times \bar{x} \right)$$

The estimated intercept is computed using the means of the two variables and the estimated slope. We should note at this point that we rarely have to use these equations since programs such as SPSS will conduct the analysis for us. Moreover, programs such as SPSS do not use these equations directly; rather they use matrix routines to speed up the process. It is useful, though, to practice using these equations with a small set of data points.

**Hypothesis Tests for the Slope**

Now that we’ve seen examples of slopes and intercepts, as well as where they come from, we should consider the issue of hypothesis testing. What are the most common hypotheses we wish to examine using a linear regression model? Generally speaking, the most common is that there is a linear association between the independent and the dependent variables. But we should make this more precise. Using your imagination, common sense, understanding of the research on your topic, and perhaps even colleagues’ ideas as you discuss your research plans, you should try to specify the reason there might be an association and then write down the null and alternative hypotheses. Let’s assume that all of these things tell us that per capita income should be positively associated with the number of robberies at the state-level. This means that we think there is a positive slope in a regression model that considers these variables. Our hypotheses are therefore
Although these are reasonable hypotheses, for some reason most researchers who use linear regression models define the hypotheses more generally as saying either the slope is zero or the slope is not zero. This, of course, means that the alternative hypothesis is that the slope is simply different from zero and that the null hypothesis is that the slope is zero: \( H_a: \beta_1 \neq 0 \) vs. \( H_0: \beta_1 = 0 \).

It may appear obvious that we can simply look at the slope coefficient and determine whether or not it is zero. But, let’s not forget a crucial issue. Remember that we are assuming that we have a sample that represents some target population. Ours is only one sample among many possible samples that might be drawn from the population. Perhaps we were just lucky (or unlucky, depending on how you look at things) and found the one sample among countless others that had a positive slope. How can we be confident that our slope does not fall prey to such an event? Although we can never be absolutely certain, we may use a significance test or confidence intervals to estimate whether the slope in the population is likely to be zero.

It is now time to return to standard errors. In Chapter 1 we learned how to compute the standard error of the mean and how this statistic is interpreted. There is also a statistic known as the *standard error of the slope* that is interpreted in a similar way. That is, the standard error of the slope estimates the variability of the estimated slopes that might be computed if we were to draw many, many samples. Let’s say we have a population in which the association between age and alcohol consumption is zero. That is, the correlation is zero and the population-based linear regression slope is zero. Therefore, the null hypothesis of a zero slope is true \( H_0: \beta_1 = 0 \). If we were to draw many samples, can we infer what percentage of the slopes from these samples would fall a certain distance from the true mean slope of zero? It turns out we can because the linear regression slopes from samples, if many samples are drawn randomly, follow a \( t \)-distribution. We therefore know, for instance, that in a sample of, say, 1000, we would expect only about five percent of the slopes to fall more than 1.96 \( t \)-
values (similar to z-values in the standard normal distribution) from the true mean of zero. We should only occasionally find a slope this far from zero if the null hypothesis is true. If the null hypothesis is not true (or if we may reject it, to use a common phrase), then a slope this many t-values from zero should be relatively common. So, the question becomes, how can we compute these t-values?

The t-value in a linear regression model is computed by dividing the estimated slope by the estimated standard error. We’ve already seen how to compute a slope in a simple linear regression model. Here is the formula for the standard error:

\[
se(\beta_1) = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2 / n - 2}{\sum(x_i - \bar{x})^2}} = \sqrt{\frac{SSE/n - 2}{SS[x]}}
\]

This equation provides the standard error formula for the slope in a simple linear regression model only. As with the slope equation, the standard error equation includes some familiar elements. In the numerator, for instance, we see the sum of squared errors (SSE) and the sample size (n); in the denominator we find the sum of squares of x. As the sum of squares of x gets larger, the standard error becomes relatively smaller. As the SSE becomes larger, the standard error also becomes larger. This should not be surprising if we think about our scatterplot: Larger SSEs indicate more variation about the regression line. Therefore, our uncertainty about whether we have a good prediction of the population slope should also increase. But what happens to the standard error as the sample size increases? Working through the algebra implied by the equation, we can see that the standard error decreases with larger samples. Again, this should make sense: The larger the sample, the more it is like the population from which it was drawn, and the more confidence we should have about the sample slope. In fact, some observers complain because if we have a larger enough sample, even one that wasn’t drawn randomly, we can make claims of certainty that are not justifiable.

As mentioned in the beginning of the last paragraph, a t-value in a linear regression model is computed by dividing the slope by the standard error:
Occasionally, you might see one beta value minus another in the t-value equation (e.g., $\hat{\beta}_1 - \beta_0$). Although it may be used in such a manner, it is much more common to assume that we wish to compare the slope from the regression model to the slope implied by the null hypothesis, which is typically that the slope is zero in the population. Each t-value is associated with a p-value that depends on the sample size. This is the basis for using significance tests to determine the importance of the slope. Fortunately, we do not need to use t-tables to determine the p-value. Statistical software provides the associated p-values.

Recall that we used SPSS to estimate a model with robberies per 100,000 as the dependent variable and per capita income as the independent variable. Here is the table that SPSS produced.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>-165.181</td>
<td>86.847</td>
<td>-1.902</td>
<td>.063</td>
</tr>
<tr>
<td>Per Capita Income</td>
<td>14.459</td>
<td>3.907</td>
<td>.475</td>
<td>3.701</td>
</tr>
</tbody>
</table>

a Dependent Variable: Robberies per 100,000

We have already learned to interpret the slope and intercept in this model. But notice that SPSS also computes the standard errors (listed as Std. Error) of the slope and intercept, the t-values (under the column labeled t), and the p-values (under the column labeled Sig.; an abbreviation for significance). The t-value for the slope is $14.459/3.907 = 3.701$. A t-value of this magnitude from a sample of 49 observations has a p-value of less than 0.001. SPSS rounds p-values to three decimal places. If you wish to see a more precise estimate double-click the number in the SPSS output. You should find that the p-value is about 0.00056.
Although it is rarely used, there is also a \( p \)-value for the intercept. It compares the intercept to the null hypothesis value of zero.

We usually don’t have need to interpret the standard error of the slope or intercept directly, nor do we interpret the \( t \)-value. Rather, we may focus on the \( p \)-value. Unfortunately, \( p \)-values are frequently misused. You might recall that most statistical analyses use a threshold or rule-of-thumb that states that a \( p \)-value of 0.05 or less means that the statistic is somehow important. Actually, this is an arbitrary value that has, for various reasons, become widely accepted. In fact, when one finds a \( p \)-value of 0.05 or less, it is common to claim that the slope (or other statistic) is *statistically significant*. But what precisely does the \( p \)-value mean? Can we come up with an interpretation? The answer is yes, but we must again assume that the sample slope says something reasonable about the true population slope. Suppose, for instance, that we find a \( p \)-value of 0.03. The following interpretation is reasonable (even if it is not entirely persuasive):

If the actual slope in the population is zero, we would expect to find a sample slope this far or farther from zero only three times out of every 100 samples, if we were to draw many samples from the population.

You should notice that this interpretation construes the \( p \)-value as a probability. Another way of thinking about this \( p \)-value is that it suggests (but, of course, does not prove) that, if the null hypothesis is true in the population, we would reject it only about three percent of the time given a slope this far or farther from zero. Once again, we are assuming that we could draw many samples to reach this conclusion.

In the model of robbery and per capita income, the \( p \)-value suggests that if the population slope is zero, we would expect to find a slope of 14.459 or some value farther from zero less than one time out of every one-thousand samples drawn. This is clearly a small \( p \)-value and offers a large degree of confidence that there is a non-zero statistical association between the number of robberies per 100,000 and per capita income.

However, note that we use the phrase “something farther from zero.” This implies that we are using what we labeled in Chapter 1 as a *two-tailed significance*
Imagine a bell-shaped curve with the value of zero as its middle point. We may then picture the area under the curve as representing the frequency of slopes from a large number of samples. The mean of zero represents the null hypothesis. Hence, this bell-shaped curve signifies the distribution of slopes for the null hypothesis. A certain percentage of the total area of the curve lies in the tails of this distribution. For instance, if we have a large sample (say, 200 or more), then we know that 2.5% of the total area falls in each tail marked by the values ±1.96. A two-tailed $p$-value of 0.05 implies a $t$-value of 1.96 (for large samples). When we rely on programs such as SPSS for $p$-values, they typically provide two-tailed tests: They assume that the slope can be in either of the tail areas, above or below the mean. In other words, a negative slope and a positive slope are equally valid for rejecting the null hypothesis as long as their $t$-values are sufficiently large (in absolute value).

But suppose our alternative hypothesis indicates direction, such as that the slope is greater than zero. We are then justified in using a one-tailed test: We are concerned only with the conceivable slopes that fall in the upper tail of the $t$-distribution. The 5% area that most people use falls above the mean of zero; in large samples this has a threshold at 1.64 $t$-units above the mean. How can we use a one-tailed test when programs such as SPSS assume we want two-tailed tests? We may take the $p$-value provided by SPSS and divide it in half in order to reach a conclusion about the statistical significance of the slope. This can quickly become confusing, especially as we think about taking only one sample and then trying to infer something about a population. Perhaps it is easiest to simply remember that we wish to have slopes far from zero if we want to conclude that there is a non-zero linear association between an independent and a dependent variable.

As with other statistical tests that compare a value computed from the data to a hypothetical value from a null hypothesis, there are many researchers who argue that significance tests that use only $p$-values are misleading: They deceive us into thinking that our estimates are precise and fail to account for the uncertainty that is part of any statistical model. Many recommend that we should therefore construct confidence intervals (CIs) for slopes, much like they are constructed for means. Since we have point estimates and standard errors,
constructing confidence intervals for linear regression slopes is relatively simple. In fact, the same general formula we used in Chapter 1 is applicable:

\[
\text{Point estimate} \pm [(\text{confidence level}) \times (\text{standard error})]
\]

As discussed in Chapter 1, the confidence level represents the percentage of the time we wish to be able to say that the interval includes the point estimate. Given the special nature that the 0.05 \( p \)-value has taken on, it is not surprising that a 95% confidence interval is the most common choice. If we wish to use a two-tailed test (again, the most common choice), then, assuming a large sample, the value of 1.96 is used for the confidence level. For smaller samples, you should consult a table of \( t \)-values or ask the statistical software to compute CIs for the slopes. Here is an example from our linear regression model:

\[
14.459 \pm [2.009 \times 3.907] = \{6.61, 22.308\}
\]

In this example, since we have a sample size of 49, we use a confidence level of 2.009. This is found in a table of \( t \)-values and is associated with 2.5% of the area in each tail of the distribution. In other words, our confidence level is 95%. The interpretation of this confidence interval is similar to the interpretation offered in Chapter 1: Given a sample slope of 14.459, we are 95% confident that the population slope representing the association between per capita income and the number of robberies falls in the interval bounded by 6.61 and 22.308. (Note: SPSS will compute confidence intervals for slopes if you check the appropriate box in the Statistics window. The resulting intervals are slightly different than those we just computed because of rounding error.) If we wish to obtain a higher degree of confidence – for example, if we want to be 99% confident about where the population slope falls – then the interval is wider because the \( t \)-value used in the equation is larger. For example, a confidence level of 99% with a sample size of 49 is represented by a \( t \)-value of approximately 2.68, so the CIs are \{3.99, 24.93\}.

An interesting relationship between \( p \)-values and CIs occurs when using the same \( t \)-value for determining the threshold and for the confidence level. Suppose we rely on 1.96 to compute a \( p \)-value of 0.05 (two-tailed test) and for our
confidence level. Then, if the slope divided by the standard error – i.e., the \( t \)-value from our regression model – is less than 1.96, the \( p \)-value is greater than 0.05. In this situation, the CIs that use 1.96 to represent the confidence level include zero. Thus, we would conclude that the null hypothesis cannot be rejected. In other words, our conclusions about statistical significance are generally the same whether we use \( p \)-values or confidence intervals. The main advantage of confidence intervals is that they are a better reflection of the uncertainty that is a fundamental part of statistical modeling.
3 Multiple Linear Regression Analysis

We have now learned about some of the basic aspects of linear regression analysis, in particular where the slopes and intercepts come from, how to interpret them, and how to determine whether or not they are meaningful. However, we considered only the simplest case, that of a model with only one independent variable. We are now prepared to consider linear regression models that include more than one independent variable. We'll continue to consider many of the same issues as with the simpler model. But, as we shall learn, a key difference between simple linear regression models and multiple linear regression models is in the interpretation of the coefficients.

A simple linear regression model is usually not very interesting. For instance, we saw in the last chapter that there is a positive association between per capita income and the number of robberies that occurred among states in the U.S. in 1995. But you don’t need to be a criminologist to realize that there might be other factors that are related to the frequency of crimes. Some of these factors include the unemployment rate, the imprisonment rate, the percent living in poverty, the migration rate, and the age structure; we could go on and on. The main point is that there are usually many potential independent variables used to predict a dependent variable. Theories of various stripes should be used to guide the selection of variables; yet it is rare that these theories can be narrowed down to only one independent variable.

A key reason to include other variables in the model is because they may account for the association between one of the independent variables and the dependent variable. This issue is known as confounding, since we say that if variable $x_2$ accounts for the association between $x_1$ and $y$, then $x_2$ is said to confound the association. For example, suppose we are interested in the association between the frequency of purchasing lighters and the rate of lung disease across a sample of U.S. counties. We would likely find a positive association between these two variables. Would we therefore conclude that purchasing lighters leads to lung
disease? Probably not, especially when we realize that cigarette smoking is related to both purchasing lighters and lung disease. We say that cigarette smoking is a confounding variable and that the association between lighters purchased and lung disease is spurious (see Chapter 7). Hence, smoking should be included in a regression model that predicts lung disease, especially if the model also includes the frequency of lighter purchases. (Actually it should always be used in models predicting lung disease.)

An Example of a Multiple Linear Regression Model

Understanding the complex way that multiple linear regression models estimate slopes and other statistics can be difficult. To estimate this model we continue to use ordinary least squares (OLS), but the formulas become more complex as we add independent variables. It is perhaps best for now to see one of these models and then discuss how to use it. We’ll use the SPSS file usdata.sav once more, but consider some other variables. Specifically, the data set has a measure of each state’s overall violent crimes per 100,000 (violrate), unemployment rate (unemprate), and gross state product in $100,000s (gsprod), a measure of overall economic productivity. Before estimating a multiple linear regression model, it is useful to check the correlations among the variables to see, in a rough sense, the strength of their associations. SPSS returns the output shown in the table on the next page when we use the following drop-down menu: Analyze-Correlate-Bivariate; and place the three variables in the Variables box.

There are significant and positive correlations among two pairs of variables: Violent crimes and unemployment rate \( r = 0.413; p = .003 \) and violent crimes and gross state product \( r = 0.478; p < .001 \) (Recall that SPSS rounds the \( p \)-values to the third decimal place. This means that very small \( p \)-values appear as .000. It is impossible to obtain a \( p \)-value of zero since probabilities are never zero; this should be read as a “\( p \)-value of less than 0.001.”). However, there is not a statistically significant correlation between the two proposed independent variables \( r = 0.178; p = .216 \). Thus, if our hypotheses propose that the unemployment rate and the gross state product are independent variables that predict the violent crime rate, we have some preliminary evidence that the key variables are associated.
Let’s now see what a linear regression analysis tells us about these associations. We’ll begin with a simple linear regression model using only the unemployment rate as an independent variable. SPSS generates the following table:

### Coefficients(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>76.686</td>
<td>151.334</td>
</tr>
<tr>
<td>Unemployment Rate</td>
<td>88.782</td>
<td>28.247</td>
</tr>
</tbody>
</table>

a Dependent Variable: Violent crimes per 100,000

How do we interpret the slope associated with the unemployment rate? It should be easy by now. We may say “Each one unit increase in the unemployment rate is associated with 88.78 more violent crimes per 100,000 people.” The $p$-value suggests that, if the actual population slope is zero, we’d expect to get a slope of
88.78 or something farther away from zero less than three times out of every one-thousand samples. The intercept indicates that the expected value of violent crimes is 76.69 when the unemployment rate is zero (is this a reasonable number?).

You may have noticed that there is one entry in the table that we haven’t yet discussed. Under the column labeled Standardized Coefficients, SPSS lists the term Beta and, below it, the number 0.413. Have you seen this number before? Look at the correlation matrix shown on the previous page. The Pearson’ correlation between the unemployment rate and violent crimes per 100,000 is 0.413. The standardized coefficient measures the association between the variables in term of standard deviation units or $z$-scores; when there is only one independent variable this statistic is the same as the Pearson’s correlation (recall from Chapter 1 that we identified the correlation as the standardized version of the covariance.). As we shall soon see, this will change once we move to a multiple linear regression model.

Standardized regression coefficients – or beta weights as they are often called – may be interpreted using standard deviations. One way to consider them is to imagine transforming both the independent variable and the dependent variable into $z$-scores (i.e., standardize them), running the regression model, and looking at the slope. This slope represents the association between variables in standard deviation units and is identical to the standardized regression coefficient that SPSS produces. Hence, our interpretation of the standardized regression coefficient in the violent crimes model is “Each one standard deviation increase in the unemployment rate is associated with a 0.413 standard deviation increase in the number of violent crimes per 100,000 people.”

The standardized regression coefficient and the unstandardized regression coefficient are related based on the following formula:

$$
\beta^* = \beta \left( \frac{s_x}{s_y} \right)
$$

where $\beta^*$ is the standardized coefficient

Recall that $s_x$ denotes the standard deviation of $x$ and $s_y$ denotes the standard deviation of $y$. In the earlier example, we may use this formula to compute the standardized coefficient for the unemployment rate:
There are some researchers who prefer standardized coefficients because they argue that these statistics allow comparisons of independent variables within the same linear regression model. An independent variable with a standardized coefficient farther away from zero, they argue, is more strongly associated with the dependent variable than are independent variables with standardized coefficients closer to zero. However, this assumes that the distributions of the independent variables are the same. Many times, though, one variable is much more skewed than another, so a standard deviation shift in the two variables is not comparable. When we learn about dummy variables in Chapter 6, we'll also see that standardized regression coefficients are of limited utility. Moreover, standardized regression coefficients should not be used to compare coefficients for the same variable across two different linear regression models.

We'll now estimate a multiple linear regression model by adding the gross state product to the previous model. In SPSS we simply move the new variable into the Independent Variable(s) box and click OK. You should find the following table in the SPSS output window.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>63.514</td>
</tr>
<tr>
<td></td>
<td>Unemployment Rate</td>
<td>72.806</td>
</tr>
<tr>
<td></td>
<td>Gross State Product in $100,000s</td>
<td>67.173</td>
</tr>
</tbody>
</table>

There are now two slopes, one for each independent variable. We'll first offer some interpretations for the numbers in this table and then try to understand how a multiple linear regression model comes up with them.
Try to imagine the association among these three variables in three dimensions. If this is difficult, begin with a two variable scatterplot with x- and y-axes and then picture another axis, labeled z, coming out of the page. The three axes are at 90 degree angles to one another. Now, how is the intercept interpreted in this representation? Perhaps you can guess that it is the expected value of the dependent variable when both of the independent variables are zero. If we had a highly unusual state that had no unemployment and no economic productivity (we’ll call it Nirvana), we’d expect it to have 63.51 violent crimes per 100,000 (Okay, we’ll call it Club Med). Of course, the intercept in this model is meaningless since it falls so far outside of any reasonable values of the independent variables.

The slopes are interpreted the same way as before, but we add a phrase to each statement. Here’s an example: “Controlling for the effects of the unemployment rate, each $100,000 increase in the gross state product is associated with 67.17 more violent crimes per 100,000 people.” Notice that we’ve used the phrase “controlling for.” Other phrases that are often used include “statistically adjusting for…” or “partialling out….” Because we are statistically controlling for or partialling out the effects of a third variable, multiple linear regression coefficients are also known as partial regression coefficients or partial slopes.

Statistical control or adjustment is a difficult concept to grasp. One way to think about it is to claim that we are estimating the slope of one variable on another while holding constant the other independent variable. For instance, when we claim that each $100,000 increase in the gross state product is associated with 67.17 more violent crimes per 100,000 people, we are assuming that this average increase occurs for any value of the unemployment rate, whether 1% or 12%. In other words, the unemployment rate is assumed to be held constant. To further our understanding of this important idea, we’ll discuss four ways of approaching this topic with the hope that perhaps one or two of these will be useful. However, the first way may only be useful if you have superior spatial perception skills. This is to use a three-dimensional plot to visualize the relationship among the three variables. SPSS has a 3-D graphing option. By plotting three variables and then using the regression option to calculate the slopes, you may be able to visualize the meaning of statistical control. SPSS’s 3-D option allows you to rotate the graph, so it can be seen from different angles. Looking at how the points follow a particular
variable’s slope may allow you to visualize the meaning of statistical control.

A second way to understand statistical control assumes experience with calculus. A partial slope is synonymous with a partial derivative. Suppose that $y$ is a function of two variables, $x$ and $w$, that we wish to represent in a regression equation. If $w$ is held constant (e.g., $w = w_0$), then $y$ becomes a function of a single variable $x$. Its derivative at a particular value of $x$ is known as the partial derivative of $y$ with respect to $x$ and is represented by $\frac{\partial y}{\partial x}$ or $\frac{\partial f(x, w)}{\partial x}$ where $y = f(x, w)$. We won’t go any further than this, even though it might be valuable to some readers. We’ll assume that if you’ve had exposure to partial derivatives – and you find them useful – then you can explore the relationship between partial derivatives and partial slopes. Many calculus textbooks include graphical representations of partial derivatives that may help you visualize the notion of holding one variable constant while allowing another to vary.

The next two ways of approaching the concept of statistical control or adjustment in multiple linear regression models should be helpful to most readers. The first involves computing residuals from two distinct linear regression models and then using these residuals to compute the partial slope. Before showing how to do this, it is useful to consider the nature of the residuals. As mentioned in Chapter 2, residuals measure the distance from the regression line to the actual observations. A basic formula for a residual is $\text{resid}_i = (y_i - \hat{y}_i)$. (As mentioned earlier, you’ll often find the residuals represented as $e_i$.) This means that we take the observed values and subtract the values we predict based on the linear regression model. As indicated by the subscripts, we do this for each observation. Another way to think about residuals is that they measure the variation that is left over in the dependent variable after we have accounted for the systematic part that is associated with the independent variable. Part of what is left over may be accounted for by another independent variable. This part is represented by the partial slope.

In order to show how this is done, take the following steps in SPSS:

1. Estimate a simple linear regression model with violent crime as the dependent variable and the unemployment rate as the independent variable
LINEAR REGRESSION ANALYSIS

(leave gross state product out of the model). After placing these variables in the dependent and independent variable boxes, click the lower-right hand box that is labeled Save and then, in the window that opens, check the box near the top that is labeled Residuals–Unstandardized. This will compute a new variable that is made up of the residuals from the model for each observation. These residuals represent the variability in violent crimes that is not accounted for by the unemployment rate.

2. Estimate a second regression model with gross state product as the dependent variable and the unemployment rate as the independent variable. Once again, save the residuals from this model. These residuals represent the variability in the gross state product that is not accounted for by the unemployment rate.

3. Estimate a third regression model that uses the residuals from (1) as the dependent variable and the residuals from (2) as the independent variable. There is no need to save the residuals from this model.

Now, look at the unstandardized regression coefficient from the third model. It is 67.17, the same number we found when we estimated the multiple linear regression model earlier. This is because the slopes from these two models represent the same thing: The covariability between the gross state product and violent crime that is independent of the unemployment rate.

The fourth way to consider statistical control is similar to using residuals, except it is visual. The figure on the next page shows three overlapping circles that are labeled A, B, and C. We could have labeled them y, x₁, and x₂ to show that they are variables we are interesting in using in a regression model. Let’s assume that A represents the dependent variable and B and C are independent variables. The total area of the circles represents their variability (if you find it more comforting, think of the sums of squares or the variances as represented by circles). The overlapping areas signify joint variability, such as a measure of covariance between two variables. Notice the darkened area of overlap between A and B. It does not include any part of C. This area denotes the joint variability of A and B that is not accounted for at all by the independent variable C. It represents the partial slope or the regression coefficient from a multiple linear regression model.
Notice that there is no need for B and C to be completely independent. Rather we can assume that they covary. Hence, the term independent variable can be misleading: It does not mean that the predictor variables are independent from one another, only that they might – assuming our conceptual model is a good one – independently predict the dependent variable. Extending multiple linear regression models to include more than two independent variables is easy and requires no deeper level of understanding than that which we already possess. The phrase “controlling for the effects of…” must now either specify each other independent variable or state something such as “controlling for the effects of the other variables in the model, each one unit increase in variable $x_1$ is associated with a [partial slope] unit increase in variable $y$.”

**A Brief Overview of the Assumptions of Multiple Linear Regression Models**

Now that we have some sense of what multiple linear regression models produce in the way of slopes, intercepts, and $p$-values, it is a good time to review the
assumptions that we make when using the model. We saw some of these assumptions in Chapter 2, but in the interests of completeness, we’ll revisit them. Moreover, this is a brief overview because several of the subsequent chapters review these assumptions in painful detail.

1. **Independence**: The Y values are statistically independent of one another. We saw that this assumption applies to the simple linear regression model. Now that we have seen a couple of regression models, we can think about how this assumption might operate. For example, when analyzing state-level data we assume that our measures are independent across states. But, is this true? It is likely that states that share borders are more similar in a lot of ways than states that are in different parts of the country.

2. **Distribution of the Error Term**: The error term is normally distributed with a mean of zero and constant variance. The first part of this assumption – that the mean is zero – is important for estimating the correct intercept. But this assumption taken as a whole is also important for making inferences from the sample to the population. In other words, it is important for significance tests and confidence intervals.

3. **Specification Error**: The correlation between each independent variable and the error term – or residuals – is zero. If there is a correlation, we’ve left something important out of the model. In Chapter 7, we’ll learn that this problem involves whether or not we have specified the correct model; hence, it is known as specification error.

4. **Measurement Error**: The x’s and y are measured without error. Notice that we are using the lower case to indicate the independent and dependent variables. This is because we are concerned primarily with sample measures. As the name implies, we must consider whether we are measuring the variables accurately. When we ask people to record their family incomes, are they providing us with accurate information? When we ask people whether they are happy, might some interpret this question differently than others? We’ll learn in Chapter 8 that measurement errors are a plague on the social and behavioral sciences.
5. **Collinearity**: There is not perfect collinearity in any combination of the X’s. If you consider the overlapping circles that we saw in the last section, you can visualize what this means. Suppose that circles B and C overlap completely. Is it possible to estimate the covariance between A and B independent of C? Of course not. There is no variability left over in B once we consider its association with C. We’ll discover in Chapter 9 that even a higher degree of collinearity, or overlap, among independent variables can cause problems in multiple linear regression models.

6. **Linearity**: The mean value of Y for each specific combination of the X’s is a linear function of the X’s. For example, the regression surface is flat in three dimensions. Note that in a simple linear regression model we only had to assume a straight line relationship, but now that we have two or more independent variables we must move to higher dimensions. One way to think about this is to imagine a three-dimensional space and then visualize the difference between a flat surface and a curved surface. We assume that the relationship between the X’s and Y is not curved for any X. In Chapter 10, we’ll find out about some of the tools for analyzing relationships that are not linear.

7. **Homoscedasticity**: The variance of the error term is constant for all combinations of X’s. The term homoscedasticity means “same scatter.” Its antonym is **heteroscedasticity** (“different scatter”). We’ll learn much more about this issue in Chapter 11.

8. **Autocorrelation**: There is not an autocorrelated pattern among the errors. This is a huge problem in times series and spatial regression models. These involve data collected over time, such as months or years, or across spatial units, such as counties, states, or nations. Another name for this problem in over-time data is serial correlation. The key issue that leads to autocorrelation is that errors – or residuals – that are closer together in time (year 1 and year 2) or space (Los Angeles County, CA and Orange County, CA) share more unexplained variance than errors that are farther apart in time (year 1 and year 10) or space (Los Angeles County, CA and Fairfax County, VA). Chapter 11 includes a discussion of autocorrelation.
Although not normally listed as an assumption, we also assume that there are not any wildly discrepant values that affect the model. This is related to assumptions about measurement error, specification error, and linearity. Suppose, for instance, that you look at a variable such as annual income and find the following values: \([25,000; 35,000; 50,000; 75,000; 33,000,000]\). As we'll see in Chapter 12, the last entry is labeled as either a leverage point (if income is an independent variable) or an outlier (if income is the dependent variable). In either case, these types of discrepant values – which are known generally as influential observations – can affect, sometimes in untoward ways, linear regression models. The question you should ask yourself when confronted by such values is why have they occurred? Did someone record the wrong value, perhaps by placing a decimal place in the wrong spot? Or is this value accurate; does someone in your sample actually earn that much money per year? If the value is simply coding error, then fixing it is usually simple. If it is an actual value, then perhaps a linear model predicting income, or using it as a predictor, is not appropriate. In this case, we should search for other regression techniques to model income. Fortunately, multiple linear regression models are flexible enough to accommodate many such problems; you just need to know where to look for solutions.

There are two general issues to consider: First, how do we test these assumptions; and, second, what do we do if they are violated. There are tests – most of them indirect – for all of these assumptions. Furthermore, if they are violated – if our model does not meet one or more of these assumptions – there are usually solutions that still make linear regression models viable. Several of the subsequent chapters in this presentation discuss these tests and the solutions. The term regression diagnostics is used as an umbrella term for these tests.

Several of these assumptions are stringent, but, one argument goes, our models are saved by statistical theory’s famous Central Limit Theorem (CLT) even when one or more of these assumptions is violated. You might recall from introductory statistics that the CLT states that, for relatively large samples, the sampling distribution of the mean of a variable is approximately normally distributed even if the distribution of the underlying variable is not normally distributed (For a nice review, see Neil A. Weiss (1999), *Introductory Statistics, Fifth Edition*, Reading, MA: Addison-Wesley, p.427; a standard proof is provided in Bernard W. Lindgren...
For random variables with finite sample variance, the sampling distribution of the standardized sample means approaches the standard normal distribution as the sample size approaches infinity.

This concept concerns the sampling distribution of the mean, which, as we’ve seen before, assumes that we take many samples from the population. If these samples are drawn randomly, the distribution of means tends to approximate the normal distribution after only about 30 samples. However, if the underlying variable’s distribution is highly skewed, it may take more samples to approximate the normal. Since intercepts and slopes are similar to means, they also may be shown to follow particular normal-like distributions (such as the t-distribution).

Therefore, given a large enough sample, we can infer that even if the dependent or independent variables are not normally distributed, the results of the regression model will be accurate most of the time. How large a sample needs to be to be large enough is a difficult question to answer, though. Thus, it is important to learn techniques appropriate for when assumptions like linearity or normality are not met. And this discussion should also reinforce the idea that using random samples or having control over independent variables is important.

Let’s assume, though, that the assumptions are met. Then, according to something known as the Gauss-Markov theorem (see Lindgren (1993), op. cit., p.510), the OLS estimators for the slopes, intercepts, and standard errors, offer the best linear unbiased estimators (BLUE) among the class of linear estimators. No other linear estimator has less bias, or is more accurate on average, than OLS estimators.

Some Important Aspects of Multiple Linear Regression Models

As mentioned earlier in the chapter, OLS is the main technique that we use to estimate a multiple linear regression model. As we’ve now learned, OLS has some nice aspects that make it especially useful in regression analysis. These aspects are important because the OLS regression equation \( \hat{y}_i = \hat{\alpha} + \hat{\beta}_1x_{1i} + \hat{\beta}_2x_{2i} + \cdots + \hat{\beta}_kx_{ki} \)
LINEAR REGRESSION ANALYSIS

results in what is known as a *linear combination* of the $x$’s that has the largest possible correlation with the $y$ variable. This is an ideal because we are attempting, with this model, to explain as much of the variability in the dependent variable as possible with the independent variables. We’ll learn in Chapter 4 about some ways to estimate this correlation.

We also mentioned in the last chapter that statistical software uses matrix routines to come up with slopes, standard errors, and other particulars of linear regression models. For those of you familiar with vectors and matrices, think of the $y$-values as a vector of observations and the $x$-values as a matrix of observations:

$$
Y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix} \quad X = \begin{bmatrix}
1 & x_{11} & \cdots & x_{1p} \\
1 & x_{21} & \cdots & x_{2p} \\
\vdots & \vdots & \cdots & \vdots \\
1 & x_{n1} & \cdots & x_{np}
\end{bmatrix}
$$

Given this short-hand way of expressing the independent and dependent variables (translate the vector and matrix into spreadsheet format, such as SPSS’s data view, if it’s not clear how this works and it is easy to see the utility of representing data this way), we may express the multiple linear regression equation as $\hat{Y} = X\beta$, where the $\hat{\beta}$ represents a vector of the intercept (denoted in the $X$ matrix with 1’s) and slopes for each independent variable. The $X$ is listed first in this equation because we say it is *postmultiplied* by the vector of slopes.

Given the $X$ and $Y$ matrices, the following matrix formula may be used to estimate the vector of slopes:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

The $X$ with the accent next to it is known as the *transpose* of the matrix and the superscripted $-1$ indicates that we should take the inverse of the product in parentheses. In matrix terminology we say the vector of slopes is estimated by postmultiplying the inverse of $X$ times its transpose by $X$-transpose times $Y$. Try it out with a very small data set (if you’ve got sufficient time and patience to compute
MULTIPLE LINEAR REGRESSION ANALYSIS

the inverse) and see if your results match what SPSS comes up with. Matrix algebra is also useful for estimating several other features of multiple linear regression models. For example, the standard errors of the coefficients may be estimated by taking the square roots of the diagonal elements of the following matrix:

$$V = (X'X)^{-1} \sigma^2$$

where $$\sigma^2 = \frac{1}{n-k} \times \epsilon' \epsilon$$

The term $$n$$ refers to the sample size and the term $$k$$ refers to the number of independent variables in the model. We’ll learn more about $$\sigma^2$$ in Chapter 4 when we discuss goodness-of-fit statistics. For now, note that it measures the amount of variability of the residuals around the regression line.

An alternative way of computing the standard errors that does not use matrix algebra is

$$se(\hat{\beta}_i) = \sqrt{\frac{\sum (y_i - \hat{y}_i)}{\sum (x_i - \bar{x})^2 (1 - R^2)(n-k-1)}}$$

We’ll learn more about the $$R^2$$ value in Chapter 4. For now, we'll say only that it is the $$R^2$$ from a linear regression model that includes $$x_i$$ as the dependent variable and all the other independent variables as predictors (e.g., $$x_i = \hat{a} + \hat{\beta}_2x_{2i} + \cdots + \hat{\beta}_kx_{ki}$$).

This model is known as an auxiliary regression model. The quantity in parentheses that involves the $$R^2$$ is called the tolerance. Moreover, notice that the standard error is affected by the sum of squares of $$x$$, with larger values tending to decrease the standard error, and the SSE, with larger values tending to increase the standard error. A larger sample size also tends to decrease the standard error.

4 The ANOVA Table and Goodness-of-Fit Statistics

As discussed earlier, there are some similarities between regression analysis and analysis-of-variance (ANOVA) models. Both are useful, for example, if we wish to estimate the means of specific groups represented by the independent variables. Both models provide significance tests to determine if there is a statistically significant difference between the mean $y$ values for two or more groups signified by the $x$ values. However, reach into your memory to recall the fundamental purpose of ANOVA. This is to partition the variance of a variable into component parts, one due to its systematic association with another variable and the second due to random error.

You probably noticed when we asked SPSS to estimate the model predicting violent crimes per 100,000 that the top of the output included a table labeled ANOVA. Why is it called this? Let’s look at an example and then we can answer this question.

<table>
<thead>
<tr>
<th>Model Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA(b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
These are two of the tables that SPSS provides by default when it estimates a linear regression model. Let’s discuss the second table first since we’ve already mentioned ANOVA. As in a standard ANOVA, there is a set of numbers known as the sum of squares. Recall from Chapter 1 that we discussed a measure of dispersion known as the sum of squares of X, abbreviated SS[X]. The formula for this measure is \( SS[X] = \sum (x_i - \bar{x})^2 \). But we now have a dependent variable, labeled \( y \), and, as seen in the last chapter, we wish to assess something about the overlap between the independent and dependent variables. So, it makes sense to try to come up with a measure of this overlap. Now, if the sum of squares of the dependent variable measures its total area, we might also wish for a measure of the overlapping area. This is what the ANOVA table provides. Under the table cell labeled Total Sum of Squares (TSS) we have SS[y]. If we were to take all 50 values of the violent crime variable, subtract the mean from each, and add these numbers up, we would obtain an overall sum of squares of 3,551,889.533. This may be thought of as the total area of a circle representing the variation in violent crime.

What, then, are those other sums of squares? The first one, in the Regression Sum of Squares (RSS) cell, computes the variation of the mean of \( y \) around the values of \( y \) predicted from the regression equation. Recall that the predicted value is designated as \( \hat{y} \) and is calculated based on the linear regression equation. Each non-missing observation in the data set has a predicted value. We may therefore modify the sum of squares formula to reflect this type of variation:

\[
RSS = \sum (\hat{y}_i - \bar{y})^2
\]

This represents the total area of overlap between the variability of the \( x \)'s and the variability of \( y \) (since the values of \( y \) are predicted based on the \( x \)'s). It may also be thought of as the improvement in prediction over the mean with information from the independent variables. Its value in the linear regression model is 606,247.35.

The other sum of squares, in the Residual Sum of Squares cell (better known as the sum of squared errors, or SSE [see Chapter 2]), is the area of \( y \) that is left over after accounting for its overlap with the independent variables. It is computed using the following equation:
Its value in the SPSS table is 2,945,642.183. Notice that the RSS and the SSE sum to equal the TSS: 606,247.35 + 2,945,642.183 = 3,551,889.533. This is because we assume we have accounted for all of the variation in \( y \) with either the variability in the \( x \)'s or random variability (this does not, however, rule out that there might be other variables that account for portions of the variability of \( y \)). Here's a picture of what we are examining with the ANOVA portion of our linear regression model of violent crimes per 100,000:

Another way of representing the equations inherent in the ANOVA table is to combine the sums of squares equations into one general equation. Such an equation should represent the partitioning of the area of the left-hand circle, or the
total sum of squares of violent crimes, into its two component parts. The diagram below illustrates this concept.

The next issue to address is whether we can use this information on overlapping or joint variability to come up with a measure of the overall fit of the model. In regression analyses of all types there is a concern with what is known as goodness-of-fit statistics. How do we know if we are accounting for a substantial amount of the variability in the dependent variable?

\[ \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \]

\[ \text{SS}[y] \text{ or TSS} \]
\[ \text{Sum of squares due to regression or RSS} \]
\[ \text{Sum of squares due to residuals /errors or SSE} \]

3,551,889.5 = 606,247.35 + 2,945,642.2

One idea is to use the sums of squares to estimate the amount of variability. Since we have area measures of a type, we can estimate the proportion or percentage of variation in the dependent variable that is accounted for by the independent variables. This is the purpose of the model summary table that SPSS prints. There is a number in the column labeled R Square; this is typically labeled simply R². Known in some circles as the coefficient of determination or the squared multiple correlation coefficient, it indicates the proportion of variability in the dependent variable that is accounted for (some use the phrase “explained by”) the independent variables. It is simple to transform it into a percentage. In the violent crime model, the R² is 0.171, or 17.1%. In other words, the linear regression model accounts for 17.1% of the variability in violent crimes. The number labeled R is known as the multiple correlation coefficient (note that it’s the square root of
R\(^2\)). It indicates the correlation between the values predicted by the regression model and the actual values of \(y\), or \(\text{corr}(y_i, \hat{y}_i)\).

Returning to our sums of squares equation, it should be apparent how to compute this \(R^2\) value: Take the RSS and divide it by the TSS. Therefore, from our model that predicted violent crimes we find \(606,247.35/3,551,889.533 = 0.171\). Sometimes the following formula is used, but it is equivalent to simply dividing the RSS by the TSS since RSS + SSE = TSS:

\[
R^2 = 1 - \frac{\text{SSE}}{\text{TSS}}
\]

An important property of the \(R^2\) measure is that it falls between zero and one, with zero indicating that the RSS is zero and a one indicating that the RSS = TSS, with the SSE equal to zero.

An interesting phenomenon occurs when we add independent variables to the model: The \(R^2\) increases whether or not the independent variables add anything of substance to the model’s ability to predict \(y\). Many researchers are understandably bothered by this problem since they do not wish to be misled into thinking that their model provides good predictive power when it does not. Thus, most use a measured known as the adjusted \(R^2\). The term adjusted is used to indicate that the \(R^2\) is adjusted for the number of independent variables in the model. If we simply add useless independent variables, the adjusted \(R^2\) can actually decrease. The formula for this statistic is

\[
\bar{R}^2 = \left( R^2 - \frac{k}{n-1} \right) \left( \frac{n-1}{n-k-1} \right)
\]

In this equation, as before, \(n\) indicates the sample size and \(k\) indicates the number of independent variables in the regression model.

The other term in the Model Summary table is labeled Std. Error of the Estimate. Although it is not used very often in the social and behavioral sciences (although it probably should be), this standard error, which is sometimes referred to as the Root Mean Squared Error or the Standard Error of the Regression (\(S_E\) or
\( S_{y|x}, \) measures the average size of the residuals. It is also the square root of what is commonly referred to as the *Mean Squared Error* (MSE):

\[
\text{MSE} = \frac{\text{SSE}}{n-k-1} \quad \sqrt{\text{MSE}} = S_E
\]

The mean squared error is found in the ANOVA table (find the cell in the Mean Square column and the Residual row). The root mean squared error may be used as a goodness-of-fit measure since smaller values indicate that there is less variation in the residuals of the model. It is also useful for estimating what are known as *prediction intervals*: Bounds for the predictions we make with the model for particular observations within the data set. For instance, say we wish to determine the interval of predicted violent crimes for a state in which the unemployment rate is 5.2, 6.4, or 7.8. The following equation may be used:

\[
\text{PI} = \bar{y} + \hat{\beta}_1 (x_0 - \bar{x}) \pm (t_{n-2})(S_E) \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\text{Var}(x)}}
\]

A value such as 5.2 is substituted for \( x_0 \). The \( t \)-value is the confidence level we wish to use with degrees of freedom equal to \( n - 2 \). For example, with a sample size of 50, we’d use a \( t \)-value of 2.011 for a confidence level of 95% (two-tailed). As with the confidence intervals for the slopes and the residuals from the model, SPSS will compute prediction intervals (see the Save window in the *Analyze-Regression-Linear* drop-down menu).

As with other statistics derived from a sample, the \( R^2 \) has a sampling distribution. If we were to take many samples from the population, then we could compute \( R^2 \)'s for each sample. We may therefore use significance tests and confidence intervals for this statistic. Before seeing such a test, though, we have to determine how \( R^2 \) values are distributed. We’ve thus far mentioned the normal distribution, the standard normal distribution (\( z \)-distribution), and the \( t \)-distribution. \( R^2 \) values follow an \( F \)-distribution. Perhaps you remember the \( F \)-distribution from the ANOVA model. If so, you’ll recall that it has two degrees of freedom, known as the numerator and denominator degrees of freedom. Notice
that in the ANOVA table there are columns labeled df (degrees of freedom) and Mean Square. The cells in the degrees of freedom column include the number of independent variables ($k$; in the Regression row), the sample size minus the number of independent variables minus one ($n - k - 1$; in the Residual row), and the sample size minus one ($n - 1$; in the Total row). The cells in the Mean square column are the sums of squares in the rows divided by their accompanying degrees of freedom. Hence, the mean square due to regression is $606,247.35/1 = 606,247.35$ and the mean square due to residuals is $2,945,642.183/48 = 61,367.545$. This latter value is also the aforementioned mean squared error (MSE).

The column labeled F includes the $F$-value, which is the mean square due to regression divided by the mean square due to residuals (MSR/MSE). Similar to the way we computed the $t$-values for the slopes, this statistic may be thought of as a measure of how far our regression predictions are from zero, or how well our model is doing in predicting the dependent variable. The degrees of freedom for the $F$-value are found in the df column: $k$ and $(n - k - 1)$. If we have a program that will compute it, such as SPSS, we’ll rarely have to consult the $F$-values in a table at the back of a statistics book or on the internet. Rather, we find in the ANOVA table a column labeled Sig. that contains the $p$-value we need. This is the significance test for whether our model is predicting the dependent variable better with information about the independent variables than with simply the mean of the dependent variable. A better way, perhaps, of looking at the $F$-test ($F$-value and its $p$-value) is that it compares the null hypothesis that $R^2$ equals zero versus the alternative hypothesis that $R^2$ is greater than zero. If we find a significant $F$-value, we gain confidence that we are predicting a significant amount of the variability in the dependent variable. It should by now be rather unsurprising that we may also compute confidence intervals for the estimated $R^2$. However, this is done so rarely that we won’t learn how to do it (but see Kleinbaum et al., op. cit., Chapter 10).

Most researchers from the social and behavioral sciences use the $R^2$ and accompanying $F$-test to determine something about the importance of the model. You will often find statements such as “the model accounts for 42% of the variability in the tendency to twitch, and this proportion is significantly greater than zero.” However, you should be cautious about such conclusions. Some
statisticians argue that we should rely on either the substantive conclusions recommended by the partial regression slopes or the MSE (or $S_e$) to judge how well the set of independent variables predicts the dependent variable (see, in particular, Franklin A. Graybill and Hariharan K. Iyer (1994), *Regression Analysis: Concepts and Applications*, Belmont, CA: Duxbury Press, Chapter 3, for a discussion of this issue).

**An Example of a Multiple Linear Regression Model**

Before completing our discussion of the ANOVA table and goodness-of-fit statistics, let's see another example of a multiple linear regression model. The data set *gpa.sav* has information on first-year college grade point average (*gpa*), scholastic aptitude test (SAT) scores (*sat_math*, *sat_verb*), and high school grades in English and mathematics (*hs Engl*, *hs math*) for 20 students. We'll use SPSS to estimate a model with *gpa* as the dependent variable and the following independent variables: *sat_math*, *sat_verb*, and *hs_math*. The results are in the following tables:

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.922(a)</td>
<td>.850</td>
<td>.822</td>
<td>.26211</td>
</tr>
</tbody>
</table>

*a Predictors: (Constant), Grades in HS math; SAT score, verbal; SAT score, math*

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>6.247</td>
<td>3</td>
<td>2.082</td>
<td>30.307</td>
<td>.000(a)</td>
</tr>
<tr>
<td>Residual</td>
<td>1.099</td>
<td>16</td>
<td>.069</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>7.346</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a Predictors: (Constant), Grades in HS math; SAT score, verbal; SAT score, math*

*b Dependent Variable: First year college GPA*
First, let’s look at the Model Summary information. The $R^2$ value is 0.85, the adjusted $R^2$ value is 0.822, and the standard error of the estimate is 0.262. Hence, we may claim that, using the three independent variables, the model accounts for 85% of the variability in first-year college GPA. This percentage explained is not affected much by the number of independent variables in the model. The ANOVA table indicates that the $R^2$ value is statistically distinct from zero, with an $F$-value of 30.307 ($df = 3, 16$) that is accompanied by a $p$-value of less than .001.

The table of coefficients shows several results. It is best to ignore the intercept in this model because, first, SAT scores do not have a zero value (the minimum is 200) and, second, it unlikely that a student who had a grade of zero in high school math would be attending college! Since SAT scores for math and verbal abilities are measured on a similar scale, we may compare their coefficients. It appears that SAT math scores matter more than SAT verbal scores for predicting GPA. We may say “controlling for the effects of SAT verbal scores and high school math grades, each one unit increase in SAT math scores is associated with a 0.218 increase in first-year college GPA.” The Beta coefficients (beta-weights), which measure the associations in standard deviation units, support the idea that SAT math scores are a better predictor than verbal scores or high school grades.

However, it is important to always remember the issue of sampling error: We do not know based only on the point estimates whether the two SAT coefficients are significantly distinct. There are methods for determining this, but they are
beyond the scope of this presentation. Relatively simple methods involving $F$-tests or $t$-tests are shown in Samprit Chatterjee and Bertram Price (1991), *Regression Analysis by Example, Second Edition*, New York: Wiley, pp.76-79. (See also Graybill and Iyer, *op. cit.*, Chapter 3) A crude way to compare them, though, is to construct confidence intervals for each coefficient and then see whether there’s any overlap. For instance, the confidence intervals for the slopes indicate that the 95% CI for `math_sat` is \{0.122, 0.315\} and for `verbal_sat` is \{0.020, 0.243\}. Since these overlap by a substantial margin, we should not be confident that math scores are a better predictor than verbal scores. This conclusion should not be surprising, however, because the sample size is small ($n = 20$).

The coefficient for high school math grades offers an interesting problem. The $p$-value for this coefficient is 0.057, just over the usual threshold for making claims of statistical significance. Yet we could clearly make a claim here that this result is statistically significant. It seems perfectly reasonable to create, *a priori*, a hypothesis that states that as math grades increase, first-year GPA should also increase. If this is the case (although you might contend “it’s too late for that!”), then my hypothesis is directional and I may use a one-tailed significance test or confidence interval. As mentioned earlier, to switch from a two-tailed test (the SPSS default) to a one-tailed test we divide the $p$-value in half. Thus, the one-tailed $p$-value is $0.057/2 = 0.0285$. Voilà! We now have a significant result. This demonstrates the uncertainty involved in many statistical endeavors. Some might see this statistical sleight-of-hand as entirely justifiable, whereas others may argue that we are being disingenuous. There is no simple answer, though, so we should be careful to set up our hypotheses and describe the types of significance tests we plan to use before we estimate the regression model.
5 Comparing Linear Regression Models

Now that we know something about ANOVA tables and goodness-of-fit statistics, we might be tempted to claim that it is a simple matter to use this information to compare different linear regression models. For example, suppose we take the model in the last chapter that sought to predict first-year college grades and add another variable to the model. How do we determine if the new model, to use a common phrase, “fits the data better” than the initial model? Perhaps your reply is “that’s easy; we simply look to see whether the additional variable is a significant predictor of the dependent variable.” Or, “I’ll look to see whether the $R^2$ has increased from one model to the next.” Well, sometimes these approaches work quite well; other times they don’t. If we decide that all three variables in the first model are significant predictors – perhaps because we have used a one-tailed significance test – then simply adding another variable and checking whether its coefficient is significant works all right in most situations.

However, regression models can be funny things and you may find various differences arise when you begin to add and take away variables from a model. We’ll discuss this issue in greater detail in Chapter 7. For now, we’ll discuss model fitting procedures, mainly by comparing models with different variables. First, we need to introduce a couple of terms. A nested model is one that has the same subsets of variables, but fewer overall independent variables than another model. For example, suppose we have two regression models designed to predict serum cholesterol levels among a sample of adult males. The first model includes the following independent variables:

{resting heart rate, exercise frequency, age}

The second model includes the following independent variables:

{resting heart rate, exercise frequency, age, triglyceride levels, smoking status}
We say that the first model is *nested* within the second model because it includes a subset of the variables. Often, the model with the additional independent variables is called the *full* model; the nested model is known as the *restricted* model.

Non-nested models are two (or more) regression models that do not include the same subset of variables. For example, if the first model included resting heart rate, exercise frequency, and age as independent variables, and the second included triglyceride levels, smoking status, and time spent each day watching television, we cannot say that one model is nested within another. Since comparing models in the non-nested situation is difficult (see Graybill and Iyer, *op. cit.*, Chapter 4), we will not pursue this issue here.

**The Partial F-Test and Multiple Partial F-Test**

The nested situation is pretty easy to deal with. Let’s begin with the simplest scenario: We add one more independent variable to the model and see if it does a better job predicting the dependent variable. The two general tests that we may use are a *t*-test and a type of *F*-test known as a *Partial F-Test*. The partial *F*-test relates the extra sum of squares attributable to adding one independent variable to the model to the Mean Square Error (MSE) from the full model:

\[
F = \frac{\text{RSS}(\text{full}) - \text{RSS}(\text{restricted})}{\text{MSE}(\text{full})} \quad df = 1, \; n - k - 1 \quad \text{(full model)}
\]

One of the advantages of this test is that, as we shall see, it generalizes well when we wish to add more than one independent variable to the model. However, it may be inefficient in the current situation because we need to check a table of *F*-values to determine the *p*-value. As an alternative, most researchers use the *t*-test. The *t*-value from a two-tailed *t*-test is the square of the *F*-value from the partial *F*-test. Moreover, the *t*-test is printed in the SPSS printout: It is the *t*-value and accompanying *p*-value for the extra independent variable that has been added to the model.

But suppose we wish to compare two models, one of which is nested in the other, but the full model contains two or more extra independent variables? In this
situation we extend the partial $F$-test by using the *Multiple Partial F-Test*. As the formula for this test shows, it is a logical extension of the partial $F$-test:

$$ F = \frac{[\text{RSS (full)} - \text{RSS (restricted)}]/q}{\text{MSE (full)}} \quad df = q, n-k-1 \text{ (full model)} $$

In this equation we compute an $F$-value based on the difference between the regression sums of squares from the two models. But note that this difference is divided by $q$ – the difference in the number of independent variables between the two models [(\(k\text{(full)} - k\text{(restricted)})] – before dividing by the mean square error of the full model. The degrees of freedom for the $F$-value are $q$ and the degrees of freedom associated with the error sum of squares for the full model.

As an example, let’s return to the *usdata.sav* data set. The models we estimate include the number of robberies per 100,000 as the dependent variable (*robbrate*). The independent variables in the restricted model are the unemployment rate (*unemprate*) and per capita income in $1,000s (*percap*). The full model includes these two variables plus the gross state product in $100,000s (*gsprod*) and migrations per 100,000 (*mig_rate*). The results of the restricted (or nested) multiple linear regression models are found in the following tables:

**Model Summary – Restricted**

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.576(a)</td>
<td>.331</td>
<td>.302</td>
<td>84.61226</td>
</tr>
</tbody>
</table>

a Predictors: (Constant), Per Capita Income in $1,000s, Unemployment Rate

**ANOVA(b)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>163236.416</td>
<td>2</td>
<td>81618.208</td>
<td>11.400</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>329324.796</td>
<td>46</td>
<td>7159.235</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Total</td>
<td>492561.212</td>
<td>48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Predictors: (Constant), Per Capita Income in $1,000s, Unemployment Rate

b Dependent Variable: Robberies per 100,000
These results indicate that the independent variables account for about 33% of the variability in robberies per 100,000. Moreover, there is a positive association between robberies and each of the independent variables: States with higher rates of unemployment or higher per capita incomes tend to experience more robberies.

Now, let’s see what happens when we re-run the model, but add the gross state product and migrations per 100,000. The following tables contain the results of the full model:

Model Summary – Full

<table>
<thead>
<tr>
<th>Model</th>
<th>R</th>
<th>R Square</th>
<th>Adjusted R Square</th>
<th>Std. Error of the Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.746(a)</td>
<td>.557</td>
<td>.517</td>
<td>70.41369</td>
</tr>
</tbody>
</table>

a Predictors: (Constant), Migrations per 100,000, Unemployment Rate, Per Capita Income in $1,000s, Gross State Product

ANOVA(b)

<table>
<thead>
<tr>
<th>Model</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regression</td>
<td>274405.373</td>
<td>4</td>
<td>68601.343</td>
<td>13.836</td>
</tr>
<tr>
<td></td>
<td>Residual</td>
<td>218155.839</td>
<td>44</td>
<td>4958.087</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>492561.212</td>
<td>48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Predictors: (Constant), Migrations per 100,000, Unemployment Rate, Per Capita Income in $1,000s, Gross State Product

b Dependent Variable: Robberies per 100,000
### COMPARING LINEAR REGRESSION MODELS

**Coefficients (a)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-218.301</td>
<td>83.955</td>
</tr>
<tr>
<td>Unemployment Rate</td>
<td>18.673</td>
<td>8.647</td>
</tr>
<tr>
<td>Per Capita Income in $1,000s</td>
<td>10.052</td>
<td>3.382</td>
</tr>
<tr>
<td>Gross State Product</td>
<td>31.995</td>
<td>6.845</td>
</tr>
<tr>
<td>Migrations per 100,000</td>
<td>.006</td>
<td>.003</td>
</tr>
</tbody>
</table>

*a Dependent Variable: Robberies per 100,000*

It appears we’ve added some important information to the model. The gross state product, in particular, has a highly significant association with robberies per 100,000 after controlling for the effects of the other variables in the model. In fact, if the beta weights can be trusted (and we cannot know without checking the distributions of all the independent variables whether they are suitable for comparing the associations), the gross state product has the strongest association with robberies among the four variables.

But can we also determine whether the addition of the two independent variables significantly improves the predictive power of the model? Perhaps it would be useful to compare the $R^2$ values. After all, they measure the proportion of variability in $y$ that is explained by the model. The $R^2$ for the restricted model is 0.331 and for the full model 0.557. Unfortunately, it is not a good idea to compare these statistics. Recall that in Chapter 4 we learned that the $R^2$ goes up simply by adding independent variables, whether they are good predictors or not? An alternative is to compare the adjusted $R^2$ values since these are not as affected by the addition of independent variables. The values are 0.302 for the restricted model and 0.517 for the full model. It appears that this is a substantial increase: The adjusted $R^2$ went up by about $(0.517 - 0.302) / 0.302 \times 100 = 71\%$. But is the increase in explained variance statistically significant? We may use the multiple
part F-test to answer this question:

\[
F = \frac{[274,405.373 - 163,236.416]/2}{4,958.087} = 11.211 \ (df = 2, 44)
\]

An F-value of 11.211 with \(df = (2, 44)\) has a \(p\)-value of approximately 0.00012, which is much lower than the threshold recommended by most rules-of-thumb. In this situation, it is safe to conclude that the addition of the gross state product and migrations per 100,000 rate increases the explained variance in the model.

But is this all we would wish to do with this model? We added two variables, one of which, the gross state product, appears to be strongly associated with the number of robberies. The other variable, though, does not appear to be as strongly associated with robberies as the other independent variables. However, assuming our hypothesis says we should use a one-tailed significance test (i.e., “Migrations per 100,000 are positively associated with the number of robberies because…”), we may claim that the \(p\)-value is \(0.063/2 = 0.0315\), below the common rule-of-thumb of \(p < 0.05\). Perhaps we should compute a partial F-test to determine whether the addition of migrations adds anything to the model’s explained variability.

However, at some point – some argue it should be very early in the analysis process – you as the analyst must decide about which variables are important and which variables are not. This should be guided, preferably, by your theory or conceptual model, rather than by rote estimation of regression models or correlation matrices.

**Confounding Variables**

There is one other issue to discuss about our multiple linear regression models. This returns us to a point made in Chapter 3 about the need for independent variables to check for confounding. As mentioned there, a confounding variable accounts for the association between an independent and a dependent variable (recall the lighter purchases and lung disease example). We’ll discuss confounding a bit more in Chapter 7, but for now it is useful to consider the two regression models designed to predict the number of robberies per 100,000. Looking back over the coefficients from the two models, notice that in the restricted model the unemployment rate has a partial slope of 27.388, whereas in the full model it has a
partial slope of 18.637. In other words, the partial slope decreased by about 32% when we included the gross state product and migrations per 100,000 in the model. We cannot tell at this point whether this decrease is statistically significant (although there are tools for this), but let’s assume it is. We may then claim that one or both of the new independent variables included in the model confounded the association between the unemployment rate and the number of robberies. It did not confound it completely – the slope for the unemployment rate would be statistically indistinct from zero if this was the case – but changed it enough to draw our attention. We usually look for variables that completely account for the association between two variables, but even those that only partially account for it can be interesting. The key question you should always ask yourself, though, is why. Why does the association change when we add a new variable? It could be a random fluctuation in the data, but it might be something important and worthy of further exploration. In this example, perhaps the unemployment rate and the gross state product are associated in an interesting way.
6 Dummy Variables in Linear Regression Models

The term *dummy variable* is not meant as an insult. We are not claiming that one of our variables is dimmer, duller, or has a lower IQ than another. Rather, it is related to the idea that mannequins or ventriloquists’ puppets – often known as *dummies* – represent or take the place of real people. Dummy variables are used to represent something of particular interest in the regression model, such as a treatment effect or a specific group in the sample (e.g., high school graduates; married people). A dummy variable is defined as a type of variable that takes on only a limited number of values so that different categories of a categorical variable may be represented in the regression model. If you’ve ever conducted an ANOVA to determine the mean differences in some outcome for two or more groups, you’ve probably seen dummy variables. Often, these are called *indicator variables* to designate that they indicate a group represented in the sample. It is important to remember that the groups represented by the dummy variables must be mutually exclusive.

Research in the social, behavioral, and health sciences is full of dummy variables. For instance, suppose we wish to determine whether gender or ethnicity is associated with some continuous outcome variable. Clearly, gender, ethnicity, religious group affiliation, marital status, family structure and many other variables cannot be represented by continuous, let alone normally distributed, variables. Yet we often wish to include these types of items in regression models. Dummy variables provide an efficient way of including categorical or binary (two-category) variables. Here are some examples of the coding associated with dummy variables:

- **Gender:** 0 = male; 1 = female

- **Ethnicity (Caucasian, African-American, Hispanic):**
  - $x_1$: Caucasian = 1; other = 0
  - $x_2$: African-American = 1; other = 0
  - $x_3$: Hispanic = 1; other = 0
The first variable, \textit{gender}, is straightforward since we place people into one of only two groups. The second set of variables requires an explanation, though. Assume we have a variable in our data set that is designed to measure ethnicity (we’ll limit this example to three ethnic groups to simplify things). It has three outcomes – Caucasian, African-American, and Hispanic – that are coded 1, 2, and 3. Since this variable is not continuous, nor can it even be ordered in a reasonable way, we could not include it as is in a linear regression model (what would a one unit increase or decrease mean?). The solution is to create three dummy variables that represent the three groups. The main group represented by each is coded as 1, with any sample member not in this group coded as 0. We can extend this form of dummy variable coding to a variable with any number of unique categories. In the data set (spreadsheet format), our new variables appear as

<table>
<thead>
<tr>
<th>Observation</th>
<th>Ethnicity</th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Caucasian</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Caucasian</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>African-American</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>African-American</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>Hispanic</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Hispanic</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

The variables $x_1$, $x_2$, and $x_3$ are dummy variables that represent the three ethnic groups in the data set. There are several other types of coding strategies possible, such as effects coding and contrast coding, but the type shown in the table is flexible enough to accommodate many modeling needs. We could also add additional groups (e.g., East Asian, Native American, Aleut) if our original variable included codes for them.

There is a rule you should always remember when using dummy variables to represent mutually exclusive groups (e.g., married people, never married people, widowed people, etc.) in a regression model: When setting up the model, include all the dummy variables in the model except one. The one dummy variable that is omitted from the model is known as the \textit{reference category} or \textit{comparison group}.

Suppose our variable has five categories (represented by $k$) and we create five
dummy variables to represent these categories. We then use four (or \( k - 1 \)) dummy variables in the regression model. If you try to use all five groups the model will not run correctly because there is perfect collinearity among the dummy variables (see Chapter 9).

How should you choose the reference category? This is up to you, but there are some guiding principles that may be helpful. First, many researchers use the most frequent category as the reference category, although it is usually preferable to let your hypotheses guide the selection. Second, it is not a good idea to use a relatively small group as the reference category. Third, many statistical software programs have regression routines that automate the creation of dummy variables. When these are used you should check the software documentation to see which dummy variable is excluded from the model. Some programs exclude the most frequent category; others exclude the highest or lowest numbered category.

How are dummy variables used in a regression model? Just like any other independent variable. It is often valuable to give the variables names that are easy to recognize. So, for example, if we were to create a set of dummy variables representing marital status, we might wish to name them \textit{married}, \textit{divorced}, \textit{cohabit}, \textit{single}, and \textit{widowed}.

Let’s first consider a regression model with only one dummy variable. This dummy variable is labeled \textit{gender}. The dependent variable is annual personal income in $10,000s (\textit{pincome}). If only gender is in the model, it is identical to an ANOVA model and, as we shall see, may be used to determine the mean income levels for males and females. The SPSS data set \textit{gss96.sav} contains these variables as well as a host of others from a representative sample of adults in the United States. Notably, there are two variables that measure gender: \textit{sex} and \textit{gender}. The first is coded as 1 = male and 2 = female and the second is coded as 0 = male and 1 = female. This illustrates a frequent coding situation: There are many data sets that code dummy variables as \{1, 2\}. Although we may still use these variables in a regression model, the \{0, 1\} coding scheme is preferable for reasons that will become clear later on. In addition, the data set includes a marital status variable that is nominal (\textit{marital}) and a set of dummy variables representing its categories (e.g., \textit{married}, \textit{widow}); and a race/ethnicity variable (\textit{race}) along with a set of dummy variables (\textit{Caucasian}, \textit{AfricanAm}, \textit{othrace}) that represent its categories.
There are also several other variables in this data set that measure family income, religion, volunteer activities, and various demographic characteristics. (Note: The income variables are not truly continuous, nor are they coded correctly. Rather, they include categories of income that are not in actual $10,000 units, even though they are labeled as such. Nonetheless, in the interests of unfettered learning, we’ll treat them as continuous and pretend that the units are accurate.)

Setting up a linear regression model with dummy variables in SPSS requires no special tools or unique approach. If you’re using the drop-down menus, the set-up is the same as with continuous independent variables: Simply place the dummy variable(s) in the box labeled Independent(s). Let’s run the model specified in the last paragraph and see what SPSS provides.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>10.595</td>
<td>.093</td>
<td>113.515</td>
</tr>
<tr>
<td></td>
<td>gender</td>
<td>-1.347</td>
<td>.132</td>
<td>-.229</td>
</tr>
</tbody>
</table>

a Dependent Variable: annual personal income in $10,000s

The output looks the same as before, with coefficients, standard errors, t-values, and p-values. In fact, most of the interpretations are similar, except the most important: The coefficients. Recall that we learned earlier that the intercept is sometimes not very useful in linear regression models because independent variables are frequently coded so as not to have an interpretable zero value. Think about how a dummy variable is different: It clearly has an interpretable zero value since it is coded as \{0, 1\}. But, because it no longer makes much sense to refer to a one-unit increase in the independent variable – at least not in the same way as with continuous independent variables – we need to modify our thinking about the slope. Perhaps writing out the linear regression model will help us move toward a clearer interpretation of the results:

Personal income, predicted = 10.595 – 1.347(gender)
Suppose we wish to compute predicted values from this equation. This should be pretty easy by now:

Predicted income for males: \[10.595 - 1.347(0) = 10.595\]
Predicted income for females: \[10.595 - 1.347(1) = 9.248\]

Consider that the mean of personal income is 9.93, which looks remarkably close to the average of these two predicted values. It should be for the simple reason that, as foreshadowed earlier, these two predicted values are the mean values for personal income among males and females (confirm this using SPSS’s Analyze-Compare Means-Means menu).

Given that we may predict these two means with the linear regression model, what do the intercept and slope represent? It should be apparent that the intercept, when there is only one \{0, 1\} coded dummy variable, is the mean of the dependent variable for the group represented by the zero category. The slope represents the average difference in the group means of the dependent variable. Hence, we may say that the expected difference in personal income between males and females is 1.347. Without even computing predicted values, we immediately see that females, on average, report less personal income than males.

One experienced with various types of statistical analysis might ask: Isn’t this the type of issue that \(t\)-tests are designed for? Of course it is. In fact, using SPSS’s Analyze-Compare Means-Independent Samples \(T\)-Test option we find the same results as in the linear regression model: An average difference between males and female of 1.347, a \(t\)-value of 10.24, and a \(p\)-value of less than 0.001. An advantage of SPSS’s \(t\)-test routine is that it provides a statistical test that addresses whether the variances in the population are identical between males and females (they probably are not, according to SPSS’s Levene test).

**Dummy Variables in Multiple Linear Regression Models**

Considering that the two sample \(t\)-test provides the same information as the dummy variable approach in a simple linear regression model, are there any advantages to using dummy variables? The main advantage is that we may include additional independent variables in the regression model to come up with predicted
means that are adjusted for continuous or other dummy variables. Perhaps you’ve heard of analysis of covariance (ANCOVA)? A linear regression model offers a straightforward way to estimate this type of model. But first let’s see another nice property of the linear regression approach by adding some more dummy variables, only this time we’ll add dummy variables to explore differences among several groups.

Let’s return to the personal income model and add the race/ethnicity dummy variables. First, we need to decide which of this set to exclude as the reference category. The frequencies for the three groups are 1,547 (Caucasian), 254 (African-American), and 98 (other race/ethnicity). We have not set up a hypothesis that guides the selection of the reference category, so we’ll have to make a somewhat arbitrary selection. Let’s choose the most frequent category – Caucasian – as our comparison group. We’ll then estimate a linear regression model that includes \textit{gender} but adds the two dummy variables that identify African-Americans and members of other racial/ethnic groups to predict personal income.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
Model & Unstandardized Coefficients & Standardized Coefficients & t & Sig. \\
\hline
(1) & (Constant) & 10.613 & .097 & 108.930 & .000 \\
& gender & -1.349 & .132 & -2.229 & -10.209 & .000 \\
& African-American & -.029 & .195 & -.003 & -.150 & .881 \\
& other race/ethnicity & -.267 & .299 & -.020 & -.892 & .372 \\
\hline
\end{tabular}
\caption{Dependent Variable: personal income in $10,000s}
\end{table}

Considering the way we’ve treated dummy variables so far, what do you think these slopes and the intercept measure? As before, let’s write out the regression model:

\[
\text{Predicted income} = 10.613 - 1.349(\text{gender}) - 0.029(\text{African-American}) - 0.267(\text{other race/ethnicity})
\]
Since the intercept in a multiple linear regression model is the expected value of the dependent variable when all the independent variables are zero, 10.613 represents the expected mean for males who are not African-American and are not in the other racial/ethnic category. So who’s left? Caucasians. In other words, the intercept represents the mean for the omitted category from both sets of dummy variables: gender and race/ethnicity. As we found earlier, the slopes represent average differences among the groups. These are also known as deviations from the mean. Here we are interested in computing deviations from the mean of the reference category (Caucasian males). But we must now combine coefficients to estimate the means for these different groups, as shown in the following table.

<table>
<thead>
<tr>
<th>Group</th>
<th>Computation</th>
<th>Expected Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian males</td>
<td>10.61</td>
<td>10.61</td>
</tr>
<tr>
<td>Caucasian females</td>
<td>10.61 – 1.35</td>
<td>9.26</td>
</tr>
<tr>
<td>African-American males</td>
<td>10.61 – 0.03</td>
<td>10.58</td>
</tr>
<tr>
<td>African-American females</td>
<td>10.61 – 1.35 – 0.03</td>
<td>9.23</td>
</tr>
<tr>
<td>Other race/ethnic males</td>
<td>10.61 – 0.27</td>
<td>10.34</td>
</tr>
<tr>
<td>Other race/ethnic females</td>
<td>10.61 – 1.35 – 0.27</td>
<td>8.99</td>
</tr>
</tbody>
</table>

We’ve estimated six means for six mutually exclusive groups defined by the dummy independent variables. However, before being too pleased with ourselves, consider the $t$-value and $p$-values in the regression output. As in the earlier model, the gender coefficient is statistically significant. Another way of thinking about this is that the gender difference is statistically significant. However, the race/ethnicity coefficients are not statistically significant. This means that we cannot say that the income differences among Caucasians, African-Americans, and members of other racial/ethnic groups are substantial enough that we could conclude that they differ in the population of adults in the U.S. Also notice that the income difference between males and females in each racial/ethnic group is identical (1.35). Whereas this difference is statistically significant, the income difference between males across racial/ethnic groups and females across racial ethnic groups is not statistically significant (e.g., the difference between Caucasian females (9.26) and Other
race/ethnic females (8.99) is not statistically significant).

**Linear Regression Models with Dummy and Continuous Variables**

The last situation we’ll address in this chapter involves linear regression models that include both dummy and continuous independent variables. As mentioned earlier, the well-known ANCOVA model is the analogue to this type of regression model. ANCOVA models serve several useful purposes. The main advantage is that we may compare groups regarding some variable after adjusting for (controlling for) their associations with some set of continuous variables. Recall in the last chapter that we discussed briefly the issue of confounding variables. Here’s another example of this phenomenon. Suppose we are studying rates of colon cancer among cities in the U.S. We draw a sample of cities and find that those in Florida, Arizona, and Texas have higher rates of colon cancer than cities in the Northeast, Midwest, and Northwest areas of the country (perhaps we have set up a series of dummy variables that indicate region of the country). Are there unique environmental hazards in warm weather cities that affect the risk of colon cancer? We cannot tell without much more information about the environmental conditions in these cities. Nevertheless, we have not yet considered a key confounding variable that affects analyses of most rates of disease: Age. It is likely that many warm weather cities – especially in so-called sunshine states such as Florida and Arizona – have an older age structure than cities in colder climates. Age is thus associated with both colon cancer rates and region of the country. ANCOVA models are designed to correct or adjust for continuous variables, such as age, that act as confounders.

Multiple linear regression models may be used in a similar manner. If we are interested in analyzing differences in some outcome among groups defined by dummy variables, we should consider adjusting for the possible association with conceptually-relevant continuous variables. Here is an example that is similar to our earlier dummy variable example, but changes the dependent variable. There is a perception in some circles that fundamentalist or conservative Christians tend to come from poorer families than do other people. Hence, we may surmise – and this is admittedly a crude hypothesis – that family income among fundamentalist Christians is lower on average than in other families in the U.S. Let’s test this
rough hypothesis using the gss96.sav data. Here are the results of a linear regression model with family income (fincome) as the dependent variable and a dummy variable (fundamental, coded 0 = non-fundamentalist, 1 = fundamentalist Christian). The model also includes gender as an independent variable.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>16.802</td>
<td>.155</td>
</tr>
<tr>
<td>fundamentalist Christian</td>
<td>-.672</td>
<td>.216</td>
</tr>
<tr>
<td>gender</td>
<td>-1.238</td>
<td>.200</td>
</tr>
</tbody>
</table>

a Dependent Variable: total family income in $10,000s

The model provides support for the hypothesis: On average, adults in the fundamentalist Christian category report 0.672 units less on the family income scale than do other adults. This coefficient is statistically significant at the $p = 0.002$ level, with 95% confidence intervals (not shown) of $\{ -1.097, -0.248 \}$. Let’s think about this association a bit more. Are there other variables that might account for this negative association? Without even knowing much about fundamentalist Christianity in the United States, I’m sure we can come up with some informed ideas. Let’s consider income. What accounts for differences in income in the United States? A prime candidate is education. If we were to read a bit about religion in the U.S., we might come across literature that suggests that conservative or fundamentalist Christians tend to have less formal education than others (although this has changed in recent decades).

A reasonable step is to include a variable that assesses formal education to determine whether it affects the results found in the previous model. The gss96.sav data set includes such a variable, educate, which measures the highest year of formal education completed by sample members. What happens when we place it in the model? Re-run the regression model in SPSS but include the education variable. You should discover the following results:
Note that the coefficient associated with the fundamentalist Christian variable is no longer statistically significant. Its \( p \)-value is now 0.301. We cannot determine with complete confidence that this coefficient (–0.217) is statistically distinct from the earlier model’s coefficient (–0.672) (once again, confidence intervals might help). But, given the literature on formal education among fundamentalist Christians in the U.S., we should strongly suspect that education confounds the association between considering oneself a fundamentalist Christian and family income.

To illustrate another important issue, we’ll cheat a bit and pretend that the fundamentalist Christian coefficient is statistically significant (although we could, perhaps, just ignore for purposes of this modest exercise the issue of statistical significance). How can we estimate means for different groups using this regression model? Hopefully, by now it is apparent that writing out the implied prediction equation is useful:

\[
\text{Predicted income} = 9.94 - 0.22(\text{fundamental}) - 1.3(\text{gender}) + 0.49(\text{educate})
\]

We’ve already seen what to do with the dummy variables: Simply plug in a zero or a one and compute the predicted values. But what should we do with education and its coefficient? Obviously we should do something similar with this variable as we did with the dummy variables: Plug in some value, compute the products, and add them up. Suppose we think that putting a zero or a one for education is the way to go. What would these numbers represent? Looking over the coding of
educate, perhaps by asking SPSS to construct a frequency table (use the drop-down menu Analyze-Descriptive Statistics-Frequencies), we find that it has a minimum value of zero and a maximum value of 20. But notice there is only one sample member who reported one year of education; the next value in the sequence is five. Therefore, placing a zero or a one in the equation is unwise since these values are not represented well in the data set. A value of 20 might be reasonable, but, for various reasons, it is better to use an “average” value of education in the computations (e.g., mean, median, or mode) or some other sensible value. For example, in the U.S. a well-recognized educational milestone is graduation from high school. This is normally denoted as 12 years of formal education. Twelve years is also the modal category in the distribution of educate. Hence, there is justification for using 12 in our computations if we wish to estimate means for the categories of fundamental and gender that are adjusted for years of education.

Here’s what the computations show us:

<table>
<thead>
<tr>
<th>Group</th>
<th>Computation</th>
<th>Adjusted Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamentalist males</td>
<td>9.94 – 0.22 + (0.49×12)</td>
<td>15.60</td>
</tr>
<tr>
<td>Fundamentalist females</td>
<td>9.94 – 0.22 – 1.3 + (0.49×12)</td>
<td>14.30</td>
</tr>
<tr>
<td>Non-fundamentalist males</td>
<td>9.94 + (0.49×12)</td>
<td>15.82</td>
</tr>
<tr>
<td>Non-fundamentalist females</td>
<td>9.94 – 1.3 + (0.49×12)</td>
<td>14.52</td>
</tr>
</tbody>
</table>

The average estimated difference in family income between fundamentalist Christians and other adults is 0.22, which is substantially smaller than the difference between males and females. Moreover, as we’ve already mentioned, a difference of 0.22 is not statistically significant. Nonetheless, it is instructive to compare the means after adjustment for education with the unadjusted means. These latter values, based on the first family income regression model, are shown in the table on the following page.
<table>
<thead>
<tr>
<th>Group</th>
<th>Computation</th>
<th>Expected Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fundamentalist males</td>
<td>16.8 – 0.67</td>
<td>16.13</td>
</tr>
<tr>
<td>Fundamentalist females</td>
<td>16.8 – 0.67 – 1.24</td>
<td>14.89</td>
</tr>
<tr>
<td>Non-fundamentalist males</td>
<td>16.8</td>
<td>16.80</td>
</tr>
<tr>
<td>Non-fundamentalist females</td>
<td>16.8 – 1.24</td>
<td>15.56</td>
</tr>
</tbody>
</table>

To estimate the difference between adjusted and unadjusted values, it is helpful to calculate the percentage difference. Choosing, for example, males, the means indicate that adjusting for the education reduces the predicted difference in family income between fundamentalists and non-fundamentalists by

\[
\left(\frac{(16.80 - 16.13) - (15.82 - 15.60)}{(16.80 - 16.13)}\right) \times 100 = \left(\frac{0.67 - 0.22}{0.67}\right) \times 100 = 67.2\% 
\]

You should notice that the slopes may be used to compute the percentage difference between the unadjusted and adjusted means.

Let’s consider one more dummy variable regression example. This time we’ll consider the association between marital status and family income. It is reasonable to hypothesize that married people have higher incomes, on average, than single people, whether they are never married, divorced, or widowed. Families with a married couple have at least the potential to earn two incomes; in fact, more than half of married couple families in the U.S. have two wage earners. Therefore, let’s run a regression model to assess the association between marital status and family income. As mentioned earlier, marital status is represented by a set of four dummy variables: married, widow, divsep (divorced or separated), and nevermarr (never married). We’ll use married as the reference category.

The resulting model has an $R^2$ of 0.176, an adjusted $R^2$ of 0.175, and an $F$-value of 134.929 ($df = 3, 1895; p < .001$). The regression coefficients or slopes support the hypothesis that married people report more family income than others. The average differences are three to four units with, it seems, never married people reporting the least family income.
However, think some more about variables that are associated with marital status and income. We’ve seen, for instance, that education is strongly associated with income, but it also may be associated with marital status. Another variable mentioned earlier is age. Is age related to income? The answer is yes since middle aged people earn more on average than younger people. What about age’s association with marital status? It makes sense to argue that there should be an association, especially when we have a never married category. Never married people tend to be younger, on average, than married people. Given these important issues, it is prudent to add education and age to the model. The table provided by SPSS is shown on the next page.

The new model, which we may call the full model to distinguish it from the earlier restricted model, has an $R^2$ of 0.271, an adjusted $R^2$ of 0.269, and an $F$-value of 140.829 ($df = 5, 1893; p < .001$). Although we’ll skip the computations, a nested $F$-test indicates that we’ve significantly improved the predictive power of the model by adding age and education (i.e., the $R^2$ has increased by a statistically significant amount). You should note that a nested $F$-test is not the same as subtracting the $F$-values from the two regression models. This latter procedure is not appropriate for comparing regression models.

Although we may have suspected that the results for marital status change when adding age and education to the model, there is actually little shift in the size of the coefficients. For example, the never married coefficient changes from $-4.112$ to $-3.922$, a difference of only about three percent. Hence, we may conclude that
the differences in family income associated with marital status are not accounted for (nor confounded) by age or education.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>9.795</td>
<td>.559</td>
<td></td>
<td>17.529</td>
</tr>
<tr>
<td>widowed</td>
<td>-3.044</td>
<td>.469</td>
<td>-.132</td>
<td>-6.485</td>
</tr>
<tr>
<td>divorced or separated</td>
<td>-3.130</td>
<td>.223</td>
<td>-.293</td>
<td>-14.065</td>
</tr>
<tr>
<td>never married</td>
<td>-3.922</td>
<td>.230</td>
<td>-.384</td>
<td>-17.058</td>
</tr>
<tr>
<td>age in years</td>
<td>.027</td>
<td>.008</td>
<td>.078</td>
<td>3.538</td>
</tr>
<tr>
<td>highest year of school completed</td>
<td>.492</td>
<td>.033</td>
<td>.298</td>
<td>15.121</td>
</tr>
</tbody>
</table>

*Dependent Variable: total family income in $10,000s*

There is one other issue that we should consider before closing this chapter. As mentioned in Chapter 3, many researchers use standardized coefficients to compare the associations within a model. For instance, since education and age are measured in different units, trying to compare their association with a dependent variable such as family income in a multiple linear regression model is difficult. By using the standardized coefficients, so the argument goes, education and age may be compared in standard deviation units. Although there is some merit to this approach, it falls apart when considering dummy variables. A one standard deviation unit shift in a dummy variable makes no sense (unless, I suppose, it equals exactly one, which doesn't happen with {0, 1} dummy variables). Dummy variables can shift only from zero to one or one to zero. So interpreting the standardized coefficients associated with dummy variables is not appropriate. For example, the variable *gender* has a mean of 0.50 and a standard deviation of 0.50. In the regression model shown earlier, the standardized coefficient associated with *gender* is -0.148. So the mindless interpretation of this coefficient is “Controlling for the effects of the other variables in the model, a one standard deviation unit
increase in gender is associated with a 0.148 standard deviation decrease in family income. A one standard deviation unit increase in gender shifts, for instance, from 0 to 0.50. This is a nonsensical value; there is no gender value of 0.50, nor can there be.
7 Specification Errors in Linear Regression Models

We discussed briefly in the last two chapters an important issue that concerns multiple linear regression models: We should carefully consider the variables we place in our models. As shown the Chapter 5, there are statistical tools to determine whether we have explained significantly more of the variance in the dependent variable by adding additional independent predictors. We’ve also learned about the importance of confounding variables. If we fail to consider variables that account (partially or fully) for the association between our key independent variables and the dependent variable, we may easily reach the wrong conclusion. Both of these issues relate to the topic of this chapter: Specification Errors. Specification error is a mistake when deciding the correct model to use for predicting the dependent variable. When establishing the term “correct model” we should consider several issues. First, is a linear model appropriate? Suppose we have the following association between an independent and a dependent variable:

\[
y \quad x
\]

A linear regression model is going to attempt, using the least squares formulas, to fit a straight line to this set of data. But it is clear that the association between \( x \) and \( y \) is not linear; rather, it is curved or \textit{curvilinear}. Estimating this association with a linear model is one example of specification error. We’ll discuss curvilinear associations in much more detail in Chapter 10.
For now, we’ll learn about three common sources of specification error in multiple linear regression models: (1)Including irrelevant variables; (2) leaving out important variables; and (3) misspecifying the “causal ordering” of variables within the model. The first problem is called *overfitting*; the second problem is called *underfitting*; and the third problem involves what is known as *endogeneity*, or *simultaneous equations bias*. All three present problems for linear regression models, but to a different degree. Before learning about these problems, though, it is important to discuss for a moment variable selection.

Variable selection is a frequently discussed topic in regression modeling. How do we know which variables to put in a regression model? There are a large number of variable selection procedures available. We’ll discuss some of these later in the chapter. At this point, it is important to realize that much of the work on variable selection by mathematical statisticians overemphasizes obtaining the best *prediction* equation at the expense of testing a reasonable *explanatory* model.

Regression models are clearly useful as predictive tools. Most of us would prefer, for instance, that medical researchers use statistical tools to predict with a high degree of certainty that a drug will cure us (or won’t injure us!). Perhaps we don’t care much about explaining the underlying biological mechanisms that link the drug with the cure. But in the social and behavioral sciences explanations are the driving force at the core of research. (In fact, some physicists argue that explanation is also at the core of their scientific endeavors. See, e.g., David Deutsch (1997), *The Fabric of Reality*, New York: Penguin Books.) We wish to be able to explain why one variable is associated with another (e.g., why are neighborhood unemployment rates associated with crime rates?) rather than to say only whether one predicts another (where’s the fun in that?). Therefore, when we select variables for a regression model, we should use our explanatory framework – our theory or conceptual model, as well as common sense – to decide which variables to include in the regression equation. This presupposes that we have a solid understanding of the research literature, especially previous studies that have examined our dependent variable, and that we understand fully the variables in our data set, including how they are constructed, what they purport to measure, and so forth.
**Overfitting – Or the Case of Irrelevant Variables**

It is very easy to find examples of overfitted models in the literature. Browse through an issue of just about any journal in the social and behavioral sciences and you will see regression models with variables that add nothing to their ability to predict the dependent variable. Of course, there may be various reasons for including seemingly irrelevant variables in a model. Perhaps most previous studies have included them. Often we seek to guard against omitting potential confounding variables and so include just about any variable we think might be important. This does not normally present a problem for linear regression models (except that it violates the sacrosanct principle of parsimony and can cause problems if we have a small sample size), but there are situations where we must be careful about putting too many variables in a model.

Overfitting does not bias the slopes; in other words, the slopes on average are still accurate. The main problem with overfitting is that it leads to inefficient standard errors. Recall from elementary statistics that an inefficient estimator has, on average, a larger variance than does an efficient estimator. To understand the consequences of overfitting, it may be useful to consider the standard errors of the slopes from a multiple linear regression model. In Chapter 3 we learned that the standard errors may be computed using the following equation:

\[
se(\hat{\beta}_i) = \sqrt{\frac{\sum(y_i - \hat{y}_i)}{\sum(x_i - \bar{x})^2(1 - R_i^2)(n - k - 1)}}
\]

The quantity \((1 - R_i^2)\) is called the *tolerance* and is estimated from an auxiliary regression equation that is defined as \(x_{i4} = \hat{\alpha} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik}\). Suppose that an extraneous variable \((x_4)\) included in the model is strongly associated with an important variable \((x_2)\) but is not associated with the dependent variable \((y)\). Then, all else being equal, the tolerance for \(x_2\) will be smaller (closer to zero) than if \(x_4\) were not in the model and \(x_2\’s\) standard error in the linear regression model will be relatively larger. The figure on the next page represents this situation.

The tolerance in the standard error computation for \(x_2\) is represented by the overlap between the two independent variables. If the overlap is large, then the
standard error can shift considerably, becoming large enough in some cases to affect the conclusions regarding the statistical significance of $x_2$’s coefficient. (Keep in mind that we want substantial variation (i.e., large circles) in the independent and dependent variables.) However, if there is not a statistical association between the independent variables (e.g., $\text{corr}(x_2, x_4) = 0$), then the standard error associated with the important variable ($x_2$) is unaffected. In the figure, this would be presented by non-overlapping circles between $x_2$ and $x_4$.

\[ y \hspace{1cm} x_2 \hspace{1cm} x_4 \]

**Underfitting – Or the Case of the Missing Variables**

A more serious situation arises when we have left important variables out of the model. In this situation, we may reach the wrong conclusions about our ability to predict – or explain, for that matter – the dependent variable. You may find this problem referred to as *omitted variable bias*. We saw this in Chapter 6 when we estimated the association between fundamentalist Christians and family income.
There appeared to be a significant association until we included education in the model. Then we learned that the association between fundamentalist Christians and family income is, as is often heard, spurious. As we have seen, we may also say that education confounded the association between this limited measure of religious affiliation and income. There is an assumption of multiple linear regression models that applies here. In Chapter 3 we learned the following:

The correlation between each independent variable and the error term – or residuals – is zero. If there is a correlation, we’ve left something important out of the model.

Keep in mind that the error term in a regression model includes all of the factors, random or otherwise, that are associated with the dependent variable. Therefore, it includes information from omitted independent variables. Suppose one of these omitted variables is associated with an independent variable that is included in the model. For instance, in the following equation we’ve omitted $x_3$ and $x_4$. If, say, $x_3$ is associated with $y$ and with $x_2$, then the slope of $x_2$ will be biased. In general terms, we have violated a key assumption of the model. Unfortunately, we cannot know for certain the direction of the bias: Is the slope too large or too small? Is the true slope negative or positive? The answer to these questions depends on the joint association among $x_2$, $x_3$, and $y$.

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

where $\epsilon_i = x_{3i} + x_{4i} + \text{random error}$

We saw an example of underfit in the last chapter, but let’s explore another. The gss96.sav data set includes a variable labeled lifesatis. This is a measure of life satisfaction that is based on some questions that ask about satisfaction in one’s marriage, at work, and in general. Higher values on this variable indicate a greater sense of life satisfaction. There are studies that suggest that life satisfaction is associated with education, occupational prestige, involvement with religious organizations, and several other variables. We’ll examine some of these associations in a linear regression model. (Note: This example is slightly revised from the following source: William D. Berry and Stanley Feldman (1985), *Multiple*...
Regression in Practice, Newbury Park, CA: Sage.) First, though, let’s look at a bivariate correlation matrix comprised of these variables. (Note: Values for life satisfaction are recorded for only 908 respondents in the gss96.sav data set. In order to analyze models using lifesatis, it is a good idea to restrict the analytic sample to only those respondents who have non-missing values on this variable. SPSS’s Data-Select Cases option allows you to filter out the observations that are missing on lifesatis, but be careful not to delete them from the data set. You may wish to use them later.)

The correlations suggest that all three of the proposed independent variables –
occupational prestige, education, and religious service attendance – are positively associated with life satisfaction. No pair of variables has a noticeably larger correlation than any other when we consider life satisfaction. But notice also that there is a pretty substantial correlation between education and occupational prestige ($r = 0.553$). This is not surprising: Highly educated people tend to be employed in positions that are considered more prestigious (e.g., judges, bank presidents, even college professors!); and occupations that require more formal education tend to be judged as more prestigious. Pay attention to education and occupational prestige as we consider a couple of multiple linear regression models. Unlike what we saw in the last chapter, we'll begin with the full model and work backwards.

The first thing to notice in the regression table on the next page is that, once we account for the association between occupational prestige and life satisfaction, the significant association between education and life satisfaction that appears in the correlation matrix disappears. The association between religious service attendance and life satisfaction remains, however. In this situation, we say that the association between education and life satisfaction is spurious; it is confounded by occupational prestige. One way to represent this is with the figure that appears on the next page. It shows that although it seems there is an association between education and life satisfaction (identified by the broken arrow), occupational prestige is associated with both of the other variables in such a way as to account completely for their presumed association.

Suppose, though, that for some reason (perhaps we failed to read the literature on life satisfaction carefully) we omit occupational prestige from consideration. Perhaps we think that highly prestigious occupations demand so much of people that they can’t possibly report high satisfaction with their lives. Or we simply don’t care much about employment patterns because, we argue, education and religious practices are much more important. The table on page 101 provides the regression results from a model that excludes occupational prestige.
## LINEAR REGRESSION ANALYSIS

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>39.839</td>
<td>2.904</td>
<td>13.720</td>
</tr>
<tr>
<td></td>
<td>occupational</td>
<td>.165</td>
<td>.049</td>
<td>.132</td>
</tr>
<tr>
<td></td>
<td>prestige</td>
<td>.168</td>
<td>.239</td>
<td>.027</td>
</tr>
<tr>
<td></td>
<td>highest year of</td>
<td>.892</td>
<td>.213</td>
<td>.137</td>
</tr>
<tr>
<td></td>
<td>school completed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>religious service attendance</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Dependent Variable: life satisfaction
In this model we find a highly significant association between education and life satisfaction. However, we know that this regression model is misspecified. We have a model that is underfit because it does not include a variable that is associated not only with the dependent variable, but also with education in such a way as to account for the presumed association between education and life satisfaction. Not only is the education slope inflated \((0.168 \rightarrow 0.613)\) when we omit occupational prestige from the model, but its standard error is also too small.

Another model to consider is one that omits religious service attendance. We have seen that this variable is significantly associated with life satisfaction whether we include occupational prestige or not. In fact, its slope, standard error, \(t\)-value, and \(p\)-value are similar in both models. But does it affect the other variables?

We see that the results aren’t much different than when religious service attendance is in the model (see the table on the next page). The adjusted \(R^2\) (not shown) is reduced from 0.039 to 0.022 (it was relatively small to begin with), but the slopes and standard errors associated with education and occupational prestige are not affected much. If you look back at the correlation matrix, you can see why: Religious service attendance has low correlations with the other two independent variables. Therefore, although this last model is also underfit, the consequences of specification error are less severe than when we leave out occupational prestige.

One lesson to learn from this exercise is that specification error is a virtual certainty. After all, we cannot hope to include all variables that are associated with the dependent variable in the model. The goal is to strive for models with a low or
manageable amount of specification error and hope that we don’t reach the wrong conclusions about the associations that the models do represent.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>42.271</td>
<td>2.871</td>
<td>14.725</td>
</tr>
<tr>
<td></td>
<td>highest year of school</td>
<td>.244</td>
<td>.241</td>
<td>.040</td>
</tr>
<tr>
<td></td>
<td>completed occupational</td>
<td>.162</td>
<td>.049</td>
<td>.130</td>
</tr>
<tr>
<td></td>
<td>prestige</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Dependent Variable: life satisfaction

The next question to ask is whether the full model, which includes education, religious service attendance, and occupational prestige, is overfit. After all, education is not significantly associated with life satisfaction once we consider occupational prestige. On the one hand, it may not hurt our conclusions to include it in the model. On the other, it may make the other standard errors larger than they should be. At this point the best way to check for overfit is to estimate a linear regression model that omits education and look over the slopes and standard errors. The results of such a model show that the slope for occupational prestige increases slightly when education is omitted. Nonetheless, unless we were extreme in our desire to figure out the precise association between occupational prestige and this measure of life satisfaction, it is a relative judgment call whether or not education needs to be in the model.

**Endogeneity or Simultaneous Equations Bias**

Recall that at the beginning of Chapter 2 we mentioned that another name for the dependent variable is the *endogenous* variable. The term *endo* refers to something that is inside or within. Thus, when referring to an endogenous variable or outcome we mean that it is produced within the system, or within the equation. Our linear regression equation, you’ve noticed by now, assumes that the dependent variable, but not the independent variables, are “produced” within the system. The
independent variables are considered *exogenous*, or produced by factors outside of the model or system. We may also say that the independent variables are *predetermined*; *pre* implying they are formed outside the system implied by the model.

But think about a set of variables such as occupational prestige, race/ethnicity, and life satisfaction. Models are often proposed that define these three variables, or many other sets of variables, as producing one another in some way. For example, we could revise our earlier life satisfaction model to include race/ethnicity:

\[
\text{Life satisfaction}_i = \alpha + \beta_1(\text{occup. prestige}_i) + \beta_2(\text{race/ethnicity}_i) + \varepsilon_i
\]

This model implies that occupational prestige and race/ethnicity independently combine to affect life satisfaction. But suppose that occupational prestige and race/ethnicity are not independent; rather, because of various historical and social factors that are associated with race/ethnicity in the United States and elsewhere, occupational prestige is, in part, a product of an individual’s race/ethnicity. In this situation, we say that occupational prestige is endogenous in the system specified by the equation. The problem for the linear regression model is that the estimated slope for occupational prestige in the above equation is biased. As an exercise, estimate a model with occupational prestige as the dependent variable and the race/ethnicity dummy variables (*AfricanAm*, *othrace*) as independent variables. Include education as a control variable. What do the results tell you about the possible endogeneity of occupational prestige?

A second endogeneity issue involves whether we have specified the correct ordering of variables. Does the dependent variable truly “depend” on the independent variables? Or could one or more of the independent variables “depend” on the presumed dependent variable? For instance, suppose we wish to estimate a model with adolescent drug use as the dependent variable and friends’ drug use as the independent variable. We may assume that one’s friends influence one’s behavior to a certain degree, thus leading to this model specification. But it may also be true that one’s choice of friends depends on one’s behavior. Hence, those who use drugs are more likely to choose friends who also use drugs. This issue implies that drug use and friends’ drug use are involved in what’s known as
We won’t pursue the issue of endogeneity in any more detail at this point because it involves (1) a thorny theoretical issue and (2) a complex statistical issue. It is a theoretical issue because we need to think carefully about the models we wish to estimate and consider whether one or more of the independent variables is potentially endogenous. When we estimate models with variables such as education, income, and life satisfaction; or one’s own behavior and one’s friends’ behavior; we should think about the order and intra-system relationships among the variables. Are the independent variables truly independent? Are they determined outside the model? Or could one or more “depend” on one or more of the other independent variables? Moreover, could variables outside the system affect the independent variables? The answer is almost always yes, so correct model specification and lack of endogeneity require a properly fit model.

Endogeneity is a complex statistical issue because it requires complicated systems of equations to specify correctly. We’ll learn in Chapter 8 about two statistical models, instrument variables with two-stage least squares and structural equation models, that are designed to address endogeneity in regression models.

How Do We Assess Specification Error and What Do We Do About It?

Clearly, specification error is a problem we wish to avoid. But, as implied earlier, considering all its sources is not an easy task. The best advice is to always strive for a clear and convincing conceptual model or theory to guide the analysis. Of course, conceptual models do not spring forth fully grown like Athene from Zeus’s skull. They are the result of previous studies and reasoned thought about the processes involved in creating an association between two or more variables. Therefore, you should not even begin to estimate regression models until you have read carefully the research literature that addresses the dependent variable. It is almost always a good idea to read the literature on your key independent variables also. You may find some unexpected variables that are associated with the independent variables you plan to use in your regression model.

Unfortunately, we cannot even hope to include all the variables that might be associated with the independent and dependent variables. So, as mentioned earlier, specification error is always lurking, we just hope to minimize it. There are,
however, some tools that are handy for assessing whether specification error is affecting the results of the regression model. The first such tool may be understood by considering, as an example, the equation shown earlier in the chapter:

\[ y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \]

where \( \varepsilon_i = x_{3i} + x_{4i} + \text{random error} \)

When we initially saw this equation, we were interested in what happens to the regression model when \( x_2 \) and \( x_4 \) are correlated. Although a correlation might affect the conclusions, how might we test whether \( x_4 \) has an influence if we do not measure it directly? Can we figure out a way to measure the error terms (residuals) so that we may at least indirectly examine the possible association with between \( x_2 \) and \( x_4 \)? As we learned in earlier chapters, we may use the predicted values from the model to compute the residuals:

\[ \{\text{residual}_i (\varepsilon_i) = (y_i - \hat{y}_i)\} \]

In Chapter 3 we saw that SPSS allows us to save the residuals from a linear regression model by using Save-Predicted Values in the Linear Regression drop-down menu and then checking the box labeled Residuals-Unstandardized. (You may also save the standardized residuals, which are the residuals transformed into z-scores; or the studentized residuals, which are the residuals transformed into Student’s t-scores.) We may then assess whether the independent variables included in our model are correlated with the residuals. Unfortunately, because of the way the OLS estimators are derived, there is rarely, if ever, a linear association between the \( x \)'s and the residuals. Sometimes, though, you will find a nonlinear association. We’ll learn more about examining potential nonlinear associations in Chapter 10.

There are several sophisticated tests for specification error. Most econometrics textbooks include many pages devoted to these tests. One of the more popular is called Hausman’s test. This test is used often to examine measurement error (see Chapter 8), one specific type of specification error. Another type of test designed to assess endogeneity is known as the Durbin-Wu-Hausman test. Unfortunately, to use the Hausman test in SPSS takes a lot of special programming. And, we must make several assumptions about our variables that may not be feasible. We’ll see an
example of the Durbin-Wu-Hausman test in Chapter 8.

Testing for underfit or overfit is much easier when you have access to all (or most) of the important independent variables in the data set. We obviously know that a regression model is underfit if we add a variable that is significantly associated with the dependent variable (a $t$-test or nested $F$-test may be used to confirm this; see Chapter 5). Moreover, some researchers recommend using the adjusted $R^2$ to compare models. This should be used only when one of the models is nested within the other. If we add, say, a set of independent variables and find that the adjusted $R^2$ increases, then we may conclude that the nested model is underfit. If, on the other hand, we take away a set of variables (perhaps because their associated $p$-values are large) and find that the adjusted $R^2$ does not decrease, then we may conclude that the model is overfit.

However, it is not a good idea to rely only on the adjusted $R^2$. It is better to combine an evaluation of this statistic with a nested $F$-test and a careful review of the slopes and standard errors of the remaining independent variables before concluding that a particular regression model is overfit or underfit. But we must also beware of overdoing it; of testing so many models by adding and subtracting various sets of independent variables that we cannot decide on the appropriate model.

**Variable Selection Procedures**

The aim of minimizing underfit and overfit is related to an important topic in regression modeling: Variable selection procedures. Unfortunately, many, if not most, statisticians’ guides to variable selection are so automated as to violate a point made earlier in this chapter: Let your theory or conceptual model guide the selection of variables. Nonetheless, these more automated procedures are discussed so often they’re worth a brief review.

The first type we’ll mention is known as *forward selection*. In this approach, independent variables are added to the model based on their correlations with the unexplained component of the dependent variable. Yet this relies on biased estimates at each point in the selection process, so it should be avoided. A type of forward selection that you might encounter is stepwise regression wherein partial $F$-statistics are computed as variables are removed and put back in the model. For
various reasons that we won’t go into, it should also be avoided. The second type is known as *backward selection*. In this approach, the analyst begins with all the independent variables in a model and then selectively removes those with slopes that are not statistically significant. This selection procedure is reasonable if you wish to have the best predictive model, but it is overly automated and takes the important conceptual work out of the hands of the researcher. Nonetheless, we’ll consider some steps that help make backward selection somewhat better than more arbitrary or biased approaches.

There are several steps to selecting variables if we wish to rely on backward selection (these are described in Kleinbaum et al., op. cit., Chapter 16). First, as mentioned earlier, estimate a regression model with all the independent variables. (Kleinbaum et al. suggest that we consider the higher order terms (e.g., $x^2$, $x^3$) and interactions in this model; we’ll save a discussion of these for Chapter 10.) Second, choose a fit measure to compare models. Third, use backward selection to remove variables – this may be done one at a time or in chunks – from the model. Fourth, assess the reliability of the preferred model with a split sample procedure. This last step is often reasonable if you wish to verify the reliability of a regression model regardless of how you decided which variables to include. If your software allows it (and SPSS does using the drop-down menu *Data-Select Cases-Random sample of cases*), select a random subsample of observations from the data set (normally 50%) for the variable selection stage and then, once you have chosen the best model, estimate it using the remaining subsample of observations. If the models are the same (or quite similar) then you have confidence that you’ve estimated the best fitting model.

These steps are fairly simple using statistical software. But we should say a few more words about the second step: Choosing a fit measure to compare models. We’ve seen a couple of these measures already: the adjusted $R^2$ and nested $F$-tests. If we compare the full model to its nested models using these approaches, we can guard against including extraneous variables in the model. Moreover, some statisticians recommend comparing the Mean Squared Error (MSE) or its square root ($S_E$) (see Chapter 4):

...
\[ \text{MSE} = \frac{SSE}{n-k-1} \quad \sqrt{\text{MSE}} = S_E \]

Looking for the smallest MSE or \( S_E \) from a series of nested models is a reasonable step since, assuming we wish to make good predictions and find strong statistical associations, we’d like to have the least amount of variation among the residuals.

There is another common measure that is related to these other measures: Mallow’s \( C_p \). An advantage of Mallow’s \( C_p \) is that it is minimized when the model has only statistically significant variables. Mallow’s \( C_p \) is computed as

\[ C_p = \frac{\text{SSE}(p)}{\text{MSE}(k)} - [n - 2(p+1)] \]

where \( p \) = number of predictors in restricted model

The \( \text{SSE}(p) \) is from the restricted model and the \( \text{MSE}(k) \) is from the full model. Using this measure does not require the model with all the independent variables and some model nested in it. It may include any two models that use subsets of the independent variables as long as at least one is nested within another. Nonetheless, Mallow’s \( C_p \) is used most often to compare the complete or full model with various models nested within it.

Here is an example of three models that use three of these model selection fit statistics: The adjusted \( R^2 \), \( S_E \), and Mallow’s \( C_p \). The models use the \( \text{usdata.sav} \) data set and are designed to predict violent crimes per 100,000 (\( \text{violrate} \)). Suppose, based on previous research and our own brilliant deductions, that we think that the following independent variables are useful predictors of violent crimes across states in the U.S.: The unemployment rate, population density, migrations per 100,000, per capita income in $1,000s, and the gross state product in $100,000s. But someone convinces us that we should use a backward selection procedure to test several combinations of these variables. The table on the following page shows what we find after estimating several linear regression equations.

The table shows that Model 2 provides the best fit. It has the largest adjusted \( R^2 \), the smallest \( S_E \), and the smallest Mallow’s \( C_p \). However, we did not test all the possible models, which would be a formidable task even with only five independent
SPECIFICATION ERRORS IN LINEAR REGRESSION

variables. (Try to figure out how many models we could fit.)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Adjusted $R^2$</th>
<th>$S_E$</th>
<th>Mallow’s $C_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 1:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>0.357</td>
<td>215.86</td>
<td>6.00</td>
</tr>
<tr>
<td>population density</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migrations per 100,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>per capita income</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gross state product</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model 2:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>0.371</td>
<td>213.59</td>
<td>4.06</td>
</tr>
<tr>
<td>population density</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Migrations per 100,000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gross state product</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model 3:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>0.311</td>
<td>223.40</td>
<td>6.34</td>
</tr>
<tr>
<td>gross state product</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent variable: Violent crimes per 100,000

SPSS offers a procedure that is similar to the one we just used. If you set up the model with all five independent variables in the Independent(s) box, select Method: Backward (the default is Enter) in the middle of the window, and click OK, SPSS returns an output window that estimates three models, the same models as Model 1 and 2 from our table, and a third that includes only three variables: The unemployment rate, the gross state product, and migrations per 100,000. The Model Summary information in the middle of the printout shows the Adjusted $R^2$ and $S_E$ (labeled Std. Error of the Estimate), which supports the conclusion that Model 2 is the best fitting model. It is interesting to note that choosing Method: Stepwise suggests that a model with the unemployment rate, migrations per 100,000, and the gross state product is the best fitting model. For some reason that you may wish to explore, the stepwise selection procedure (as well as the forward selection procedure) in SPSS omits population density from its preferred model. Yet, as mentioned earlier, most statisticians strongly advise against using stepwise selection.
The lesson from using these automated procedures is that some of them, such as backward selection, are reasonable tools if your goal is to come up with a model that includes the set of independent variables that offer the most predictive power. But these procedures cannot substitute for a good theoretical or conceptual model that not only predicts the dependent variable, but, more importantly, explains why it is associated with the independent variables. Moreover, they may provide misleading results because they are designed to fit the sample data, thus overstating how precise the results appear to be when we wish to make inferences to the population from which the sample was drawn. (For more on this point, see John Fox (1997), *Applied Regression Analysis, Linear Models, and Related Methods*, Thousand Oaks, CA: Sage, Chapter 13.)
8 Measurement Errors in Linear Regression Models

How do we measure concepts in the social and behavioral sciences? There are numerous things we might wish to measure that don’t have a clear definition, or that are difficult to examine in an unambiguous fashion. Concepts such as self-esteem, depression, happiness, social capital, antisocial behavior, aggression, impulsivity, political ideology, marital satisfaction, social support, and dozens of other phenomena don’t lend themselves well to the types of measuring instruments available to social scientists. We do a better job of measuring concepts such as monthly income in dollars, official rates of crime and suicide, or years of education, since these are more tangible and there are widely accepted ways of assessing the quantity of each. But even these are often measured poorly since we must rely on people to report them accurately. Sometimes, people do not report these things with care or there are sources of error that are out of their control.

Researchers often measure various concepts using self-report instruments. To measure adolescent drug use, for example, they usually ask a sample of adolescents to report if and how often they’ve used alcohol, cigarettes, marijuana, cocaine, or other illicit drugs. To measure happiness, researchers ask people whether they are happy in particular spheres of their lives (e.g., job, family), or they ask a global question such as “Do you consider yourself a happy person?” There are probably many more ways that happiness or some similar concept such as life satisfaction is measured, but generally we must be content with asking people about themselves. Only rarely is there an external instrument that is useful for verifying what these people tell us. An interesting exception involves recent research on illicit drug use. There are now widely available saliva and hair tests used to determine whether people have recently used illegal substances such as marijuana or cocaine. Of course, we could hook respondents up to lie detector machines – which are actually quite accurate – and then ask questions. This would likely deter most people from participating in surveys, however.
The literature on measurement in the social and behavioral sciences is huge and we cannot even begin to cover all the important topics. Our concern is with a narrow statistical issue – but one of utmost importance – that involves measurement. Known in general terms as measurement error or errors in variables, we are interested in learning about what happens to the regression model when the instruments used to measure the variables are not accurate. (For a much broader overview, see Paul Bieder et al. (Eds.) (2004), *Measurement Errors in Surveys*, New York: Wiley). There are various sources of measurement error, including the fact that we simply don’t have accurate measuring instruments (such as scales or rulers) (known as method error) or that people sometimes misunderstand the questions that researchers ask or don’t answer them accurately for some reason (lack of interest or attention, exhaustion from answering lots of questions; this is known as trait error). Sometimes, people don’t wish to tell researchers the truth or provide information about personal topics. One of the most personal topics, it seems, is personal income. If you ask questions about a person’s income, be prepared for many refusals or downright deceptions.

Another source of measurement error is recording or coding errors, such as when a data entry person forgets to hit the decimal key so an income of $1,900.32 per month becomes an income of $190,032 per month. Normally, this type of measurement error is easy to detect during the data screening phase. If we make it a point to always check the distributions of all the variables in our analysis using means, medians, standard deviations, ranges, box-and-whisker plots, and the other exploratory techniques, and are familiar with the instruments used to measure the variables, we will usually catch coding errors. (Don’t forget to always check the missing data codes in your data set! See Appendix A.)

But suppose that we have carefully screened our data for coding errors and fixed all the apparent problems. We may still suspect that there is measurement error in the variables for all the other reasons we’ve discussed so far. Is there anything we can do about it? We’ll discuss two situations that lead to distinct solutions: (a) the dependent variable, \( y \), is measured with error; (b) one or more of the independent variables, \( x \), is measured with error.

Before discussing these two situations, let’s revisit some of the assumptions (see Chapter 3) of linear regression analysis that relate to measurement error. First,
there is an assumption that is almost never followed in the social sciences that the $x$ variables are fixed (the technically proficient say nonstochastic). This means that researchers have control over the $x$ variables. They may then try out different values of the independent variable (such as by applying different amounts of fertilizer and water to plants and examining their growth) and determine with quite a bit of accuracy the association between $x$ and $y$. Second, we assume that the $x$ and $y$ variables are measured without error. We'll learn in this chapter what to do when this assumption is violated. Third, $x$ is uncorrelated with the error term. We saw one type of violation of this assumption in Chapter 7 when addressing specification error.

The first assumption is not normally met in the social sciences. There is certainly experimental research – especially in psychology, but also in sociology and economics – that manipulates independent variables, but most social science research relies on random samples to consider a variety of values of independent variables. There are important ethical limits faced by social and behavioral researchers. For example, we should not – perhaps cannot – manipulate an adolescent’s friends in studies of peers and delinquency. Another limitation involves the fact that there are many interesting variables that cannot be manipulated. For instance, we cannot control or manipulate a person’s race or gender. Although deliberate manipulation of the $x$’s is preferred, statistical control is often the best substitute.

Since the second assumption is the main focus of Chapter 7, much of our discussion in this chapter centers on the third assumption. We'll learn that measurement error can cause several problems for linear regression models. As with specification error, you should realize that we always have measurement error either because our concepts are not well defined or for some systematic reason (e.g., our instruments are not accurate enough). It is thus important that we seek to reduce it to a manageable level.

**The Dependent Variable is Measured with Error**

Suppose there is a dependent variable, such as happiness, that, in all likelihood, is measured with error. For now, we'll assume that our independent variables, say education and income, are measured without error. We'll furthermore assume that
somewhere within the conceptual measurement of happiness is a true measure, or what is often called the *true score*. Suppose, then, that we represent the dependent variable as the sum of two components: True score plus error:

\[ y_i^* = y(\text{true score})_i + u(\text{error})_i \]

Using these terms, what does the linear regression model look like? Here’s one way to represent it:

\[ y_i^* = \alpha + \beta_1x_{1i} + \beta_2x_{2i} + \varepsilon_i \quad = \quad y_i + u_i = \alpha + \beta_1x_{1i} + \beta_2x_{2i} + \varepsilon_i \]

Now, let’s use basic algebra to subtract \( u \) from both sides of the equation. By solving, we obtain:

\[ y_i = \alpha + \beta_1x_{1i} + \beta_2x_{2i} + \varepsilon_i + (\varepsilon_i - u_i) \]

Therefore, the error in measuring \( y \) can be considered just another component of the error term, or simply another source of error in our model. As long as this source of error is not associated systematically with either of the independent variables (i.e., \( \text{corr}(x_{1i}, u_i) = 0; \text{corr}(x_{2i}, u_i) = 0 \)), we have not violated a key assumption of the model and our slope estimates tend to be unbiased. But if the \( x \)'s are correlated with the measurement error in \( y \), then we have the specification error problems described in Chapter 7. Unfortunately, there is rarely any way to include this error explicitly in the regression model to solve the specification problem.

However, even when there is no correlation between the \( x \)'s and the error in \( y \), the \( R^2 \) tends to be lower than if measurement error is not present. In the presence of error in \( y \), the SSE is larger because there is more variation in the residuals that is not accounted for by the model. Moreover, the standard errors of the slope estimates are inflated, thus making it harder to find statistically significant slope coefficients. Recall that the standard error formula is

\[
\text{se}(\hat{\beta}_i) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{\sum (x_i - \bar{x})^2(1 - R_i^2)(n-k-1)}}
\]
Notice what happens when the SSE \(\sum(y_i - \hat{y}_i)^2\) is larger. All else being equal (or \textit{ceteris paribus}, to use an urbane Latin term favored by some social scientists), the standard error is also larger. All is not lost, though. If you obtain statistically significant slope coefficients, you then have a measure of confidence that you would find significant results even if there was no measurement error in \(y\).

\textbf{The Independent Variables are Measured with Error}

A scenario that is much worse is when one or more of the independent variables is measured with error. For now, let’s assume that the dependent variable is measured without error, so we have only \(y\) (true score) to deal with. But \(x\) as we observe it is made up of a true score plus error \(\{x_{1i} = x_{1i} + v_i\}\).

\[
y_i = \alpha + \beta_1 x_{1i}^* + \varepsilon_i = \alpha + \beta_1 (x_{1i} + v_i) + \varepsilon_i
\]

Distributing by the slope we have the following regression equation:

\[
y_i = \alpha + \beta_1 x_{1i} + \beta_1 v_i + \varepsilon_i
\]

So the error term now has two components \((\beta_1 v_i\ and \varepsilon_i)\) with \(x\) correlated with at least one of them. In this situation, the estimated slope is biased, usually towards zero. The degree of bias is typically unknown, though; it depends on how much error there is in the measurement of \(x\). The standard error of the slope is also biased. Usually the \(t\)-value is smaller (because of a larger standard error) than it would be if there was no measurement error in \(x\). Combining these two problems – smaller slopes and larger standard errors – you should immediately see that it is difficult to find statistically significant slopes when \(x\) is measured with error. This does not mean, however, that there is not a true statistically significant association between \(x\) and \(y\); only that you may not be able to detect it if there is sufficient error in \(x\) to bias the slope and standard error.

But let’s suppose that we have a multiple linear regression model and several of the independent variables are measured with error. It should be apparent that this compounds the problem, sometimes tremendously. Let’s make it even worse. What happens when both the dependent variable and the several of the independent
variables are measured with error? Now we have a truly bad situation: The $R^2$ is biased, the slopes are biased, and the standard errors are biased. We can’t trust much about such a model. Yet, it is not uncommon in the social and behavioral sciences to have measurement error in all of the variables. A major problem is we don’t often know if and how much measurement error we have. Fortunately, there are several techniques designed to address this problem.

**What To Do About Measurement Error**

We’ll discuss five (partial) solutions to measurement error. The first one is easy to state, yet slippery to implement: Have good measuring instruments. This is not meant as a flip response to an important question. It should not surprise you to learn that the development and testing of instruments for use in surveys and social and behavioral experiments is a huge area of research. It is mostly within the domain of psychology and cognitive studies, but also draws from a variety of fields such as sociology, marketing, political science, and opinion research (see, e.g., Jon A. Krosnick and Leandre R. Fabrigar (2005), *Questionnaire Design for Attitude Measurement in Social and Psychological Research*, New York: Oxford University Press). We have neither the space nor the ability in this brief chapter to go over the many aspects of good measurement, including various topics such as the validity and reliability of instruments. One piece of advice, though: It is better to use a well-established instrument to measure a phenomenon than to try to make one up for your study. For instance, if we wish to measure symptoms of depression, an instrument such as the Hamilton Depression Inventory is a better choice than trying to develop a brand new scale for a study.

The next two approaches are not normally very useful, but you may run into situations where knowing about them comes in handy. The first is known as inverse least squares (ILS) regression. Its logic is pretty straightforward. Suppose you wish to estimate a regression model with only one independent variable, but you know that this variable is measured with error. Fortunately, the dependent variable is not measured with error (an admittedly rare situation). In this situation, we may switch the independent and dependent variables and estimate a regression model:
\[ x_i = \alpha + \beta y_i \]

But is this the slope you wish to use? Probably not. Rather, remembering your elementary algebra classes, we may reorder the equation (after estimating the slope) so that the \( x \) appears on the right-hand side and the \( y \) appears on the left-hand side. In this manipulation, the slope for \( x \) becomes \( 1/\beta \) and it is not biased (although the standard error is not accurate). Note also that the intercept is not correct in this type of least squares regression.

The second method assumes that we have some external knowledge that is not normally available. First, though, we need to learn about a term in measurement theory known as reliability. In general usage, reliability refers to the reproducibility of a measuring instrument. The most common type is test-retest reliability, which is a quantitative approach to determining whether an instrument measures the same thing at time one that it measures at time two. Assuming a continuous variable, researchers usually use the Pearson’s correlation coefficient of the two measurement points to determine the reliability.

A common view of reliability is that it measures the ratio of the true score to the measured score. Using the notation introduced earlier, this is written as (the subscripts are omitted):

\[
r_{xx^*} = \frac{x(\text{true score})}{x(\text{true score}) + \nu(\text{error})} = \frac{s_x^2 x}{s_{xx^*}^2}
\]

In this equation \( r \) represents the reliability, not the Pearson’s correlation coefficient (although there are statistical similarities). The \( s_x^2 \) in the numerator and denominator are variance measures. It should be clear, and this agrees with what we have been claiming throughout this chapter, that reliability increases as the error in measurement decreases.

Although rare, we may occasionally have information on the reliability from other studies. Suppose, for example, that we are conducting a study that uses Mudd’s Scale of Joy to measure happiness among teenage boys and girls (this is not an actual scale). Several very sophisticated studies have found that when Mudd’s scale is used among this population, it yields a reliability coefficient \( (r_{xx^*}) \) of 0.8.
LINEAR REGRESSION ANALYSIS

(We won’t bother trying to figure out how these other researchers came up with this reliability coefficient.) Since errors in measuring the $x$ variable tend to bias the slope downward, our observed slope will be too small if $x$ is contaminated by measurement error. But we now know the degree of bias because of the reliability coefficient. Therefore, we may multiple the observed slope by the inverse of the reliability coefficient to obtain the unbiased slope. Assume our model suggests that high levels of joy reported by adolescents predict high GPA. We estimate a linear regression model and obtain an unstandardized slope of 1.45. Multiplying this slope by $1/0.8 = 1.25$ yields the unbiased slope of $\{1.45 \times 1.25\} = 1.81$. Unfortunately, this approach can become quite difficult to use when there is more than one independent variable.

**Latent Variables as a Solution to Measurement Error**

The term *latent* means hidden, concealed, or present but not revealed to the senses. It has been adopted in applied statistical analysis to refer to variables that exist in some indistinct sense, but are not detectable using observable methods. In the social and behavioral sciences, one might argue, virtually all variables are latent: Depression, anxiety, happiness, attitudes, antisocial behavior, prosocial behavior, etc. We do not see, hear, smell, taste, or touch these variables. Perhaps we can be confident in our ability to recognize phenomena, like Supreme Court Justice Potter Stewart when he famously wrote that he could detect pornography in motion pictures because “I know it when I see it.” Most of us, though, would likely have a difficult time coming up with a universal way to put fixed boundaries around the many phenomena of interest to the social and behavioral sciences. One solution to this slippery problem is to measure concepts indirectly through the use of variables that we can measure well. Measurement of concepts using latent variables provides the most common solution.

As the name implies, latent variables are thought to exist, but are hidden from our senses (actually, we assume they are hidden from the direct capabilities of our measuring instruments). The accepted logic is that we can measure aspects of some latent concept. By accumulating information from these measurable aspects, we may construct a latent variable. The statistical area that permits this approach is known as *multivariate analysis* because it is concerned with variance that is shared
among multiple variables. Under the broad area of multivariate analysis fall statistical techniques such as principal components analysis, factor analysis, latent variable analysis, log-linear analysis, and several other statistical procedures.

We'll discuss very briefly a frequently used technique known as factor analysis. Factor analysis is designed to take observed or manifest variables and reduce them to a set of latent variables that underlie the observed variables. There are many books that focus specifically on factor analysis (see, e.g., Paul Kline (1999), *An Easy Guide to Factor Analysis*, London: Routledge; David J. Bartholomew (1987), *Latent Variable Models and Factor Analysis*, New York: Oxford University Press; and Bruce Thompson (2004), *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications*, Washington, DC: American Psychological Association).

Perhaps the simplest way of understanding this statistical technique is to consider that variables that are similar share variance. If there is sufficient shared variance among a set of observed variables, we claim that this shared variance represents the latent variable. Conceptually speaking, we assume that the latent variable predicts the set of observed variables. The diagrams on the next two pages represent two simplified models of a latent variable. The first shows the overlap of three observed variables; the shaded area of overlap represents the latent variable, or the area of shared variability. The second diagram shows how latent variables are often represented in books and articles. Notice the direction of the arrows. They imply that the latent variable (recall it is not directly observed) predicts the observed variables.

SPSS provides several factor analysis techniques. However, when used to measure latent variables, many researchers rely on specialized software such as AMOS, LISREL, COSAN, EQS, Latent Gold, or MPlus to conduct the analysis. An advantage of these programs is that they will simultaneously estimate factor analyses to come up with latent variables and regression models to test the associations among a set of latent variables. For example, if we used 10 observed variables to measure aspects of anxiety, assuming that a latent variable – or the true measure of anxiety – explains their shared variance, and 12 variables to measure self-esteem, then a latent variable analysis program allows us to test the measurement properties of these two latent variables and their statistical association using a regression model. One of the advantages of this approach – which is often
termed *structural equation modeling* – is that it may be used to estimate more than one regression equation at a time, or what are called *simultaneous equations*. This provides one solution to the endogeneity problem discussed in Chapter 7.

But, you may be asking, what happened to the measurement error in the latent variable? The idea is that the shared variance represents the true score of a latent variable, whereas the remaining unshared variability represents the error. In the diagram below, the shaded area where all three variables overlap represents the true score for the latent variable and the non-shaded areas represent measurement error. Of course, we must make the assumption that the shared variance truly represents something that is significant and that we have correctly labeled it.
The second latent variable approach has been popular in econometrics. (It is still used, but not as often as in the past.) In this approach, the latent variables are based on instrumental variables. An example is the easiest way to describe instrumental variables. Suppose we have an independent variable that is measured with error. We know that if we place it in a regression model, the slopes and standard errors will be biased. However, we think we have found another set of variables that have the following properties: (a) they are highly correlated with the error-contaminated independent variable; and (b) they are uncorrelated with the error term (which, you'll recall, represents all the other influences on \( y \) that are not in the model). If we can find such a set of variables – which are known as instrumental variables – we may use them in an OLS regression model to predict the independent variable, use the predicted values to estimate a latent variable as a substitute for the observed independent variable, and use this latent variable in an OLS regression model to predict \( y \). One common estimation technique that uses instrumental variables is known as two-stage least squares because it utilizes the two OLS regression stages just described. We assume that the latent variable is a linear
combination of the instruments, thus we can estimate the association with a linear model.

Here are the steps in a more systematic fashion. We’ll use the letter \( z \) to represent the instrumental variables and \( x \) to present the error-contaminated independent variable. The steps in two-stage least squares are as follows (the \( i \) subscripts are omitted):

1. Estimate \( x_i = \alpha_p + \beta_{p1}z_i + \beta_{p2}z_2 + \cdots + \beta_{pq}z_q \).
2. Save the predicted values from this OLS regression model.
3. Using OLS, estimate \( y = \alpha_s + \beta_{s1}x'_1 + \beta_{s2}x_2 + \cdots + \beta_{ss}x_s \); where \( x'_1 \) denotes the predicted values from (1).

The slopes from model (3) are unbiased (we’ll assume that \( x_2 \) through \( x_s \) and \( y \) are measured without error). However, the standard errors from a two-stage least squares are biased, so special techniques are required. Therefore, it is not a good idea to go through these steps on your own. Rather, use a program such as SPSS that includes a two-stage least squares routine.

An important problem that makes the use of two-stage least squares of limited utility is that it is difficult to find instrumental variables that meet the requirements outlined earlier. Most variables that are associated strongly with the independent variable are, first, also contaminated by error, and, second, are associated with the error term. A serious problem arises when the instrumental variables are contaminated by error or do not predict the \( x \) variable well: The slopes and standard errors from a two-stage least squares regression model are even worse than the estimates from an OLS regression model.

Although instrumental variables may be of limited utility because they are so difficult to find, let’s see an example of a two-stage least squares model in SPSS. The fact that it is hard to find instrumental variables is illustrated by the artificiality of this example. We’ll use the \textit{gss96.sav} data set and make the following (poor) assumptions: (a) Life satisfaction (\textit{lifesatis}) is measured without error; (b) the true score of occupational prestige predicts life satisfaction, but the observed variable (\textit{occprest}) is contaminated by measurement error; (c) the two instrumental variables, education (\textit{educate}) and personal income (\textit{pincome}), are measured without
error; and (d) education and personal income are strong predictors of occupational prestige but are uncorrelated with the error in predicting life satisfaction. So, in this regression model, we hypothesize that higher occupational prestige is associated with higher life satisfaction, but that we should use the instrumental variables in a two-stage least squares analysis to reduce the effects of measurement error.

The first step is to estimate the correlations among the independent variable and the instrumental variables. SPSS's Correlate-Bivariate provides these correlations. The table indicates that the three variables are moderately correlated, with the largest correlation between occupational prestige and education. In Chapter 7 we explored this association from a different angle. Nonetheless, at this point education and income appear to be reasonable instruments for occupational prestige. Next, we'll see the results of the OLS regression model with life satisfaction as the dependent variable and occupational prestige as the independent variable.

<table>
<thead>
<tr>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>occupational prestige</td>
</tr>
<tr>
<td>occupational prestige</td>
</tr>
<tr>
<td>personal income in $10,000s</td>
</tr>
<tr>
<td>highest year of school completed</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Coefficients(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

a Dependent Variable: life satisfaction
The $R^2$ for this model is 0.023, with a standard error of the estimate ($S_e$) of 16.742. The explained variance is low, but statistically significant ($F = 21.363, p < .001$). The coefficient for occupational prestige is also statistically significant, with, as expected, higher occupational prestige associated with higher life satisfaction. However, we assumed earlier that occupational prestige is measured with error. So let’s see what a two-stage least squares regression analysis provides.

In SPSS, this technique is estimated using the drop-down menu Analyze-Regression-2-Stage-Least-Squares. The window that opens includes boxes for the dependent, independent, and instrumental variables. After running the model, SPSS provides the following output (I’ve edited it to make the results a bit more manageable).

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE B</th>
<th>Beta</th>
<th>T</th>
<th>Sig. T</th>
</tr>
</thead>
<tbody>
<tr>
<td>occupational prestige</td>
<td>.26567</td>
<td>.0713</td>
<td>.21254</td>
<td>3.724</td>
<td>.0002</td>
</tr>
<tr>
<td>(Constant)</td>
<td>40.96852</td>
<td>3.2829</td>
<td>12.479</td>
<td>.0000</td>
<td></td>
</tr>
</tbody>
</table>

a Dependent Variable: life satisfaction

This model has an $R^2$ of 0.015, with a standard error of the estimate of 16.774. The $F$-value is 13.869, which is statistically significant ($p < .001$). The key issue is whether the slope changes from the OLS to the two-stage least squares model. If there is measurement error in occupational prestige and if the instruments are useful predictors of its true score, the slope should increase. It does and the proportional increase is substantial: 0.190 → 0.266. Thus, the slope increases by about 40% after adjusting the independent variable for measurement error.

There is a relatively simple test – identified in Chapter 7 as the Durbin-Wu-Hausman test – that allows us to determine if the instrumental variables approach is preferred. To implement this test, we first estimate a linear regression model with the instruments; in the above model this is

$$\text{Occupational Prestige}_i = \alpha + \beta_1 (\text{Education}_i) + \beta_2 (\text{Personal Income}_i) + \epsilon_i$$
Second, save the unstandardized residuals from this model; in this example, we’ll label them $res\_occ$. Third, estimate a linear regression model with the original dependent variable, the original independent variables, and the unstandardized residuals. In our example, this model is

$$\text{Life satisfaction}_i = \alpha + \beta_1 (\text{Occupational Prestige}_i) + \beta_2 (res\_occ_i) + \epsilon_i$$

If the coefficient involving the residuals is statistically significant, then there is evidence that the instrumental variables–two-stage least squares model is preferred to the OLS regression model. In the example of the model designed to predict life satisfaction, the $t$-value of $res\_occ$ is $-1.290$, which has a $p$-value of $0.197$. Since it is not statistically significant, there is no evidence that the two-stage least squares model offers any benefit over the OLS regression model. (For more information on this test, see Russell Davidson and James G. Mackinnon (1993), *Estimation and Inference in Econometrics*, New York: Oxford University Press.)

Yet, you should ask yourself, are the assumptions we made before estimating the model reasonable? Can you think of potential sources of error in either of the instruments? Suppose that you were told that occupations considered prestigious are those that require more education or that pay, on average, higher salaries? Would this change your judgment of education and personal income as instrumental variables? You may decide that these are not good instruments in this model. But finding instrumental variables is typically problematic and you will be forced to justify your choices based on reasoning that is divorced from concrete statistical evidence. If you are not content with the instrumental variables approach, there are other errors-in-variables models that researchers use, but they are beyond the scope of this presentation.

It should be obvious by now that measurement error is an almost universal problem in social science statistics. It may be thought of as one type of specification error in a linear regression model since it involves extra information (errors in measuring) that we hope to exclude from the model. Whether there will ever be tools to minimize it sufficiently so that we may convince all or most people that our models are valid is, at this time, an unanswerable question. Nonetheless, measurement techniques continue to improve as researchers study the properties of
measuring tools and the way that people answer survey questions or report information in general. (Excellent guides to constructing survey questions are Norman Bradburn et al. (2004), *Asking Questions: The Definitive Guide to Questionnaire Design*, San Francisco: Jossey-Bass; and Krosnick and Fabrigar, *op. cit.*.) At this point, we should simply do our best to measure social and behavioral phenomena accurately and reliably. Moreover, some useful tools for reducing the damaging effects of measurement error exist and are readily available to even the novice researcher. A review of the literature on factor analysis, multivariate analysis, and structural equation modeling will provide a guide for using these tools.
9 Collinearity and Multicollinearity

Assumption Five in Chapter 3 states that there is no perfect collinearity among the independent variables. The two components of the word collinearity are co, or together, and linear, or the implication of a straight line or a flat plane. This latter component is tied to the geometric relationship between two variables in space. Since we have not discussed the geometric bases of statistical analysis, it is best to refer to collinearity as an issue of covariance. We are interested in the covariability of two or more independent variables.

We’re not, at this point, interested in the covariability of any particular independent variable with the dependent variable. It is the hope of most researchers to have high covariability between at least one of the independent variables and the dependent variable. When we speak about collinearity, we’re interested in the covariability among the independent variables. Recall that in a couple of chapters we’ve represented covariability using overlapping circles. Statistical control in linear regression and covariability among several variables to demonstrate a latent variable were represented by overlapping circles. Yet they also come in handy for representing collinearity.

The figure on the following page represents what happens when two independent variables are highly correlated. Suppose we wish to determine the unique association between \( x_1 \) and \( y \), but \( x_1 \) and \( x_2 \) have this type of association. Where is the association of \( x_1 \) and \( y \) controlling for the association with \( x_2 \)? It is the thin sliver where \( x_1 \) overlaps with \( y \) but does not include any portion of \( x_2 \). But suppose we wish to estimate this overlap with a linear regression model. Do you see any potential problems? Recall, again, the equation for the standard error of the slope (perhaps you’ve memorized it by now!):

\[
\text{se}(\hat{\beta}_i) = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{\sum(x_i - \bar{x})^2(1 - R_i^2)(n-k-1)}}
\]
What will the tolerance \((1 - R_i^2)\) be if there is such a large overlap between \(x_1\) and \(x_2\)? We cannot be certain with a diagram, but rest assured that it will be large. What happens to the standard error of the slope as the tolerance increases? It becomes larger and larger. Suppose, for example, that the correlation between \(x_1\) and \(x_2\) is somewhere on the order of 0.95. Then, assuming an OLS regression model with only two independent variables, the tolerance is \(1 - 0.95^2\) = 0.0975. Next, take the correlation between another sample’s \(x_1\) and \(x_2\) that is much lower, say 0.25. The tolerance in this situation is \(1 - 0.25^2\) = 0.4375. Try plugging fixed values into the standard error equation for the other quantities with these two tolerances and you’ll see the effect on the standard errors.

The assumption of OLS regression, though, states that there is no perfect collinearity. Perfect collinearity is represented by completely overlapping circles, or a correlation of 1.0. (It could also be −1.0 and have the same consequences.) What happens to the standard error when there is perfect collinearity? The tolerance in this situation is \(1 - 1\), which means that the denominator in the standard error
COLLINEARITY AND MULTICOLLINEARITY

Collinearity and multicollinearity are concepts in statistics that refer to the phenomenon where two or more variables in a regression model are highly correlated. When this occurs, it can lead to problems in estimating the regression coefficients and their standard errors. The statistical theory literature emphasizes that the main problem incurred in the presence of high or perfect collinearity is biased or unstable standard errors. Slopes are considered unbiased, at least when the model is estimable. Yet, practice strongly indicates that slopes can be highly unstable when two (or more) variables are highly collinear.

Let's consider the usdata.sav data set which has three variables that we suspect are associated with violent crimes per 100,000: Gross state product in $100,000s (gsprod), the number of suicides per 100,000 (sucrate), and the age-adjusted number of suicides per 100,000 (asuicrat). Yet there is something peculiar about these variables: Two of them address the state’s number of suicides. Demographers and epidemiologists are well aware that several health outcomes, such as disease rates or the prevalence of some health problems, are influenced by the age distribution. Some states, for instance, may have a higher prevalence of particular types of cancer because they have, on average, older people. Since older people are more likely to suffer from certain types of cancer, states with older populations will have more cases of these cancers. Therefore, demographers often compute the age-adjusted prevalence or rate. This, in essence, controls for age before regression models are estimated. If age has a large association with the outcome, then the original frequency of the outcome and the age-adjusted frequency might differ dramatically. A correlation matrix will provide useful evidence to judge the association between the age structure of states and number of suicides.

The correlation between the number of suicides and the age-adjusted number of suicides is 0.990 (p<.001), which is close to the maximum value of one. Hence, we have two variables with much of their variability overlapping. Moreover, it appears that the age structure is only modestly associated with suicide. But suppose we wish to enter these two variables in a linear regression model designed to predict violent crimes per 100,000. This is clearly an artificial example since the
two suicide variables are measuring virtually the same thing. But let’s use it to explore the effects of collinearity in a linear regression model.

### Correlations

<table>
<thead>
<tr>
<th></th>
<th>Suicide rate per 100,000</th>
<th>Age Adjusted Suicide Rate</th>
<th>Gross State Product in $100,000s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suicide rate per 100,000</td>
<td>1.000</td>
<td>.990**</td>
<td>-.369**</td>
</tr>
<tr>
<td>Age Adjusted Suicide Rate</td>
<td>.990**</td>
<td>1.000</td>
<td>-.356*</td>
</tr>
<tr>
<td>Gross State Product in $100,000s</td>
<td>-.369**</td>
<td>-.356*</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2-tailed).**
*Correlation is significant at the 0.05 level (2-tailed).

We’ll estimate a linear regression model using these three variables as independent predictors, with violent crimes per 100,000 as the dependent variable.

### Coefficients(a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>269.909</td>
<td>156.797</td>
<td>1.721</td>
</tr>
<tr>
<td></td>
<td>Suicides per 100,000</td>
<td>-33.149</td>
<td>73.185</td>
<td>-.412</td>
</tr>
<tr>
<td></td>
<td>Age Adjusted Suicides</td>
<td>42.718</td>
<td>69.992</td>
<td>.553</td>
</tr>
<tr>
<td></td>
<td>Gross State Product in $100,000s</td>
<td>84.060</td>
<td>22.142</td>
<td>.523</td>
</tr>
</tbody>
</table>

a Dependent Variable: Violent crimes per 100,000

Look at the unstandardized coefficients for the two suicide variables: One is positive and one is negative. Although statistical theory suggests that these slopes are unbiased, you can see they are untenable. Both variables are measuring the same phenomenon, yet they indicate that as one increases violent crimes tend to decrease, whereas as the other increases violent crimes tend to increase. Clearly,
Both interpretations of the association between suicides and violent crimes cannot be correct.

Here is another linear regression model that omits one state: Utah.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>270.005</td>
<td>158.166</td>
<td>1.707</td>
<td>.095</td>
</tr>
<tr>
<td>Suicides per 100,000</td>
<td>-17.934</td>
<td>81.036</td>
<td>-.223</td>
<td>-.221</td>
</tr>
<tr>
<td>Age Adjusted Suicides</td>
<td>28.506</td>
<td>77.197</td>
<td>.370</td>
<td>.369</td>
</tr>
<tr>
<td>Gross State Product in $100,000s</td>
<td>83.883</td>
<td>22.339</td>
<td>.523</td>
<td>3.755</td>
</tr>
</tbody>
</table>

The coefficients and standard errors for the two suicide variables have shifted quite a bit from one model to the next. Although this may be because the state of Utah has unusual values on one or more of the variables in the model (see Chapter 12), it might also be the result of collinearity between the two suicide variables. It is not uncommon for collinearity to create highly unstable coefficients that shift substantially with minor changes in the model.

Of course, it is unlikely that we would have estimated a model with two variables that are so similar. Even though we used these models to illustrate a point, it is important to always consider whether two (or, as we shall see, more) independent variables are highly correlated before estimating the model. Now, let’s see what happens when we include only one suicide variable in the OLS regression model that predicts violent crimes.

The results of this model (see the table on the next page) show that there is a positive association between the number of suicides (age-adjusted) per 100,000 and the number of violent crimes per 100,000 after adjusting for the association with the gross state product. Although it is not statistically significant, notice how different the coefficient is from those in the earlier models. More importantly,
notice the standard error in this model: It is much smaller than the standard errors from the earlier models. This demonstrates how collinearity inflates standard errors substantially.

<table>
<thead>
<tr>
<th>Coefficients(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Multicollinearity

Recall that the tolerance from the auxiliary regression equation is based on the proportion of variance among one independent variable that is explained by the other independent variables. In other words, it is the $R^2$ from a linear regression model. I hope it is clear by this point that we should consider overlapping variability between not only two variables, but among all the independent variables. Suppose, for example, that we have four independent variables, $x_1, x_2, x_3,$ and $x_4$. The largest bivariate correlation between any two variables is 0.4. If we were to use only the two variables with the largest correlation in a linear regression model, the tolerance in the standard error formula would not be any smaller than $(1 - 0.4^2 = ) 0.84$. It is unlikely that a tolerance of this magnitude would bias the standard errors to an extreme degree or make the slopes unstable. However, it is possible for one of these independent variables to be a linear combination of the others. If, for instance, $x_4$ is predicted perfectly or nearly perfectly by $x_1, x_2,$ and $x_3$, then the tolerance will approach zero and the same problems result: Biased standard errors and unstable slopes. In terms of our overlapping circles (see the figure on the next page), multicollinearity appears as two or more variables encompassing another independent variable.
It should be clear, though, that looking at bivariate correlations will not always be sufficient for determining whether multicollinearity is present. Any of the bivariate correlations may be below thresholds that warn of collinearity, but they won't show if some combination of independent variables predicts another independent variable. We must use other diagnostic tools.

**How To Detect Multicollinearity**

Since we have already mentioned that bivariate correlation matrices are of limited use for detecting multicollinearity (although they are useful for determining collinearity), we should find some other diagnostic tools. First, though, it is worth mentioning a rule-of-thumb that has been used to gauge collinearity: Estimate a correlation matrix and then look for bivariate correlations among the independent variables that exceed 0.7. Any correlation that matches or exceeds 0.7 should lead you to suspect that collinearity may affect your results. (This depends on the sample size, though.)

But what about multicollinearity? Is there a test that will allow us to determine its presence? How would one go about devising such a test? You were given a hint when we discussed the tolerance formula in the standard error equation. In
Chapter 3, we mentioned that the tolerance was based on the auxiliary regression model that includes $x_i$ as the dependent variable and all the other independent variables as predictors (e.g., $x_{ii} = \hat{\alpha} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_i x_{ii}$). Can we use this information to assess the possibility of multicollinearity? The easy answer is “Of course!” We may simply regress each independent variable on all the others and compute an $R^2$ for each model. We need a rule-of-thumb, though: How high must an $R^2$ be before we suspect multicollinearity? Perhaps we should use the same rule-of-thumb as mentioned above: 0.7. However, since we square the $R$, perhaps it should be 0.49. Although there are rules-of-thumb that specify the threshold for the auxiliary $R^2$ and the tolerance, a more frequently used statistic is known as the Variance Inflation Factor (VIF). For various reasons we don’t have time to discuss, the VIF has become the test de jour for assessing multicollinearity. It is defined as the inverse of the tolerance:

$$VIF = \frac{1}{1 - R^2_{ii}}$$

The rule-of-thumb(s) is that a VIF greater than or equal to 10 (some say it is nine) is indicative of multicollinearity. Others argue that if the square root of the VIF is greater than or equal to three, we should suspect multicollinearity (see John Fox (1991), Regression Diagnostics, Newbury Park, CA: Sage Publications). A relevant issue that is not often discussed, though, is sample size: Linear regression models are affected less by collinearity or multicollinearity as the sample size increases. Hence, all else being equal, a model that uses a sample size of 10,000 is less sensitive to multicollinearity – at least using conventional rules-of-thumb – than a model that uses a sample size of 100. Unfortunately, the decision rules that have emerged for these tests do not normally consider the sample size.

Another standard test for multicollinearity involves condition indices and the variance proportions. Before demonstrating how to use this test, it is useful to take a quick journey into the intersection of linear algebra and principal components analysis for some terminology. An eigenvalue is the variance of the principal component from a correlation (or covariance) matrix. The principal components of a set of variables are a reduced set of components that are linear combinations of the
variables. They are one type of latent variable (see Chapter 8), yet the principal components from a set of variables are uncorrelated with one another. Hence, if we have a set of, say, 10 independent variables, we may reduce them to a set of, perhaps, two or three principal components that are uncorrelated with one another. The eigenvalues of these principal components measure their variances. Briefly, then, eigenvalues are a measure of joint variability among a set of independent variables. As the eigenvalue gets smaller, there is more shared variability among the independent variables, with a value of zero indicating perfect collinearity. An interesting property of eigenvalues in the following test is that they sum to equal the number of predictors in the regression equation (number of independent variables + the intercept, or $k + 1$).

But how is this useful for judging multicollinearity? We may use eigenvalues to compute condition indices (CIs; not to be confused with confidence intervals), which are another measure of joint variability among independent variables. CIs are used in conjunction with variance proportions to assess multicollinearity in linear regression models (they may also be used in other regression models). Variance proportions assess the amount of the variability in each independent variable (as well as the intercept) that is accounted for by the dimension (or principal component). Fortunately, SPSS will compute these values for us. Let's see an example of VIFs, eigenvalues, CIs, and variance proportions. In SPSS’s Analyze-Regression-Linear drop-down menu click on the box labeled Statistics. Next, place a check in the box labeled Collinearity diagnostics. Estimate the linear regression model we ran earlier that included both suicide variables. The results appear in the tables on the following two pages.

The collinearity statistics reported in the second table show the tolerances and the VIFs. You can see that two of these greatly exceed the rule-of-thumb that any VIF of 10 or more indicates collinearity or multicollinearity problems. But we may also use the CIs and variance proportions in third table to assess collinearity. Here’s how it’s done. First, look for any CI that is greater than or equal to 30. Second, scan across the row that contains this CI and look for any variance proportion that is greater than or equal to 0.50. The variables with variance proportions greater than or equal to 0.50 in these rows are involved in a collinearity or multicollinearity problem and should be checked further.
LINEAR REGRESSION ANALYSIS

Coefficients\(^{(a)}\)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>269.909</td>
</tr>
<tr>
<td></td>
<td>Suicides per 100,000</td>
<td>-33.149</td>
</tr>
<tr>
<td></td>
<td>Age Adjusted Suicides per 100,000</td>
<td>42.718</td>
</tr>
<tr>
<td></td>
<td>Gross State Product in $100,000s</td>
<td>84.060</td>
</tr>
</tbody>
</table>

\(^{(a)}\) Dependent Variable: Violent crimes per 100,000

Collinearity Statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>.020</td>
</tr>
<tr>
<td></td>
<td>Suicides per 100,000</td>
<td>.020</td>
</tr>
<tr>
<td></td>
<td>Age Adjusted Suicides per 100,000</td>
<td>.859</td>
</tr>
<tr>
<td></td>
<td>Gross State Product in $100,000s</td>
<td>.859</td>
</tr>
</tbody>
</table>

In the collinearity diagnostics shown on the next page there is one CI, associated with Dimension 4, which exceeds 30; it is 72.374. This CI is computed by taking the largest eigenvalue (Dimension 1’s = 3.379), dividing it by Dimension 4’s eigenvalue (it is rounded to .001, but is actually 0.00064845), and taking the square root of the quotient:

$$\sqrt[\frac{\text{Eigenvalue of Dimension 1}}{\text{Eigenvalue of Dimension 4}}} = \sqrt[\frac{3.379}{0.00064845}] \approx 72.3$$
COLLINEARITY AND MULTICOLLINEARITY

The equation on the left shows that this is how CIs are generally computed, with the largest eigenvalue divided by one of the other's to yield the squared value of the CI. Of course, this guarantees that the CI for Dimension 1 is always 1.00.

Collinearity Diagnostics (a)

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
<th>Variance Proportions</th>
<th>Gross State Product in $100,000s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Suicides per 100,000</td>
<td>Age Adjusted Suicides per 100,000</td>
</tr>
<tr>
<td>1</td>
<td>3.397</td>
<td>1.000</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>.570</td>
<td>2.442</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>.033</td>
<td>10.105</td>
<td>.01</td>
<td>.01</td>
</tr>
<tr>
<td>4</td>
<td>.001</td>
<td>72.374</td>
<td>.99</td>
<td>.99</td>
</tr>
</tbody>
</table>

a Dependent Variable: Violent crimes per 100,000
(Note: The column labeled (Constant) has been omitted from the printout.)

Now that we have identified a CI that exceeds 30, look across its row for variance proportions that are greater than or equal to 0.50. There are two and they are associated with the number of suicides per 100,000 (variance proportion = 0.99) and the age adjusted number of suicides per 100,000 (variance proportion = 0.99). When a large variance proportion is accounted for by a dimension and it is associated with a CI greater than 30, then we know that multicollinearity is present. For more information on using eigenvalues, CIs, and variance proportions, see Kleinbaum et al., *op cit.*, Chapter 12.

Before moving on to learn some solutions to collinearity and multicollinearity problems, it is instructive to look at one more example because occasionally VIFs fail to show problems that CIs and variance proportions uncover. The data set *passers.sav* contains information on a sample of NFL quarterbacks who played for professional football teams in the late 1970s and 1980s. We wish to predict the quarterback rating from several passing statistics. The National Football League’s (NFL) quarterback ratings are based on a formula that takes into account several statistics, some of which we use in the following linear regression model.

The regression output indicates that all of the independent variables are
significantly associated with quarterback rating. Assuming they are measured using different metrics, but are all normally distributed (you may wish to check this assumption), the beta weights suggest that no one variable stands out as particularly influential. Rather, all four independent variables appear to be important for predicting quarterback ratings. However, are there any collinearity or multicollinearity problems? The VIFs indicate that there are not; all four are well below the standard rule-of-thumb of 9 or 10. However, the CIs and variance proportions tell a different story. Dimension 5 has a very large CI, 141.188. Scanning its row shows three variance proportions that exceed 0.50: The constant or intercept, completion percentage, and yards per completion. We have not discussed the possibility of the intercept being involved in a multicollinearity problem, yet here it is. Although we may not be concerned with the intercept, studies suggest that it is part of a multicollinearity problem usually when there is relatively little variability in the dependent variable.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>-26.962</td>
<td>1.005</td>
<td>-26.827</td>
<td>.000</td>
</tr>
<tr>
<td>Completion percentage</td>
<td>1.339</td>
<td>.013</td>
<td>.727</td>
<td>106.658</td>
</tr>
<tr>
<td>Yards per completion</td>
<td>2.398</td>
<td>.037</td>
<td>.560</td>
<td>63.962</td>
</tr>
<tr>
<td>TDs per 100 attempts</td>
<td>3.366</td>
<td>.040</td>
<td>.682</td>
<td>85.167</td>
</tr>
<tr>
<td>Interceptions per 100 attempts</td>
<td>-4.236</td>
<td>.042</td>
<td>-.743</td>
<td>-101.509</td>
</tr>
</tbody>
</table>

a Dependent Variable: Quarterback rating

(SPSS printout continued on next page)

The Collinearity Diagnostics show that we should be concerned with completion percentage and yards per completion. There is something about their association (and it may also include the intercept) that is problematic. We do not know at present what this problem entails, but it may be worth exploring further if
we wish to have confidence in our regression model that predicts quarterback rating.

<table>
<thead>
<tr>
<th>Model</th>
<th>Completion percentage</th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>.544</td>
<td>1.839</td>
</tr>
<tr>
<td></td>
<td>Yards per completion</td>
<td>.330</td>
<td>3.035</td>
</tr>
<tr>
<td></td>
<td>TDs per 100 attempts</td>
<td>.394</td>
<td>2.541</td>
</tr>
<tr>
<td></td>
<td>Interceptions per 100 attempts</td>
<td>.471</td>
<td>2.123</td>
</tr>
</tbody>
</table>

Collinearity Diagnostics

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
<th>Variance Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Constant</td>
</tr>
<tr>
<td>1</td>
<td>4.955</td>
<td>1.000</td>
<td>.00</td>
</tr>
<tr>
<td>2</td>
<td>.031</td>
<td>12.644</td>
<td>.00</td>
</tr>
<tr>
<td>3</td>
<td>.011</td>
<td>21.191</td>
<td>.00</td>
</tr>
<tr>
<td>4</td>
<td>.003</td>
<td>39.932</td>
<td>.00</td>
</tr>
<tr>
<td>5</td>
<td>.000</td>
<td>141.188</td>
<td>1.00</td>
</tr>
</tbody>
</table>

What To Do About Collinearity and Multicollinearity

Now that we have learned how to detect these problems, we need to know about some possible solutions. The first solution was mentioned earlier. Collinearity and multicollinearity have less influence on linear regression models as the sample size
gets larger. Therefore, if a model is plagued by collinearity problems, then, if possible, one should add more observations. Unfortunately, this is rarely feasible outside of a laboratory. Survey data are typically of a fixed sample size because of cost constraints, so adding more observations is not practical.

Another solution was discussed in Chapter 8. Suppose that some subset of independent variables is involved in a multicollinearity problem. Look over these variables: Are they measuring a similar phenomenon? Would it be practical to combine them in some way? Oftentimes, highly collinear variables are indirectly measuring some latent construct, so using them to create a latent variable is a common solution to multicollinearity. Rather than including the collinear variables in the regression model, we may place the latent variable in the model. Then we may assess its association with the dependent variable. For instance, perhaps completion percentage and yards per completion are indicators of an underlying characteristic, such as safe or highly adept passers. They may not score a lot of TDs or have many interceptions, but they are good at getting the football to their receivers.

The third solution is to use a technique known as \textit{ridge regression}. Although it is rarely recommended as a general solution, ridge regression may offer some help when nothing else seems to work or in particular situations that we won’t go into here. The general idea underlying ridge regression is to create (artificially) more variability among the independent variables. This is usually done by considering, rather than the raw data, the variance-covariance matrix of the variables. The diagonals of this matrix are the variances, which are too small in the case of collinearity or multicollinearity to yield unbiased estimates of standard errors. Hence, in ridge regression the analyst chooses a constant to add to the variances, thus increasing variability and decreasing the effects of multicollinearity. The choice of this constant may be arbitrary or based on a specific formula. This may sound very promising, but you should realize that ridge regression results in biased slopes. But the bias may be less than the OLS estimates that are made unstable by multicollinearity. A key problem with ridge regression, though, is that it seems to change slopes that are not significant in an OLS model more than slopes that are significant. Therefore, the analyst must have solid conceptual hypotheses to determine which slopes are important and which are not. (See Norman R. Draper

The fourth solution, which is not recommended very often, is to use a variable selection procedure, such as stepwise or backward elimination, to choose the “best” set of predictive variables (see Chapter 7). These methods will typically drop one or more of the variables involved in the collinearity or multicollinearity problem. However, they also take the important conceptual work out of the hands of the analyst and present other problems (see Fox (1997), op. cit., Chapter 13).

The last solution is recommended the least often, but is probably employed most often by analysts who use regression techniques. This is to simply omit one or more of the variables involved in the collinearity or multicollinearity problem. This is recommended only as a last resort or if one of the collinear variables measures virtually the same phenomenon as other collinear variables (such as was the case with the two suicide variables we saw earlier in the chapter). It is rare, however, for two variables to measure the same thing. It is much more common to find that collinearity is caused by recoding problems in the data cleaning stage. This may occur when one variable is created from another.

It should be clear to you that collinearity and multicollinearity present some potentially serious problems for the linear regression model. Biased standard errors are a nuisance, but, when coupled with unstable slope coefficients, the results of the model are not trustworthy. What should you do about this problem? It is best to consider carefully the source of collinearity or multicollinearity and then use the best solution available. Unfortunately, there are no easy answers.
10 Nonlinear Associations and Interaction Terms

In Chapter 3 we discussed an important assumption of the linear regression model: The mean value of $y$ for each specific combination of the $x$’s is a linear function of the $x$’s. We also saw briefly at the beginning of Chapter 7 that one type of specification error involves this assumption. When the association between $y$ and one or more of the $x$’s is not represented well by a straight line or flat surface, then we have one form of specification error. A linear regression analysis is going to attempt, using least squares techniques, to fit a straight line to a set of data. But, as shown in the graph in Chapter 7 (which is reproduced below) it is clear that when the association between $x$ and $y$ is not linear – rather, when it is curved or curvilinear – the OLS estimates will not yield the best results.

Curvilinear and other associations among variables that are not represented well by a straight line are known generally as nonlinear associations. Although making comparisons across the social and behavioral sciences would be very difficult, it is likely that nonlinear associations are the norm, with linear associations the exception. Yet linear associations are much simpler to conceptualize and to model using statistical procedures such as correlations and OLS regression analysis. Although there is a growing literature on nonlinear regression techniques, and modern-day statistical software provides powerful tools for estimating nonlinear
associations, the way we think about associations among variables still tends to be linear. This may have something to do with general limitations of the human mind, but it is likely that we simply need more experience with nonlinear thinking and modeling.

There are actually two approaches to nonlinear analysis that are used in regression modeling. We’ll begin with the more complex approach, although we’ll end up learning about the simpler approach. The first type of nonlinear regression analysis proposes that the estimated coefficients are nonlinear; in other words, they do not imply a straight line pattern. An example of a model that is nonlinear in the estimated parameters is

\[ y_i = \alpha + \beta_1 x_{i1} + \beta_2^2 x_{i2} + \log_e (\beta_3 x_{i3}) + \beta_4 x_{i4} + \varepsilon_i \]

In this equation, the betas are not simple, linear terms; rather, they include \textit{higher-order terms}, with nonlinearities such as quadratics and logarithms (\(\log_e\) implies the natural or Naperian logarithm). There are nonlinear regression routines that are designed specifically for situations when the estimated parameters are hypothesized to be nonlinear. However, such hypotheses can be difficult to formulate, so we will not cover this type of nonlinear regression approach in this presentation. Another promising avenue in regression analysis, but one that we also will not discuss, is known as \textit{nonparametric regression}. An example of this model is known as generalized additive models (GAM), which allow the analyst to fit various types of associations – linear and nonlinear – among the dependent and independent variables. See John Fox (2000), \textit{Multiple and Generalized Nonparametric Regression}, Thousand Oaks, CA: Sage, for a good overview of these and other nonparametric models.

In this chapter, we’ll address a simpler approach that adapts the linear regression model so that it may incorporate nonlinear associations among the independent and dependent variables. To begin to understand how we may do this, consider that when we estimate a linear regression model and interpret the slopes, we assume that the magnitude of change in \(y\) that is associated with each unit change in \(x\) is constant. In other words, it is the same regardless of where in the \(y\) distribution the change is presumed to take place. For example, suppose we find,
like in Chapter 2, that the slope of per capita income on the number of robberies per 100,000 is 14.46. By now, we should be able to interpret this coefficient effortlessly: Each $1,000 increase in per capita income is associated with a 14.46 increase in the number of robberies per 100,000. Keeping this in mind, consider the following diagram:

![Diagram illustrating the relationship between per capita income and robberies per 100,000.]

It should be clear that the difference in the number of robberies associated with the difference in per capita income is expected to be the same regardless of where along a hypothetical continuum of robberies per 100,000 people we find ourselves. Of course, as mentioned in Chapter 2, we should only interpret these associations within the bounds of our data. Suppose, though, that when states have low per capita incomes, say, less than $20,000, each unit increase is per capita income is associated with an increase of 20 robberies per 100,000; whereas in higher income states (e.g., per capita income > $30,000) each unit increase in per capita income is associated with an increase of only 10 robberies per 100,000. Perhaps higher income states have more resources to hire police or provide job opportunities for youths at risk of becoming delinquent or growing into criminals. Whatever the hypothesized mechanism, these sorts of possibilities are legion and recommend against simple hypotheses that suggest only linear associations.
Let’s look at an example of a nonlinear association. In the gss96.sav data set there are variables that measure personal income \((\text{pincome})\) and the respondent’s age in years \((\text{age})\). Before analyzing these data, think for a moment about people’s incomes in the United States. How should the association between age and income behave? (To simplify our thinking, let’s deal with adults only.) First, it seems clear that younger adults who are still in school or just entering the workforce have less income than middle-aged or older adults who have been in the workforce longer (we’ll also ignore for now the influences of education, unemployment, job training, and other important variables). But what happens as people get older and reach their sixties? Many begin to retire and their personal income tends to decrease. Thus, we have already departed from hypothesizing a linear association between age and personal income. Rather, there is likely a curvilinear association, with income rising with age until it reaches a maximum point, after which it begins to decrease as people retire. But there are also important demographic issues to consider when dealing with age. One such issue involves death. Wealthier people, on average, tend to live longer than poorer people partly because they have access to better health care. As we begin to look at the age-income association at older ages, say 75 or 80, we may begin to see the effects of certain people dying earlier than others. If, on average, those with less income tend to die earlier, then income should appear higher among older adults. There are many complex demographic phenomena at play here that we don’t have time to address, but we should consider that the age-income association is not linear and may have more than one “bend.”

The following three graphs represent the three associations between age and income that we’ve discussed thus far. The first graph shows a linear association, with income increasing at a steady pace with increasing age. The second graph shows what is known as a quadratic association, with one “bend” in the association. The third graph shows a cubic association, with two bends – and three distinct pieces – to represent the last part of our discussion about age and income. Once we begin dealing with nonlinear associations, they can quickly blossom into some very complex shapes with numerous bends and twists. Therefore, to manage the modeling it is important to always consider conceptually the associations in your set of variables. As we saw in earlier chapters, there are routinized ways to estimate
regression models (e.g., stepwise selection, reliance on nested $F$-tests), but these should be avoided if they are not guided by theoretical considerations.

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
<th>Income</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>50</td>
<td>80</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>50</td>
<td>80</td>
<td>20</td>
<td>50</td>
</tr>
</tbody>
</table>

When thinking about associations among any set of variables, it is wise to consider whether there are nonlinear associations. You may be surprised at how often they occur. For instance, the association between age and various outcomes is often curvilinear. The impact of independent variables such as education on outcomes often have what are known as *floor or ceiling effects*: There is a negative or positive association up to a certain point; thereafter there is a flat association.

![Graph showing a hypothetical ceiling effect concerning the association between peer delinquency and one’s own delinquency.](graph)

This graph shows an example of a hypothetical ceiling effect concerning the association between peer delinquency and one’s own delinquency. As the number of friends involved in delinquency increases, one’s own delinquency also tends to
LINEAR REGRESSION ANALYSIS

increase up to a certain point. High frequency delinquents, though, are often psychologically different from others (e.g., more impulsive, less empathetic, more callous), so having more delinquent friends may not add much to their own delinquency.

If you suspect that there are nonlinear associations among your variables, there are some simple tools for checking them. Perhaps the most useful is the ordinary scatterplot. Using SPSS’s graphing options, it is easy to construct a scatterplot between two variables. For instance, we’ll use the gss96.sav data set and the following drop-down menu – Graphs-Scatterplot-Simple – to construct a two-variable scatterplot with personal income on the y-axis and age on the x-axis. The result is very busy since there are so many data points. It is impossible to visualize whether a straight line or a curved line fits this association best. Fortunately, SPSS provides additional options to fit a variety of lines to the scatterplot. By double-clicking the scatterplot and highlighting the data points (click on them in the new window that opens up until they are highlighted with the same color) you will find a set of icons that appear at the top of the window. Place your mouse pointer over the icon that looks like a scatterplot with a straight diagonal line. It should read “Add fit line.” Click this icon once and the Properties window opens up. You should now see the five types of lines that are available: The mean of y, a linear fit line, a quadratic fit line, a cubic fit line, and a loess fit line. The mean of y fits a horizontal line. A linear line fits the straight OLS regression line to the scatterplot; the quadratic fits a line with up to one bend; the cubic fits a line with up to two bends; and the loess fits a line with a varying number of bends. Loess stands for locally weighted regression. This means that it parcels the data into smaller (local) sections and fits regression lines to each section. It is useful when you’re not sure what type of nonlinearity might occur among your variables.

Try alternating between linear, quadratic, and cubic to determine the estimated association between age and personal income. You should find that there are the types of bends in the association that we discussed earlier. In particular, there is a quadratic association up until about age 80, and then the fit line flattens out. Hence, there is evidence that the association between age and personal income follows a cubic pattern. If it followed a linear pattern, the linear, quadratic, cubic, and loess would yield the same straight line.
Of course, you may not always know if your model includes these types of associations. Therefore, let's now see a way to test for nonlinearities in a regression model. Then we'll learn what to do when faced with a nonlinear association. We'll stick with the same simple example, but specify personal income as the dependent variable and age as the independent variable. However, after setting up the linear regression model but before running it, click on the Save button and check the box that is labeled Residuals – Studentized. Then run the model. You'll find there is a positive association between age and personal income. But this does not examine the nonlinear association between these two variables, which we now know exists. By constructing a scatterplot with the studentized residuals (which scale the residuals using a \( t \)-distribution) on the \( y \)-axis and age on the \( x \)-axis, however, we can learn something. This type of scatterplot is known as a partial residual plot. Many analysts recommend that we view a partial residual plot for each independent variable used in a regression model. Keep in mind that, in our example, the residuals gauge all the influences on personal income except for its linear association with age. Based on an assumption of the regression model, there should be no association between age and these residuals. But is there? Plot a fit line in the partial residual plot you've just created. Try out whichever type of fit line you'd like. But be advised that many experts recommend using the loess line because it will show many possible associations. If you choose any of the nonlinear fit lines, you'll notice the curvilinear association between age and the residuals. This is indicative of a nonlinear association that we haven’t considered in our regression model, but, as we suspect, it is the nonlinear aspects of age and personal income that are showing up.

Now that we know about the association between age and personal income, what should we do about it? There are a couple of approaches that are particularly helpful. The first is available in SPSS, but is useful only if we are limiting the model to one independent variable. Under the Analyze-Regression drop-down menu there is an option labeled Curve Estimation. Choose this and then place personal income in the box labeled Dependent(s) and age in the box labeled Independent. You'll notice at the bottom of the window a set of boxes under the heading Model. By checking these boxes you can fit a host of nonlinear associations. But, for now, let's stick with linear, quadratic, and cubic (check these boxes). After clicking OK you'll
LINEAR REGRESSION ANALYSIS

find output consisting of three regression models. In the interests of space, we’ll look only at the last model. (Note: The output that appears below is edited to make it easier to read and to keep it consistent with SPSS’s standard linear regression output presented in other chapters.)

The regression output shows three independent variables: age, age**2, and age**3. The latter two variables are age-squared (age × age; or age raised to the power of 2) and age-cubed (age × age × age; or age raised to the power of 3), and correspond to the quadratic and cubic versions of age. All three slopes are significant. Moreover, the signs of the slopes tell us something important (although we already knew this) about the nonlinear association between age and personal income. The signs are positive, negative, and positive; corresponding to an initial positive association, subsequent negative association, and finally a slight positive association. This is the same pattern – as it should be – that we saw in the scatterplot. It is actually quite simple to determine at which ages the association turns negative and then positive. We’ll see one way to do this later.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>-7.487</td>
<td>1.489</td>
<td>-5.027</td>
<td>.000</td>
</tr>
<tr>
<td>age</td>
<td>1.0242</td>
<td>.107</td>
<td>4.332</td>
<td>9.592</td>
</tr>
<tr>
<td>age**2</td>
<td>-0.018</td>
<td>0.002</td>
<td>-6.647</td>
<td>-7.447</td>
</tr>
<tr>
<td>age**3</td>
<td>0.0009</td>
<td>0.0002</td>
<td>2.535</td>
<td>-5.027</td>
</tr>
</tbody>
</table>

A question you may ask is whether this model is an improvement over models with just age or age and age**2 as independent variables. One simple way to determine this is to compare the adjusted R² values from the three models and combine them with nested F-tests to determine statistical significance (see Chapter 5), but we first need to learn a bit more about how to estimate the model.
The curve fitting option in SPSS allows only one independent variable, even though it permits this variable to take on different forms. But we may also compute these higher-order terms ourselves using either the Transform-Compute drop-down menu, or with the following commands in an SPSS syntax file:

```
COMPUTE agesq = age ** 2.
COMPUTE agecube = age ** 3.
EXECUTE.
```

The names for these new variables are arbitrary, but indicate what they represent. We may then estimate a linear regression model with age, agesq, and agecube as the independent variables. The resulting model is identical to the model using the Curve Fitting option. However, it is always a good idea when using independent variables with higher-order terms (e.g., age and age^2) to check for multicollinearity among the variables. Let’s do this and see what the results show. The VIFs for all three independent variables are very high: 454, 1773, and 477. The fourth condition index (CI) is a whopping 218, with all four variance proportions exceeding 0.50. In fact, one of them is 1.00, the maximum value! There is – and this should not surprise you – substantial multicollinearity in this model. Another symptom of multicollinearity is the large standardized regression coefficients; it is rare to find one larger than 1.0, yet all three are much larger than one (in absolute value). Recall, though, that multicollinearity inflates standard errors, making it harder to find significant results. Yet all three age variables have significant slopes, thus supporting the hypothesis of a cubic association between age and personal income.

Nonetheless, most of us don’t care for such extreme multicollinearity in our linear regression models. But is there anything we can do about it? There is a simple statistical trick that allows us to diminish the multicollinearity problem. It is based on the following statistical phenomenon:

\[ \text{Cov}(x_i - \bar{x}, \bar{x}) = 0 \]

You may substitute correlation for covariance and the results are the same. The proof of this statement is provided in many books on probability or statistical
theory (e.g., Sheldon Ross (1994), *A First Course on Probability*, New York: Macmillan). What it means for us is that by centering the independent variable before computing the higher-order terms, and then using these new terms in the linear regression model, multicollinearity is reduced substantially. You may recall that perhaps the most common form of centering a variable is to compute its $z$-scores. In SPSS the simplest way to compute $z$-scores is with the Analyze-Descriptive Statistics-Descriptives drop-down menu: Place the variable in the Variable(s) box, and check the box in the lower left-hand corner that reads “Save standardized values as variables.” Going through these steps with age, we find that a variable $Z_{age}$ has been added to the data set: It consists of the $z$-scores for age.

Next, let’s use $Z_{age}$ to create new quadratic and cubic age variables. In an SPSS syntax file, the following lines produce these variables:

```
COMPUTE zagesq = Zage ** 2.
COMPUTE zagecube = Zage ** 3.
EXECUTE.
```

After creating these variables, estimate a linear regression model with personal income as the dependent variable and $Z_{age}$, $zagesq$, and $zagecube$ as the independent variables. Make sure you ask for collinearity diagnostics. The results shown on the following page should be listed in the output window.

Keeping in mind that the age variables are now measured in different units, the results are quite similar to the earlier results: A positive association that turns negative and then positive again. Notice also that the slopes are highly significant (look at the relatively large $t$-values). More importantly, though, the VIFs and CIs are well below the thresholds indicative of multicollinearity. It is a good idea to remember this statistical trick for dealing with collinearity and multicollinearity problems; it will come in handy later in the chapter and, perhaps, throughout your career in regression modeling.
NONLINEAR ASSOCIATIONS AND INTERACTION TERMS

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>10.858</td>
<td>.087</td>
<td></td>
<td>125.512</td>
</tr>
<tr>
<td>Zage</td>
<td>.515</td>
<td>.096</td>
<td>.175</td>
<td>5.362</td>
</tr>
<tr>
<td>zagesq</td>
<td>-1.034</td>
<td>.071</td>
<td>-.483</td>
<td>-14.542</td>
</tr>
<tr>
<td>zagecube</td>
<td>.180</td>
<td>.033</td>
<td>.244</td>
<td>5.475</td>
</tr>
</tbody>
</table>

A Dependent Variable: personal income in $10,000s

A problem with using these centered independent variables emerges if we wish to determine at what point in the age or personal income distribution the slope turns negative or positive. One valuable approach uses, once again, a scatterplot. This scatterplot employs the unstandardized predicted values (these may be saved by checking the appropriate box in SPSS’s Linear Regression’s Save window) as the y-axis and age as the x-axis. The plot (it should look similar to the scatterplot below) shows that the initial downturn in the association occurs close to age 45 (change the scale marks of the x-axis from 20 to 5 to help you visualize this), with the final upturn around age 80. But consider the following question: Why is the downturn at such a relatively young age? Shouldn’t it occur much closer to the typical retirement age? (Hint: It might have something to do with the nature of cross-sectional data and birth cohorts.)
Nonlinear associations abound in the social and behavioral sciences (as well as in other scientific disciplines, such as chemistry, biology, and physics). We’ve only seen a couple of simple examples, but when we consider variables such as age, education, socioeconomic status, and many other interesting variables, we often find nonlinear shifts in their associations with numerous outcome variables. It is best to first think carefully about potential nonlinear associations among your variables (does the literature suggest any?) and then use visual techniques, such as scatterplots, to examine them. Then, once you have a sense of the likely shape of the associations, consider ways to test for them in a linear regression model. Squaring or cubing independent – or, as we shall see – dependent variables scarcely uncovers the myriad possibilities.

Is the Error Term Normally Distributed?

Another assumption that falls under nonlinearities (although indirectly) involves the dependent variable. As discussed in Chapter 3, we assume that the error term is normally distributed. Recall that the error term (or residuals) measures everything about the variability of the dependent variable once we have taken into account the part of it that may be predicted by the independent variables. We hope that all that is left is random noise, but this is not always the case. Nonetheless, we still assume that the part that is left over is normally distributed; in other words, it follows a normal (Gaussian) distribution. There are two ways to test this assumption, one occurs before estimating the regression model, the other after estimating the regression model, but both use the same type of plot.

A normal quantile-quantile or normal q-q plot is very handy for examining the distribution of the dependent variable. When using this plotting approach to test the assumption of normally distributed errors, some statisticians call it a normal probability plot or a Gaussian rankit-plot. It is rare in the social and behavioral sciences that a dependent variable is continuous; it is rarer still to find one that is normally distributed. Since this book deals with continuous dependent variables (but see Chapter 13), we’ll ignore the issue of non-continuous dependent variables. (For information about regression models for non-continuous dependent variables, see John P. Hoffmann (2004), Generalized Linear Models: An Applied Approach, Boston: Pearson.) However, because the assumption of a normally distributed
error term is so important, yet so rarely satisfied, we need to address it directly. A normal *q-q* plot provides a straightforward way to assess whether a variable is normally distributed. It also offers an indirect and preliminary way to assess whether the residuals are normally distributed. Here’s how it works. First, the statistical program checks the mean and standard deviation of the variable. Second, it generates a simulated variable that follows a normal distribution but has the same mean and standard deviation as the observed variable. Third, it orders the simulated variable and the observed variable from their lowest to highest values. Fourth, it produces a scatterplot with the simulated variable on the *y*-axis and the observed variable on the *x*-axis. If the observed variable is normally distributed, the data points line up along a diagonal line drawn by the program. If it is not, the points deviate from a diagonal line. The direction of deviation provides some guidance as to how the observed variable is distributed. Let’s first look at some *q-q* plots and then we’ll learn how to use one in SPSS with an actual variable.

This plot shows that the observed variable (it is on the *x*-axis) is skewed slightly and, if we were to investigate further, we’d find it has a longer right-tail than a normal distribution. One way to normalize such a distribution (this is simply shorthand for saying we are trying to transform it into a normally distributed variable) is to take its square root.
Here we have a variable that has an even longer right-tail than the previous variable. If we observed a histogram of this variable, we’d find a few observations at an extreme upper end of the distribution. This type of variable needs a greater transformation effort to pull in the extreme values than is available with the square root function. In this situation, most analysts take the natural logarithm (log$_e$) of the variable. However, if you do this make sure you recode any zero values to a positive number (perhaps by adding a one to the variable) since the natural logarithm of zero is undefined. A regression model with a logged dependent variable is so popular in applied statistics that it is referred to as a log-linear model.

The distribution of this variable is virtually the mirror image of the first q-q plot we saw. Its right-tail is shorter than a normal distribution; it has fewer observations
in its upper tail than does a normal distribution. In order to normalize it, we might square its values to stretch them out a bit. Finally, let’s see an even more extreme example of this short-tailed distribution.

![Graph showing normal distribution and expected normal value vs. observed value.](image)

In this situation, one might try exponentiating the variable (using the exponential and raising it to the power of the observed value, $e^{x_i}$) to stretch out its distribution.

Of course, this does not exhaust the ways we might transform variables so they are normally distributed. Moreover, it is important to realize that these transformed variables may be placed in a linear regression model just like any other dependent variable, but the interpretations of the coefficients must change to account for the transformation. For example, if we estimate a model with the square root of income as the dependent variable and age as the independent variable, the slope indicates differences in square root units of income that are associated with each one year difference in age.

Thus far, we’ve discussed what we should do to test whether the dependent variable is normally distributed before estimating a linear regression model. But we should also examine the residuals after estimating the model to make sure they follow a normal distribution. After all, this is the key assumption that needs to be examined. We still use the normal q-q plot, although SPSS calls it a normal probability plot. It is available in the Plots window of the Analyze-Linear Regression drop-down menu. To see one of these in action, let’s estimate a model with the
usdata.sav data set. We’ll once again use robberies per 100,000 people as the dependent variable.

First, create a normal q-q plot using the drop-down menu Graphs-Q-Q. Place the variable robbrate in the Variables box and make sure that Normal appears in the box labeled Test Distribution. Then click OK. The resulting graph should look like one of the graphs above. Given this, what type of transformation should be used to normalize the robberies variable? Before deciding, let’s run a linear regression model with robberies per 100,000 as the dependent variable and the unemployment rate and per capita income in $1,000s as independent variables. In setting up the model, ask SPSS to generate a normal probability plot as part of the output (check the box that appears in the Plots window).

The normal probability plot is based on the standardized values (z-scores) of the residuals. SPSS’s chart output labels it as Normal P-P Plot of Regression Standardized Residuals. Notice that the residuals of robbrate are not too far off the diagonal line. In fact, there is, if anything, a slight trend toward a compressed distribution. Nonetheless, experience suggests that these residuals are in pretty good shape. But can we make them even closer to a normal distribution? Let’s compute the natural logarithm of robberies per 100,000 and place it in the linear regression model. Use the Transform-Compute drop-down menu and look for the LN function to create this transformed variable. Or in a syntax file, use the following commands:

```
COMPUTE ln_robb = LN(robbrate) .
EXECUTE .
```

Note that ln_robb is an arbitrary, yet properly descriptive, name. Once you’ve created this transformed variable, substitute it for the original dependent variable and re-estimate the model. What does the normal probability plot show?

It appears we’ve done more harm than good. There are now a lot of residuals in the middle of the distribution that are well above the diagonal line. Go back and see if the square root of robberies per 100,000 does a better job. What do you find? It looks better than the logarithmic version, although there are still some odd-looking residuals near the middle of the distribution. Is the normal probability plot
from the original model better looking? It appears to be. Therefore, even though the dependent variable is slightly skewed, it is suitable for a linear regression model with these two independent variables.

Let’s now look at a variable that is highly skewed: Gross state product in $100,000s (gsprod). A q-q plot shows the degree of non-normality. Estimate a simple linear regression model with gross state product as the dependent variable and the population density (density) as the independent variable. What does the normal probability plot show? It looks rather odd, with some residuals well below and some well above the diagonal. Try out a few transformations of gross state product to see if you can normalize the residuals. Then compare the regression model before transforming the variable to one after transforming the variable. Do you see any differences that would make you prefer one model over the other? The natural logarithm does a pretty good job of normalizing the residuals, but can you do better? A key question is whether your interpretation of the association between the gross state product and population density changes after transforming the variable. To put another spin on it: Check the distribution of population density. Is it normally distributed? Can you find a transformation that normalizes it? Once you find this transformation, try re-running the linear regression model with normalized versions of both variables. What is your interpretation of the association between gross state produce and population density now?

It is important to keep in mind that you will rarely find a variable in the social and behavioral sciences that has observations – or residuals for that matter – that fall directly on the diagonal line of a normal q-q plot. The goal is to find a transformation, assuming one is needed, which allows the observations or residuals to come close to the diagonal line. Scanning the functions in SPSS’s transform window will give you a good idea about the many transformations that are available. For advice on how to use these various transformations and numerous other ways to think visually about the distribution of variables, an excellent resource is William S. Cleveland (1993), *Visualizing Data*, Summit, NJ: Hobart Press.

To summarize what we’ve learned so far: It is important to check the distributional properties of the variables in the model. Although we’ve emphasized the distribution of the dependent variable, it is also important to examine the
distributions of the independent variables. In fact, some statisticians recommend that analysts should first check to make sure the independent variables are normally distributed (with the exception of dummy variables; see Chapter 6). Moreover, if the association between an independent and dependent variable is nonlinear, it is better in many situations to transform the independent variable in some way so that the association is linear or so that it can accommodate a nonlinear association. We saw one example of this when we analyzed the association between age and personal income. But there are many other nonlinear associations to consider. Perhaps taking the natural log or cube root of the independent variable is needed to linearize (a sophisticated way of saying “make linear”) its association with the dependent variable. An association may be “piecemeal” linear, with different slopes needed in different areas of the independent or dependent variable’s distribution. Or, we may need to take the squared values of one variable and the cubed values of the other to linearize the association. It is easy to become overwhelmed quickly with the many possibilities, so it is best to be cautious, yet thorough, and always keep in mind that theory and past research usually provide good guidance.

After we have checked the distributions of the variables and assessed whether there are nonlinear associations among the independent and dependent variables, it is always a good idea to examine the partial residual plots for each independent variable. Sometimes, nonlinear associations emerge in multiple linear regression models that bivariate scatterplots fail to reveal. This has to do with statistical control in a way that is beyond the scope of this presentation. Nevertheless, these nonlinearities may reveal unexpected, yet important, aspects of the model. If any of these nonlinear associations emerge, you should go back and think about why they occur. Including higher-order terms in a model (such as age\(^2\) and age\(^3\)) may be interesting and important for correct model specification (see Chapter 7 and the age-income example in this chapter), but it is an empty exercise in the absence of thoughtful consideration about the theory and previous research that initially motivated your conceptual model.

Finally, it is essential that linear regression models be accompanied by normal q-q plots of the residuals. A key assumption is that the error term is normally distributed, so testing this assumption is a vital exercise. If you find that the residuals do not follow a normal distribution, there are myriad transformations
available. With additional experience looking at normal $q$-$q$ plots, you will find that selecting an appropriate transformation will become easy.

**Interaction Terms in Linear Regression Models**

This section discusses an important topic in regression analysis: The use of *interaction terms* to represent associations between an independent and dependent variable that vary based on a third variable. These are also known as *non-additive* associations because, to include an interaction term in a linear regression model, we must multiply variables. Therefore, our regression equation is no longer limited to plus signs between variables. For instance, we have thus far restricted regression equations to the following form:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

However, there is no reason we should limit the model so that the predictor variables are completely independent, which is implied by the additive terms in the equation. An interaction term – also known as a *multiplicative term* – involves multiplying independent variables. Here is a sample regression equation which we’ll learn to use shortly:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 (x_{1i} \times x_{2i}) + \epsilon_i$$

The last term in this equation is listed as $x_{1i} \times x_{2i}$, which indicates that the variables are multiplied. We typically say that they “interact” in some fashion. (For a general overview of interaction terms in linear regression models, see Leona S. Aiken and Stephen G. West (1991), *Multiple Regression: Testing and Interpreting Interactions*, Newbury Park, CA: Sage.)

A useful way of thinking about interaction terms is by considering distinct slopes for the different groups represented in the sample. Suppose, for example, that we hypothesize that self-esteem increases between the teenage years and young adulthood (e.g., from ages 15 to 25). For now, we’ll assume there is a linear association between age and self-esteem within this age range. Given the types of regression equations we’ve used thus far, we might estimate the following linear
regression model:

\[
\text{Self-esteem}_i = \alpha + \beta_1(\text{age}_i) + \beta_2(\text{gender}_i) + \varepsilon_i
\]

In this model, we treat age as a continuous variable and gender as a dummy variable. Let's say that gender is coded as 0 = female and 1 = male. Suppose the SPSS output indicates positive slopes for both age and gender, so that males, on average, report higher self-esteem than females. The assumed slopes for age by the two gender groups may be represented using the graph that appears below. If this is confusing, try making up a couple of slopes for age and gender and then estimate a few predicted scores using the linear regression equation. You'll notice that the relative distance between the males and females at each age is the same. This type of model is known as a different intercept-parallel slopes model. The slopes are the same for the two groups; they are just a constant distance apart. This distance is gauged by the intercept.

But perhaps you’ve read the literature on self-esteem and suspect that self-esteem increases more among males than among females from ages 15-25. Thus, we need a way to model different slopes for males and females. One approach we might use is to estimate separate linear regression models for males and females and then compare the slopes from each model. In SPSS, the Data-Split File menu may be used to run separate models for different groups. As we’ll learn, interaction terms are also very handy for modeling different slopes for two or more groups.
The graph below shows how this is represented. But how do we characterize this type of association in a linear regression model? With an interaction term that multiplies age and gender. In the regression equation, this appears as

\[
\text{Self-esteem}_i = \alpha + \beta_1 (\text{age}_i) + \beta_2 (\text{gender}_i) + \beta_3 (\text{age}_i \times \text{gender}_i) + \epsilon_i,
\]

As we saw earlier, the values of one independent variable are multiplied by the values of the other independent variable; in this case age is multiplied by gender. If the coefficient for \( \beta_3 \) is statistically significant, then we may infer that males and females have different age–self-esteem slopes. Hence, this type of model is known as a different intercept–different slopes model. Normally, though, we need to consider the other coefficients as well. At this point we’re assuming all three are positive and statistically significant. This also raises another important issue when using interaction terms: If interaction terms are used in the model, the constituent variables must also appear in the model for correct specification.

Let’s see an example of an interaction term using the gss96.sav data set. We’ll consider the associations among age, personal income, and gender. But, first, consider the following issue: We know that education and income are positively associated in the U.S., with people who have more years of formal education, on average, earning higher incomes than people with fewer years of formal education (an interesting exception involves air traffic controllers, but we’ll ignore this).
Many studies also indicate that women, on average, earn less than men. Let’s see if these two observations are supported in the \textit{gss96.sav} data set. We’ll include age as a control variable in the model since we know it is associated with both education and income (but, to simplify things, we’ll ignore the nonlinear association between age and income).

The results support both propositions: Controlling for the effects of age, males, on average, report more personal income than females; and more years of education is associated with more personal income (note that \textit{gender} is coded as 0 = male and 1 = female in this data set). If we were to plot slopes for education, it would appear similar to the graph shown earlier: Different intercepts but parallel slopes for males and females. The average difference between males and females is 1.328 units of income.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficients(a)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unstandardized</td>
<td>Standardized</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coefficients</td>
<td>Coefficients</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
</tr>
<tr>
<td>(Constant)</td>
<td>5.524</td>
<td>.394</td>
<td>.158</td>
<td>14.011</td>
</tr>
<tr>
<td>age in years</td>
<td>.037</td>
<td>.005</td>
<td>.158</td>
<td>7.376</td>
</tr>
<tr>
<td>highest year</td>
<td>.257</td>
<td>.024</td>
<td>.233</td>
<td>10.869</td>
</tr>
<tr>
<td>of school completed</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td>-1.328</td>
<td>.126</td>
<td>-.226</td>
<td>-10.531</td>
</tr>
</tbody>
</table>

a Dependent Variable: personal income in $10,000s

Now, let’s consider another question that involves these variables: What role does gender play, if any, in the association between education and income? Does more education tend to equalize personal income among males and females? Or is there some type of glass ceiling that results in a greater income disparity among males and females at higher levels of education? Perhaps the occupational choices of highly educated men and women lead to an income disparity. Or there may be no association between gender and education when it comes to personal income; they may only be independently associated with this outcome. We cannot even approach the answers to these questions with the linear regression model we just
NONLINEAR ASSOCIATIONS AND INTERACTION TERMS 165

It provides no evidence about whether income varies by gender and education. Fortunately, an interaction term may be used to gather some of this evidence.

In SPSS, you may use the Transform-Compute drop-down menu to calculate the interaction term. In an SPSS syntax file, this appears as

\[
\text{COMPUTE gen\_educ = gender * educate .}
\]

\[
\text{EXECUTE .}
\]

The name of the new variable, \text{gen\_educ}, is arbitrary. Let’s place this new variable in the linear regression model we just estimated. The output should appear as

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>6.619</td>
<td>.502</td>
<td>13.173</td>
<td>.000</td>
</tr>
<tr>
<td>age in years</td>
<td>.038</td>
<td>.005</td>
<td>.159</td>
<td>7.446</td>
</tr>
<tr>
<td>highest year of school completed</td>
<td>.012</td>
<td>.074</td>
<td>.011</td>
<td>.161</td>
</tr>
<tr>
<td>gender</td>
<td>-3.619</td>
<td>.666</td>
<td>-.615</td>
<td>-5.430</td>
</tr>
<tr>
<td>gen_educ</td>
<td>.165</td>
<td>.047</td>
<td>.458</td>
<td>3.500</td>
</tr>
</tbody>
</table>

*Dependent Variable: personal income in $10,000s*

How should we interpret these results? First, comparing them to the results of the previous model, we see that education no longer has a significant slope and the gender slope is larger (in absolute value). Second, the gender-by-education interaction term has a positive and significant coefficient. It takes some experience to be able to look at the three slopes – education, gender, and gender by education – and make sense of the patterns. But here is a hint that will help you interpret them: Choose one of the variables involved in the interaction to focus on and then consider what the interaction term is telling you about this slope. Let’s pick education. Its slope is 0.012 and is not significant. Hence it is flat or vertical (but at
what value of the other variables does this occur?). But the interaction term multiplies either a zero (to represent males) or a one (to represent females) by education. Therefore, for males we assume one slope, but when we consider females the slope is pulled in an upward direction (as indicated by the positive gender-by-education slope). Similarly, females begin with a detriment in income (as shown by the negative slope for gender), but then are pulled up by increasing education (as indicated by the positive gender-by-education slope).

If this is still difficult to understand, use SPSS to compute predicted values for some representative groups and then compare these values to see whether the gap between males and females changes at different levels of education. In order to do this, first re-run the regression model and ask SPSS to save the unstandardized predicted values. Then, use the Analyze-Compare Means-Means drop down menu to examine the means for the different gender and education groups represented in the data set. Place the unstandardized predicted values (these are labeled $PRE_1$ by SPSS) in the box labeled Dependent List and $educate$ in the box labeled Independent List. Then click the Next box under Layer 1 of 1. This allows you to place another layer in the Means command. Place $gender$ in the Independent List box and click OK.

There are now numerous predicted values for males and females. It is best to look for a frequent or easily recognizable category of education to use as comparison groups. We’ll use 12 and 16 since these typically represent high school graduates and college graduates in the United States. But how do we interpret the difference among these predicted values? One possibility is to look at the raw difference among males and females in the two education groups. In the 12-year group, the raw difference is $\{10.228 - 8.657\} = 1.571$. In the 16-year group, the raw difference is $\{11.024 - 9.862\} = 1.162$. Thus, it appears that the gap between males and females on education diminishes at higher levels of education. This approach to comparing groups often works, but it may break down in certain situations. It is better to compare relative differences or percentage differences within education groups to understand the effects implied by the interaction term:
It is now clear that there is a larger percentage difference between males and females with 12 years of formal education than at 16 years of formal education. This supports the idea that higher levels of education tend to diminish income differences among men and women in the United States.

One of the risks of computing and comparing predicted values in this manner is that we assume that the average age is the same in the four groups we’ve compared. This implies that age is independent of the gender-education association. However, if the average age among, say, females with 12 years of education is substantially different from the average age of males with 12 years of education, then these comparisons may break down. Therefore, it is safer to compute predicted values for those at the same age. To do this, we could compute the predicted values using the implied regression equation and a calculator. As an alternative, we could use a spreadsheet program and construct a formula. In either case, we would plug in the values of interest (e.g., age = 40; gender = 0 or 1; education = 12 or 16) and calculate predicted values for each group. We’ll let you try out this approach.

There is a problem you may have anticipated that could affect our confidence in the regression model. Recall that an interaction term is calculated by multiplying one variable by another. This is bound to lead to collinearity between the interaction term and its constituent terms. In fact, if we re-run the linear regression model and ask SPSS for collinearity diagnostics, we find that gender has a VIF of 28 and gen_educ has a VIF of 37.5. The condition indices and variance proportions tell a similar story and also implicate educate in a multicollinearity problem. Earlier in the chapter, when computing higher-order terms for age, we learned about a solution to this problem: Standardize (i.e., take the z-scores of) the constituent variables first before computing the higher-order terms. It makes no sense to standardize gender since it is a dummy variable, so let’s standardize educate and then compute the interaction term. Then place age, gender, the standardized value of education and the interaction term \{standardized value of education \times gender\}.
in the linear regression model designed to predict income.

The results of this new model show a lack of multicollinearity, but the same general patterns that we saw earlier. In fact, we may save the unstandardized predicted values from the new model and go through the Analyze-Compare Means-Means procedure in the same way as we did earlier. The results are identical. We’ll let you figure out why.

**Interaction Terms with Continuous Variables**

It is slightly more difficult when we wish to use interactions among two continuous variables, although the same logic and most of the same procedures apply. First, we need to have a reason to include interaction terms: Why do we think they belong in a regression model? Second, we may multiply the two continuous independent variables and place the interaction term and its constituent terms in the regression model. There is bound to be multicollinearity, so you may wish to calculate $z$-scores for the two variables and use them to compute the interaction term.

Let’s look at an example of an interaction term that uses two continuous variables. We’ll return to the usdata.sav data set and consider some predictors of the number of violent crimes per 100,000 people. We learned in an earlier chapter that the unemployment rate is positively associated with the number of violent crimes per 100,000. However, we suspect that the percent of males in a state is also associated with this dependent variable. As a control variable, we’ll include the number of migrations per 100,000 people. Let’s suppose, furthermore, that our sad little theory also predicts an interaction between the unemployment rate and percent male, such that we think that when there are more males and more unemployment in a state, there is likely more violent crime. We’re unsure whether this prediction is valid, so we’ll use a two-tailed hypothesis test and simply propose that the unemployment rate and percent male interact in some way.

To test this notion, let’s estimate a linear regression model. First, though, since we wish to minimize the risk of multicollinearity, we’ll use standardized versions of the independent variables (unemprat, permale, and mig_rate) and calculate the interaction term between the unemployment rate and percent male using their standardized forms. After calculating the interaction term, we may estimate the
regression model. The results are provided in the table below.

The unemployment rate is, as expected, positively associated with the number of violent crimes per 100,000. Since the unemployment rate is measured in $z$-scores, we may interpret its slope as “Controlling for the effects of the other variables in the model, each one standard deviation increase in the unemployment rate is associated with 96.7 more violent crimes per 100,000.” Somewhat unexpectedly, the percentage of males in a state is negatively associated with number of violent crimes per 100,000. But what does the coefficient of the interaction term show?

One way to think about the interaction term is to consider the slope for percent male: It is negative. However, as the unemployment rate increases this negative slope becomes, for want of a better term, “less negative.” Another way to view it is to consider the slope of the unemployment rate: It is positive. As percent male increases, this positive slope becomes more positive.

<table>
<thead>
<tr>
<th>Coefficients(a)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>(Constant)</td>
</tr>
<tr>
<td>$z$-Migrations per 100,000</td>
</tr>
<tr>
<td>$z$-Unemployment rate</td>
</tr>
<tr>
<td>$z$-Percent male</td>
</tr>
<tr>
<td>$z$-male x $z$-unemployment</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>(Constant)</td>
</tr>
<tr>
<td>$z$-Migrations per 100,000</td>
</tr>
<tr>
<td>$z$-Unemployment rate</td>
</tr>
<tr>
<td>$z$-Percent male</td>
</tr>
<tr>
<td>$z$-male x $z$-unemployment</td>
</tr>
</tbody>
</table>

a Dependent Variable: Violent crimes per 100,000

Although such interpretations provide general guidance for understanding the interaction term, some analysts find it more intuitive to compute predicted scores. SPSS’s Analyze-Compare Means-Means is not very useful, unless you wish to create categorical variables from the continuous variables, so we’ll simply compute the predicted values by hand (and calculator). But we have to decide on the values of the independent variables. Since they are all measured using $z$-scores, convenient
values include –1, 0, and 1; or one standard deviation below the mean (low), at the mean (medium), and one standard deviation above the mean (high). Let’s see what the predicted violent crime numbers at these values of the independent variables indicate (note that we’ll ignore the effects of migrations per 100,000; by not including it in our calculations, we are predicting the number of violent crimes at mean migration). We may utilize the following equation and plug in the values for the unemployment rate and percent male.

\[
\text{Violent crimes} = 529.6 + 96.7(z\text{-Unemployment}) - 132.2(z\text{-Percent male}) + 83.6(z\text{-Male} \times z\text{-Unemployment})
\]

<table>
<thead>
<tr>
<th>Percent male</th>
<th>One SD below mean</th>
<th>At mean</th>
<th>One SD above mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployment rate</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>One SD below mean</td>
<td>648.7</td>
<td>432.9</td>
<td>217.1</td>
</tr>
<tr>
<td>At mean</td>
<td>661.8</td>
<td>529.6</td>
<td>397.4</td>
</tr>
<tr>
<td>One SD above mean</td>
<td>674.9</td>
<td>626.3</td>
<td>577.7</td>
</tr>
</tbody>
</table>

What do these predicted values indicate about the association among percent male, the unemployment rate, and the number of violent crimes per 100,000? Notice, first, that the lowest predicted number occurs when a state has a low unemployment rate and a high percentage of males. The highest predicted number occurs when a state has a high unemployment rate and a low percentage of males. But how can we determine the effects of, say, the percent male on the association between the unemployment rate and the number of violent crimes? We may compare the predicted values down the columns. For instance, at low values of percent male, the predicted values increase by about 4% (648.7 → 674.9), whereas at high values of percent male, the predicted values increase by 166% (217.1 → 577.7). In other words, the slope of the unemployment rate on the number of violent crimes is much steeper when percent male is high.

One way that these types of findings are often represented is with a graph
where we plot two or more slopes based on the results of the interaction term. The graph below shows the two estimated slopes based on the predicted values. It is important to note that even though there is a steeper slope in states with a high percentage of males, the highest predicted values are in those states with a low percentage of males.

As a final note: Remember to check the normal q-q plot of the residuals from this model. Sometimes, interaction terms introduce additional complications in a model, such as non-normally distributed residuals. What does this model show? Are you more or less confident in the results after examining the distribution of the residuals? More importantly, though, what does the literature on violent crime tell us about the model? Do we trust the finding that a lower proportion of males is associated with more violent crime?

Before concluding this chapter, we should mention that regression models are not limited to interaction terms involving only two independent variables, or what are known as two-way interactions. You may also enter three-way interaction terms in a regression model, which implies multiplying three independent variables together (e.g., \( x_1 \times x_2 \times x_3 \)). However, this regression model must include all of the constituent terms as well as all the possible two-way interactions (\( x_1 \times x_2; x_1 \times x_3; x_2 \times x_3 \)). Conceivably, four-way, five-way, or even higher-order interactions are
possible, but quickly become very difficult to work with and lead to problems with interpretation.

A word of warning: There are some statisticians who claim that interaction terms are problematic because they fail to disentangle the actual ordering of the variables. Recall in Chapter 7 that we discussed endogeneity, or whether there are independent variables in the regression model that are predicted by (or, to use generally unrecommended terminology, caused by) other independent variables inside or outside the model. If one of the variables used to compute the interaction term is endogenous to the other – or to other variables – then the usual interpretation of the interaction term becomes questionable. Think about the model estimated earlier in the chapter: Perhaps gender in some way influences education, or, as we say, education is endogenous in the model. What about the violent crime model? Percent male may affect the unemployment rate (e.g., males are more likely than females to participate in the labor force, so they are at higher risk of unemployment), so the unemployment rate could be endogenous in the model. A comprehensive discussion of this problem is beyond the scope of this presentation, but you should always think carefully about the variables in your model when considering interaction terms and the likelihood of endogeneity.

Chapter Summary

This chapter has covered a lot of important material. Nonlinear associations and dependent variables – as well as residuals from linear regression models – that are not normally distributed are quite common in the world of regression analysis. Moreover, it is important to consider that many of the associations between two variables that interest you may depend upon some third variable. Interaction terms provide a valuable way to test for non-additivity in your models. As a review and conclusion to what we’ve learned, here are four suggestions:

1. Always use theory and previous literature to guide your thinking about interaction terms and possible nonlinearities among your variables.
2. Always use a normal q-q plot to check the distribution of the dependent variable before estimating a regression model. Construct bivariate scatterplots of each independent variable by the dependent variable. Plot
NONLINEAR ASSOCIATIONS AND INTERACTION TERMS

nonlinear lines in these scatterplots. Check the partial residual plots for the regression model using studentized residuals. Check the normal $q-q$ plot of the residuals. When faced with nonlinear associations, consider the various transformations that are available. It is likely you will find one that will linearize the associations or normalize the residuals.

3. When entering interaction terms, consider using standardized versions of the independent variables to diminish the risk of multicollinearity. Use predicted values to figure out what the interaction term implies about the associations in the model.

4. It is easy to get carried away by combining quadratic terms, cubic terms, interactions, and transformed variables in a single regression model. Take care to avoid doing too much since models can easily become unmanageable when there are too many of these types of variable. It is better to have a good idea – guided by theory and previous research – of why these associations exist, rather than searching for them using multiple models. If you cannot explain them, then perhaps they should not appear in your model.
Heteroscedasticity and Autocorrelation

An important assumption of the linear regression model is that the variance of the error term is constant for all combinations of x’s. Recall the artificial example discussed in Chapter 2. There is a presumed association between a nation’s percent labor union and public expenditures. But, in order to examine this association in a statistical model we assumed that the nations in the sample had labor union values that represented labor union values for the population. Similarly, when we collect information from a sample of adults in the United States, we assume that each adult represents some number of adults in the population of U.S. residents. In an important way, each adult’s value on some variable of interest represents the mean value of the variable in the subpopulation he or she represents. Let’s take it for granted that these sample members’ values are a good representation of the means for the subpopulations. What about variability? Is it the same across sample members who have different values for some set of independent variables?

To put some meat on this issue, consider the model in Chapter 4 designed to predict first-year college GPAs. In the example, we used the following variables to predict the dependent variable: SAT math scores, SAT verbal scores, and high school math grades. We now know we made several assumptions when estimating this linear regression model (e.g., lack of measurement error in the variables; correctly specified model). Let’s focus on just one independent variable – high school math grades – to understand one of the assumptions addressed in this chapter. We assumed that the variability in the errors in predicting college GPA followed the same pattern for any particular value of high school math grades. This means that we assumed the same variability for these errors whether math grades were mostly C’s or mostly A’s.

The name for this assumption is homoscedasticity. The term homo means same and the term scedastic means scatter. Hence the errors are assumed to be homoscedastic or have the same scatter. The alternative is that the errors do not follow the same pattern; hence, they are heteroscedastic (hetero means different). The figure on the next page shows an example of heteroscedastic errors. Note how the
presumed spread of the errors in predicting \( y \) increases with hypothetical values of \( x \).

The main consequence of heteroscedastic errors is that, although the slopes are unbiased, the standard errors are biased. An alternative way of saying this is that although the slopes are unbiased they are *inefficient*. The degree of bias is rarely known, although it seems to be especially severe in small samples. Yet it can throw off interpretations of statistical significance even in moderately sized samples.

Heteroscedastic errors are a common phenomenon in linear regression models. They seem to be particularly common in cross-sectional data, or data that were collected at one point in time. Here is an example found frequently in textbook illustrations of heteroscedasticity. The SPSS data set, `consume.sav`, has information on how much income the families in the sample earned in a year and how much they spent on consumable goods in a year. A general hypothesis is that families with more income spend more on goods than families with less income. The results of a simple linear regression analysis support this proposition. Each one-thousand dollar increase in family income is associated with about a nine-hundred dollar increase in spending on consumable goods. Given what you know about
standardized regression coefficients (beta-weights) in the simple linear regression model, what is the model’s $R^2$?

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficients(a)</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant) annual family income in $1,000s</td>
<td>.847</td>
<td>.703</td>
<td>1.204</td>
<td>.244</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.899</td>
<td>.025</td>
<td>.993</td>
<td>35.534</td>
</tr>
</tbody>
</table>

a Dependent Variable: consumable goods in $1,000s

One of the graphical tests we should consider after estimating this model is a partial residual plot. In Chapter 10 we learned that this type of plot is useful for checking the appropriateness of the linear model. However, it also provides a useful way to check whether the errors are homoscedastic or heteroscedastic in a simple linear regression model. Ask SPSS to construct a simple scatterplot of the standardized residuals ($y$) by the consumable goods variable ($x$). (Checking the box for all partial plots on the Analyze-Regression-Linear-Plots menu will provide this scatterplot.) The results should look like the graph on the following page.

Notice that the residuals spread out at higher values of family income. This is an exemplary example of what heteroscedasticity looks like, with the spread of the errors fanning out at higher values of the independent variable. However, an opposite pattern also indicates heteroscedasticity: The spread of the residuals narrowing down at higher values of the independent variable. Homoscedastic residuals look like a random scatter, with no recognizable association in the partial residual plot.

At this point, it is important to ask yourself the “why” question: Why does this pattern exist? Is there a reasonable explanation? Often, by answering this question in a thoughtful way, you can figure out a solution to the problem of heteroscedasticity. For instance, it seems reasonable to suggest that there are (at least) two types of high income families: Those that spend most of their money and those that save a lot of their money. Low income people don’t have this
luxury; they often must spend what they have on necessary goods such as food, with little left over to place in a savings or retirement account or spend on additional items. How does this offer a solution to heteroscedasticity, though? If we had a variable in the data set that measured saving patterns or leisure activities, then we could include it in the model and determine if it explained the heteroscedasticity among the residuals. This illustrates that, as with many problems that affect the linear regression model, a good theory goes a long way.

Another common reason for heteroscedasticity is that one or more of the variables requires transformation. This is often not apparent by looking at a scatterplot between the dependent and one of the independent variables. But previous research may provide a clue. As mentioned in an earlier chapter, income measures are often linearized by taking their natural logarithms. And many statisticians point out that taking the natural logarithm of the independent or dependent variable may be all that is needed to eliminate the heteroscedasticity problem. For example, take the natural logarithm of either the independent or dependent variable in the linear regression model we just estimated, re-estimate the model using the transformed variable, and construct a partial residual plot. What does the plot show? There does not appear to be a clear heteroscedastic pattern to
the residuals, although it does look unusual (either like a U-shaped or inverted U-shaped pattern, which may indicate other problems that we don’t have the space to investigate).

**Checking for Heteroscedasticity in Multiple Linear Regression Models**

It is only slightly more cumbersome to check for heteroscedasticity in a multiple linear regression model. If you suspect that a particular independent variable is involved in a heteroscedasticity problem (draw upon theory and your knowledge of previous research!), then construct a partial residual plot to diagnose it. However, when there are multiple independent variables it is often not clear if one or more of them may induce this problem. As an alternative to a partial residual plot, most statisticians recommend plotting some form of the residuals against the predicted values. John Fox (1991; 1997; *op. cit.*), in his books on regression modeling, recommends plotting the studentized residuals (or deleted studentized residuals) \( y \) against the standardized predicted values \( x \). This is because the studentized residuals (which are often, but not always, similar to the standardized residuals) tend to have more constant variance than the standardized residuals. Regardless of the type of residual used, we still look for a fan-shaped pattern in the scatterplot. Of course, we prefer to find a random pattern because this suggests that the errors are homoscedastic.

As an example, let’s consider a linear regression model using the `passers.sav` data set. We’ll use quarterback rating as the dependent variable and the following two independent variables: Completion percentage and interceptions per 100 attempts. As you are setting up this model, click Plots in the Linear Regression window and place ZPRED (z-scores of the predicted values) in the box labeled X and SRESID (studentized residuals) in the box labeled Y. Click OK and then estimate the model. Notice that completion percentage is positively associated with quarterback rating, but interceptions are not. Rather than the coefficients, let’s focus on the scatterplot that SPSS creates.

The scatterplot on the next page approximates the one produced by SPSS. Notice that, with the exception of the data point at about \((-1, 2)\), there is clearly a fanning-out shape to the residuals at higher quantities of the predicted values. This suggests that there is a heteroscedasticity problem among the residuals. The other
line in the graph is the loess line (locally weighted regression) that we learned a little about in Chapter 10. It is occasionally useful for determining heteroscedasticity and other problems with the residuals. If there was no pattern to the residuals, then the loess line should look like the mean line.

Now that we've see the heteroscedastic pattern in the residuals, are there other ways to detect it? One of the problems is that graphical techniques become difficult to use with large data sets (imagine trying to visualize patterns with hundreds of points in a scatterplot), so it would be nice to have some other techniques for diagnosing heteroscedasticity. Fortunately, there are several. We'll discuss two of these.

The first diagnostic approach is known as White's test (after the econometrician H. White). This test is conducted after estimating a multiple linear regression model. It is comprised of the following steps:

1. Save the unstandardized residuals from the regression model.
2. Compute the square of these values (e.g., resid²).
3. Compute the square of each independent variable (e.g., $x_1^2, x_2^2$). Do not compute squared values of dummy variables, though.

4. Compute two-way interactions using all the independent variables (e.g., $x_1 \times x_2; x_1 \times x_3; x_2 \times x_3$; etc.).

5. Estimate a new regression model with the squared values of the residuals as the dependent variable and the computed variables in (3) and (4) as the independent variables (don’t forget to include their constituent terms).

6. There are then two ways to conclude whether or not the errors are heteroscedastic: (a) Rely on the $R^2$ from the model in (5); if it is significantly different from zero then there is heteroscedasticity. (b) Use the following test statistic:

$$nR^2 \sim \chi^2_k$$

Where $n =$ sample size, $R^2$ is from the model, and $k =$ the number of independent variables. Note that this test statistic is distributed as a $\chi^2$ variable, so we need to compare its value to a $\chi^2$ value with $k$ degrees of freedom.

An advantage of White’s test is that not only does it test for heteroscedasticity, but it also provides evidence about potential nonlinearities in the associations.

The second diagnostic approach is known as Glejser’s test (after the statistician who introduced it, H. Glejser). The easiest way to understand how this test works is to consider the appearance of heteroscedastic errors in the scatterplot. Recall that residuals, whether unstandardized, standardized, or studentized, have a mean of zero, with negative and positive values. Suppose we folded over the residuals along the mean, so that the negatives were pulled up to be positive. This is, in effect, what happens when we take the absolute value of the residuals, or $|\hat{\epsilon}|$. Assuming a fan-shaped pattern, what would the association between the predicted values and the absolute values of the residuals look like? It would be positive if the residuals fan out to the right and negative if they fan out to the left. The scatterplot on the next page, based on the plot shown earlier, demonstrates this. Notice that the presumed regression line indicates a positive association. This, in brief, is how Glejser’s test is conducted: Take the absolute values of the residuals and estimate a linear regression model with these new residuals as the dependent variable and the
original independent variables. An advantage of this test is that it allows you to isolate the particular independent variables that are inducing heteroscedasticity.

Let’s continue our example of quarterback ratings. After estimating the model and saving the unstandardized residuals, we’ll test for heteroscedasticity using White’s test and Glejser’s test. The absolute value function appears on the Transform-Compute drop-down menu as ABS(numexp), or you may use the following commands in a syntax file:

```
COMPUTE absresid = ABS(RES_1) .
EXECUTE .
```

`RES_1` is the default name given by SPSS to the unstandardized residuals and `absresid` is an arbitrary name assigned to the new variable.

![Effect of Taking the Absolute Values of Residuals](image)

The results of White’s test (see the next page) indicate there is a heteroscedasticity problem. The R² from the model is 0.642 and the test statistic \( nR^2 = 26 \times 0.642 = 16.69 \ (df = 4) \sim \chi^2 \) is statistically significant (\( p = .002 \)). Moreover, it is clear from these results that the heteroscedasticity problem is linked to the
completion percentage. As this variable increases, the residuals tend to fan out. Its squared value is also associated with the residuals. This suggests that there may be a nonlinear association between completion percentage and quarterback rating that should be explored further. We’ve omitted the interaction term since it was not statistically significant and the results indicated the source of the problem so well.

Glejser’s test shows a similar result (see the table on the next page). The positive association between completion percentage and the absolute value of the residuals suggests that there is increasing variability in the residuals as the completion percentage increases. In fact, a simple scatterplot with the residuals on the \( y \)-axis and completion percentage on the \( x \)-axis shows much the same thing.

**Solutions for Heteroscedasticity**

We’ve already mentioned one possible solution to heteroscedasticity: Figure out why it occurs. What variable or variables might influence it? In the quarterback rating example, perhaps something as simple as including the squared value of completion percentage in the model would solve the problem. Or, there might be another type of nonlinear association between completion percentage and quarterback rating. In fact, there appears to be a slight U-shaped relationship.
between a quarterback’s completion percentage and his rating, with a negative association between 55% and 58%, and a positive association from about 59% and 64%. (See if you can figure out a way to verify this statement.)

Glejser’s Test

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>-38.794</td>
<td>10.384</td>
<td>-3.736</td>
</tr>
<tr>
<td></td>
<td>Completion percentage</td>
<td>.655</td>
<td>.163</td>
<td>.692</td>
</tr>
<tr>
<td></td>
<td>Interceptions per 100 attempts</td>
<td>.790</td>
<td>.505</td>
<td>.270</td>
</tr>
</tbody>
</table>

a Dependent Variable: Absolute value of residuals

As mentioned earlier in the chapter, transforming a variable by taking its natural logarithm also may reduce heteroscedasticity in the model. This is especially useful when using well-known skewed variables such as income or measures of deviant behavior. Taking the natural logarithm helps when the residuals spread out at higher values of an independent variable; if they narrow down, then squaring the variable or exponentiating it may help. Note, though, that these solutions involve the proper specification of the model. Heteroscedasticity is often only a symptom of improper model specification.

However, there are also situations when a variable that you may not be particularly interested in induces heteroscedasticity. This problem emerges quite frequently when we analyze what are known as repeated cross-sectional data. The General Social Survey (GSS), whose data we’ve used in previous chapters, is a well-known example of this type of data set. Although we’ve employed only the 1996 data, the GSS has been collected for more than 30 years. A different sample of adults in the U.S. are surveyed every couple of years; hence, it involves cross-sectional data that are repeated over time.

One of the problems with such data is that the sample size may vary from one survey to the next. Since sample size affects variability (e.g., larger samples generally lead to smaller standard errors), differing sample sizes across years –
assuming one wishes to analyze data from several years – induces heteroscedasticity. Some analysts simply include a variable that gauges year in their models, hoping that this will minimize heteroscedasticity. A preferred approach, however, is to use weighted least squares (WLS) to adjust for differing sample sizes. WLS uses a weight function to adjust the standard errors for heteroscedasticity. The general formulas for WLS are

\[ SSE = \sum \left( \frac{1}{s_i^2} \right) (y_i - \hat{y}_i)^2 \]

\[ \hat{\beta}_1 = \frac{\sum \frac{1}{s_i^2}(x_i - \overline{x})(y_i - \overline{y})}{\sum \frac{1}{s_i^2}(x_i - \overline{x})^2} \]

where \( \overline{y}' = \frac{\sum y_i}{\sum s_i} \) and \( \overline{x}' = \frac{\sum x_i}{\sum s_i} \)

The formula for the slope is for simple linear regression models, but extending it to multiple linear regression models is not difficult. The key for WLS is the weight function, or \( 1/s_i^2 \). In the example of repeated cross-sectional data, the inverse of the sample size often works well to correct standard errors for heteroscedasticity.

However, in most situations – especially those that do not involve repeated cross-sectional data – we must come up with the weight function conceptually, or by thinking about the associations implied by the model. For instance, suppose that our completion percentage – quarterback rating example was not solved by considering nonlinearities. Are there other variables in the data set that might be inducing heteroscedasticity? There aren’t a lot of options, so let’s propose the following hypothesis: When thinking about NFL quarterbacks and the percent of the time they complete their passes, we should also consider what constitutes a safe passer and a risky passer. Safe passers throw short, accurate passes and therefore have high percentages. But there’s also the rare passer who is accurate when throwing the ball farther; he appears to be a risky passer, but perhaps he is not. Hence, heteroscedastic errors might be induced by combining shorter and longer distance passers in the evaluation of completion percentage and quarterback rating. If this argument has any merit (and it probably doesn’t), then a measure of average distance per pass may be just the weight variable we’re looking for. In the
passes.sav data set there is a variable that assesses average yards per completion (avercomp) that may prove useful.

In SPSS WLS is estimated using the Analyze-Regression-Weight Estimation drop-down menu. Notice that it has areas for the dependent variable, independent variables, and the proposed weight variable. Before going any farther, let’s consider how we will know if the WLS estimates are better than the OLS estimates. Recall that the slopes are unbiased but the standard errors are biased in the presence of heteroscedasticity. Therefore, the standard errors should shift in the WLS analysis. But the degree of bias is rarely known. Let’s see the original OLS model again:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>15.229</td>
<td>20.755</td>
<td>.734</td>
<td>.471</td>
</tr>
<tr>
<td>Completion percentage</td>
<td>1.148</td>
<td>.326</td>
<td>.623</td>
<td>3.520</td>
</tr>
<tr>
<td>Interceptions per 100 attempts</td>
<td>.104</td>
<td>1.009</td>
<td>.018</td>
<td>.103</td>
</tr>
</tbody>
</table>

a Dependent Variable: Quarterback rating

Now, let’s see what the WLS regression model with average yards per completion as the weight variable provides. The results are on the following page. It does not appear as though the WLS model is any improvement over the OLS model. You might try other variables in the data set, although it can quickly turn into a search for the unobtainable. In fact, it is not a good idea to simply hunt willy-nilly for weight variables without strong theoretical justification since using the wrong weight can lead to misleading results.
The other solutions to heteroscedastic errors do not require us to find the source of the problem. Rather, they are general solutions. One is named White’s correction (after the same econometrician who developed White’s test) or the Huber-White sandwich estimator (since it was also described by statisticians P. Huber and F. Eicker; I don’t know why Eicker isn’t recognized!), and the other is named the Newey-West approach (after the econometricians who developed it). Both involve a lot of matrix manipulation that is beyond scope of this chapter. Moreover, neither is available in SPSS, at least not directly. However, there are SPSS user-written programs available on the internet and elsewhere that will do these corrections. Rather than listing the syntax of these programs, the following results were obtained through a statistical program titled Stata. It includes both White’s correction (Huber-White sandwich estimator) and the Newey-West approach as part of its regression routines. The Stata table on the next page shows the results with Huber-White estimates of the standard errors. The table has been revised to look similar to an SPSS table. Note that the general conclusion does not change: Completion percentage, but not interceptions per 100 attempts, is significantly associated with quarterback rating. Note, though, that the standard error for the completion percentage coefficient is larger after correction. However, given what we learned earlier, this result does not preclude the possibility of a nonlinear association involving completion percentage.

Some experts suggest that, because heteroscedasticity is such a common problem in linear regression models, we should always use a correction method. If
there is no heteroscedasticity, then the results of the corrected and uncorrected models are the same. But if there is heteroscedasticity, then the results of a standard linear regression analysis can be highly misleading. Substantially more information on this issue is provided in the following article: J. Scott Long and Laurie H. Ervin (2000), “Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model,” *American Statistician* 54:217-224.

Regression Model Using Huber-White Sandwich Estimators (a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>15.229</td>
<td>26.491</td>
<td>.575</td>
</tr>
<tr>
<td></td>
<td>Completion percentage</td>
<td>1.148</td>
<td>.462</td>
<td>.623</td>
</tr>
<tr>
<td></td>
<td>Interceptions per 100</td>
<td>.104</td>
<td>.971</td>
<td>.018</td>
</tr>
<tr>
<td></td>
<td>attempts</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a Dependent Variable: Quarterback rating

**Autocorrelation**

Another problem arises when the error terms follow an autocorrelated pattern. This problem is especially acute in two types of data: Those that are collected over time and those that are collected across spatial units. Data that are collected over time come in several types. First, we’ve already learned about repeated cross-sectional data. Second, data that are collected from the same individuals (these can be people, animals, or other individual units) over time are known as longitudinal or panel data. Third, data that are collected on the same aggregated unit (e.g., cities, counties, states) over time are known as time-series data. Time-series data are often limited to one unit, such as the city of Detroit, the state of Illinois, or the entire United States; data collected on different aggregate units over time are known as cross-sectional time-series data. No matter the type of data collected over time, autocorrelation is virtually a constant problem.

One way to think about this problem is to consider the nature of residuals or errors in prediction. When we collect data over time, there is typically a stronger
association among errors in time periods that are closer together than in those that are farther apart. For example, if we collect information on crime rates in New York City over a 25-year period and try to predict them based on unemployment rates over the same period, our errors are likely to be more similar in 1980 and 1981 than in 1980 and 2000. Hence, the errors are correlated differentially depending on time. Another name for this type of autocorrelation is serial correlation.

A similar problem occurs when collecting data across spatial units. Suppose we collect data on suicide rates from cities across the United States. We then use percent poverty and the amount of air pollution to predict these rates. The errors in prediction are likely to be more strongly related when considering Los Angeles and San Diego, CA, than when considering San Diego and Providence, RI. Los Angeles and San Diego probably share many more characteristics, some of which we do not measure, than do San Diego and Providence. These unmeasured characteristics add to the error in the model. Another name for this type of problem is spatial autocorrelation.

As with heteroscedasticity, the main result of autocorrelation is biased standard errors. The slopes, on average, are still correct, but the standard errors and, ultimately, the $t$-values and $p$-values, are not correct. Hence, autocorrelation makes it much more difficult to determine whether a slope is statistically distinct from zero in the population.

The scatterplot on the following page shows an example of autocorrelation, or, more accurately, serial correlation. Notice the snaking pattern of the residuals around the linear regression line. This is the typical appearance of autocorrelation; it is the consequence of the stronger association among errors closer together in time than among those farther apart in time. Of course, such a pattern is rarely so clear; it is more common, especially with larger sample sizes, to find an unrecognizable pattern to the residuals. Hence, we need additional tools to determine if autocorrelation is a problem in linear regression models.
Let’s see an example of serial correlation. The SPSS data set, *detroit.sav*, includes variables that assess the annual number of homicides in Detroit, MI from 1961 – 1973. It also has additional variables that were collected on an annual basis. This is an example of a classic time-series data set. Looking at homicides over time indicates an increasing trend. But let’s look at a scatterplot of the homicide rate (*y*) by year (*x*) to see if we can find evidence of autocorrelation. After constructing this scatterplot, fit a linear regression line to the data points. It should look similar to the graph on the following page. You may wish to fit a loess line to the scattersplot; it provides a nice picture of autocorrelation.

Suppose, though, that we wish to use one of the variables in the data set to predict the number of homicides. A key issue in criminal justice management is whether the number of police officers affects crime patterns. We’ll thus consider whether the number of police per 100,000 people predicts the number of homicides. What does a linear regression model show (we’ll ignore year for now)? The results on the next page indicate that the number of police per 100,000 is positively associated with the number of homicides that Detroit experienced over these years. But, we also know that autocorrelation is a likely problem. How should we check for this problem, though? As with heteroscedasticity a plot of the residuals by the predicted values is a useful diagnostic tool.
HETEROSCEDASTICITY AND AUTOCORRELATION

Number of homicides per 100,000 people

<table>
<thead>
<tr>
<th>Year of data</th>
<th>1960</th>
<th>1964</th>
<th>1968</th>
<th>1972</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficients (a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unstandardized Coefficients</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td>t</td>
<td>Sig.</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-77.630</td>
<td>8.631</td>
<td>-8.994</td>
<td>.000</td>
</tr>
<tr>
<td>Police per 100,000 people</td>
<td>.337</td>
<td>.028</td>
<td>.964</td>
<td>12.035</td>
</tr>
</tbody>
</table>

As before, ask SPSS to plot the studentized residuals by the predicted values from this regression model. The resulting plot indicates what appear to be heteroscedastic errors (note the fanning shape in the left part of the plot). But is there evidence of autocorrelation? It is difficult to tell based on this plot, but try fitting a cubic or a loess line. The cubic line indicates an autocorrelated structure. Moreover, the loess line shows a huge bend in the lower left-hand portion of the plot, and then some more curvature in the right-hand side of the plot. With additional experience, you will learn that this is also evidence of autocorrelated residuals.

As mentioned earlier, looking for patterns among the residuals and predicted values can be difficult, especially with large data sets. Fortunately, there are also
statistical tests for autocorrelation. The most common of these is known as the *Durbin-Watson test*. It is calculated by considering the squared difference of the residuals adjacent in time relative to the sum of the squared residuals:

\[
d = \frac{\sum_{t}^{n} (\hat{e}_{t} - \hat{e}_{t-1})^2}{\sum_{1}^{n} \hat{e}_{t}^2}
\]

The subscript indicates time. This test is so common that most statistical programs, including SPSS, compute it as part of their linear regression routines. A rule-of-thumb is that the closer the \(d\) value is to two, the less likely that the model is affected by autocorrelation. The theoretical limits of \(d\) are zero and four, with values closer to zero indicating positive autocorrelation (the most common type) and values closer to four indicating negative autocorrelation (a rare situation). To use the Durbin-Watson test, we first compute the \(d\) value and then compare it to the values from a special table of \(d\) values available in many regression textbooks and on the internet. Some statistical programs, such as Stata (but not SPSS) provide values from these tables as part of their output. The values in the table include an upper limit and a lower limit based on the number of coefficients (independent variables + 1) \(\{k + 1\}\) and the sample size \((n)\). The following rules-of-thumb apply to the use of the \(d\) values from a linear regression model:

1. If \(d_{\text{model}} < d_{\text{lower}}\) then positive autocorrelation is present
2. If \(d_{\text{model}} > d_{\text{upper}}\) then there is no positive autocorrelation
3. If \(d_{\text{upper}} < d_{\text{model}} < d_{\text{lower}}\) then the test is inconclusive

Notice that this test does not assess negative autocorrelation, although, as mentioned earlier, if a \(d\) value from a model is close to the upper theoretical limit of four then there is negative autocorrelation that affects the regression model. Returning to the Detroit homicide example, we may ask SPSS for the \(d\) statistic by checking the box labeled Durbin-Watson in the *Analyze-Regression-Linear-Statistics* menu. The results provide, in the Model Summary table, a \(d\) value of 0.66. It is
substantially lower than two, but we should always check a table of Durbin-Watson values to determine if it falls outside the boundaries for \( k + 1 \) = 2 and \( n = 13 \). A table found on the internet indicates the following boundaries for a five percent test: \( \{1.01, 1.34\} \). Hence, the results of the police – homicide linear regression analysis shows evidence of autocorrelation among the residuals.

**Solutions for Autocorrelation**

As with heteroscedasticity, autocorrelated errors are usually a symptom of incorrect model specification. Thus, a transformation of one or more of the variables may eliminate the problem. Unfortunately, it is difficult to find a suitable transformation in the presence of autocorrelation. Since we are concerned with correct specification, perhaps an important variable has been left out of the model. The most obvious possibility is the variable that gauges time. What happens in the Detroit example when we add *year* to the model? The \( d \) value increases to 1.23. The boundaries for a five percent test with \( k + 1 \) = 3 and \( n = 13 \) are \( \{0.86, 1.56\} \). Since 1.23 falls within these boundaries, there is inconclusive evidence about autocorrelation. Nonetheless, looking at a scatterplot of the studentized residuals by the standardized predicted values indicates that there is still a problem with autocorrelation.

The best approach to solving the autocorrelation problem is to use a statistical technique designed for data that are collected over time. The statistical area of time-series analysis is substantial and is beyond the scope of this presentation (a lucid introduction is Charles W. Ostrom (1990), *Time Series Analysis: Regression Techniques, Second Edition*, Newbury Park, CA: Sage). SPSS has the capabilities to estimate time-series models, but it requires more background information than we have at present to utilize.

There are two popular regression-based approaches to diminishing the effects of autocorrelation: Prais-Winsten regression and Cochran-Orcutt regression. Both of these models use the lagged period of the data (e.g., \( t – 1 \)) to come up with better estimates of the coefficients. The SPSS Trends module includes these and other techniques for handling data collected over time. Let's see an example of a Prais-Winsten regression model and compare it to the original OLS regression model.
**LINEAR REGRESSION ANALYSIS**

### OLS Regression Model (a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-77.630</td>
<td>8.631</td>
</tr>
<tr>
<td>Police per 100,000 people</td>
<td>.337</td>
<td>.028</td>
</tr>
</tbody>
</table>

- Dependent Variable: Number of homicides per 100,000 people

\[d = 0.6656\]

### Prais-Winsten Regression Model (a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>-63.373</td>
<td>14.304</td>
</tr>
<tr>
<td>Police per 100,000 people</td>
<td>.290</td>
<td>.045</td>
</tr>
</tbody>
</table>

- Dependent Variable: Number of homicides per 100,000 people

\[d = 1.4688\]

The Prais-Winsten model has all but eliminated the autocorrelation problem. Notice that the \(d\) statistic is now above the upper bound of 1.34, thus indicating that positive autocorrelation is no longer a problem. Moreover, we continue to find a significant association between the number of police and homicides in Detroit. Yet the slope is smaller and the standard error is larger in the new model. A question remains, however: Why is the association positive?

There are also some useful techniques for other types of data collected over time, in particular longitudinal data. A set of techniques that is growing in popularity is known as generalized estimating equations (GEEs). One of the drawbacks of the Cochrane-Orcutt and Prais-Winsten regression approaches is that they are limited to models with a first-order autoregressive process, or AR(1). This means that the immediate previous value has a much stronger direct influence on
the current value (e.g., 1970 $\rightarrow$ 1971) than other preceding values. Yet, it is often the case that other preceding values (e.g., 1968, 1969) also have a strong effect on the current value. GEEs offer substantial flexibility to model various types of correlation structures among the residuals, not just AR(1). There are syntax files available on the internet to conduct GEE analyses in SPSS. Some statistical software, such as Stata, include GEE routines as part of their standard suite of techniques. (However, SPSS’s Advanced module includes a *Linear Mixed Models* program that may allow the estimation of some of these models.)

Let’s look at a couple of examples of GEEs. The data set *delinq.sav* includes information collected over a four-year period from a sample of adolescents. It includes information on family relationships, self-esteem, stressful life events, and several other issues. Before considering a regression model, look at how the data are set up: Each adolescent provides up to four years of observations (if the id number is the same then it is the same adolescent). We can consider that there are two sample sizes: The number of people represented ($n = 651$) and the number of total observations, or $n \times t$ ($651 \times 4 = 2,904$). Moreover, a key assumption of the linear regression model is violated in this type of data set: The observations are not independent because the same people contribute more than one observation.

Let’s set up a linear regression model. We’ll predict self-esteem (*esteem*) among these adolescents using family cohesion (*cohes*; a measure of family closeness) and stressful life events (*stress*). An OLS regression model provides the results shown in the first table of the following page.

Now let’s see what a GEE provides. The first GEE model is set up (in Stata) to assume that the errors follow an AR(1) process. Therefore, it is similar to a Prais-Winsten regression model. Note that the coefficients are quite a bit closer to zero in the GEE model than in the OLS model. Hence, it is likely that the OLS model leads to biased coefficients when the observations are not independent. Autocorrelation does not affect the standard errors much, though.
### Linear Regression Analysis

#### OLS Regression Model (a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>25.240</td>
<td>.422</td>
<td></td>
<td>59.824</td>
</tr>
<tr>
<td>Family cohesion</td>
<td>.158</td>
<td>.007</td>
<td>.397</td>
<td>22.019</td>
</tr>
<tr>
<td>Stressful life events</td>
<td>-.376</td>
<td>.048</td>
<td>-.142</td>
<td>-7.833</td>
</tr>
</tbody>
</table>

*Dependent Variable: Self-esteem score*

#### GEE Model – AR(1) (a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>z</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>26.604</td>
<td>.444</td>
<td></td>
<td>59.919</td>
</tr>
<tr>
<td>Family cohesion</td>
<td>.127</td>
<td>.007</td>
<td>.319</td>
<td>18.143</td>
</tr>
<tr>
<td>Stressful life events</td>
<td>-.268</td>
<td>.046</td>
<td>-.101</td>
<td>5.826</td>
</tr>
</tbody>
</table>

*Dependent Variable: Self-esteem score*

The AR(1) model may not be the best choice. Is it reasonable to assume that the value in previous year affects the current year’s value, but other years don’t have much of an effect? Perhaps not; other years may also be influential. In a GEE model we may estimate what is known as an **unstructured** pattern among the residuals. This allows the model to estimate the correlation structure rather than assuming that the residuals are uncorrelated (OLS) or follow an AR(1) pattern. The GEE model with an unstructured pattern provides the results shown in the table on the next page.

The first thing to notice is that there are few differences between the AR(1) model and the unstructured model. Therefore, the assumption of an AR(1) pattern among the residuals is reasonable. But keep in mind that if we do not take into consideration the longitudinal nature of the data, we overestimate the association between family cohesion and self-esteem, and between stressful life events and self-esteem.
GEE model – unstructured (a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>z</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>(Constant)</td>
<td>26.691</td>
<td>.441</td>
<td>60.524</td>
<td>.000</td>
</tr>
<tr>
<td>1</td>
<td>.125</td>
<td>.007</td>
<td>.313</td>
<td>17.857</td>
</tr>
<tr>
<td>Family cohesion</td>
<td>-.276</td>
<td>.046</td>
<td>-.104</td>
<td>6.000</td>
</tr>
</tbody>
</table>

a Dependent Variable: Self-esteem score

If available, GEE models provide a good alternative to OLS linear regression models. If you are analyzing longitudinal data – which offer some important advantages over cross-sectional data – GEE models are the preferred approach. Other models that you may come across, such as fixed-effects models and random-effects models, are actually specific cases of GEE models. A review of these various models is beyond the scope of this chapter, but an excellent start is James W. Hardin and Joseph M. Hilbe (2003), *Generalized Estimating Equations*, Boca Raton, FL: Chapman & Hall/CRC.

**A Brief Overview of Spatial Autocorrelation**

Before concluding our discussion of autocorrelation, it is important to recall that another type is known as spatial autocorrelation. This problem occurs when the units in the data are spatial areas, such as counties, states, or nations. It should be evident that errors in prediction are likely to be more similar in adjacent or nearby units than in more distant units. Therefore, whenever spatial data are analyzed it is a good idea to assess the likelihood of spatial autocorrelation because, as with serial correlation, the standard errors are affected. Moreover, as with longitudinal data, the assumption of independence is violated since adjacent areas usually share characteristics much more than distant areas. (Note: The information in this section is derived mainly from Peter A. Rogerson (2001), *Statistical Methods for Geography*, Thousand Oaks, CA: Sage.)

The Durbin-Watson statistic ($d$) is not useful for assessing spatial autocorrelation. However, a standard test is known as Moran’s $I$ (named after its
originator, the Australian statistician Patrick A.P. Moran). Assuming we transform the variable into \( z \)-scores (this simplifies the formula), Moran’s \( I \) is calculated by

\[
I = \frac{n \sum_{i} \sum_{j} w_{ij} z_i z_j}{(n-1) \sum_{i} \sum_{j} w_{ij}}
\]

In this equation, there are \( n \) spatial units and \( w_{ij} \) is a measure of how close together the specific units, indexed by \( i \) and \( j \), are to one another. If two spatial units that are close together exhibit similar scores on the variable (\( z \)), there is a positive contribution to Moran’s \( I \). Hence, if the nearest units tend to have scores that correlate more than units that are far apart, Moran’s \( I \) is larger. In fact, its theoretical values are –1 and 1, much like Pearson’s correlation coefficient, with higher values indicating positive spatial autocorrelation. Negative spatial autocorrelation also may occur, with Moran’s \( I \) closer to –1, although it is rare. Notice that this measure is concerned conceptually with lack of independence across spatial units; hence, we may recognize autocorrelation as a problem of the dependence of observations.

An important issue in calculating Moran’s \( I \) is to figure out the \( w_{ij} \)’s because the \( I \) value is highly dependent on them. The simplest approach is to create a dummy variable with a value of one indicating that the spatial units are adjacent (or share a border) and a zero indicating that they are not. This is known as binary connectivity. Another way is to compute distance measures, although we then have to decide distances between which particular points. A commonly used distance measure is from the center of one region to the center of another (e.g., the city center of Los Angeles to the city center of San Diego). If we use distance, \( w_{ij} \) in Moran’s \( I \) is the inverse of the distance so that units closer together receive a larger weight.

As an example of how to compute Moran’s \( I \), consider the following map. It consists of data from five counties in the state of Hypothetical. The numbers represent a measure of the number of crimes committed in each county in the last year adjusted for their population sizes. We wish to determine the degree of spatial autocorrelation before going any farther in our analysis.
To compute Moran’s $I$, we first need to decide on a system of weights for the distances. To ease the computations, let’s use the binary connectivity approach where a one is assigned to adjacent counties (e.g., A and C) and a zero is assigned to non-adjacent counties (e.g., A and B). A simple way to see these weights is with a matrix where the entries are the $w_{ij}$’s:

$$W = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

This matrix is symmetric, with the same pattern above the diagonal as below the diagonal. For example, there is a 1 listed in row 1, column 3; and a 1 listed in row 3, column 1. This is an indicator that counties A and C share a border. The overall mean of the crime variable is 22.4, with a standard deviation of 6.43. To compute Moran’s $I$, we may convert specific crime values into $z$-scores to save us some steps. The $z$-scores (calculations omitted), listed in order from counties A to
E, are \{1.494, -0.685, 0.56, -0.84, and -0.529\}. We then add the products of each pair of \(z\)-scores, and multiply this sum by the sample size (5) to obtain the numerator (note that the non-adjacent pairs, since they have zero weights, are omitted from the equation):

\[
5 \times [AC + BD + CA + CD + CE + DB + DC + DE + EC + ED] =
\]

\[
5 \times [(1.494 \times 0.56) + (-0.685 \times -0.84) + (0.56 \times 1.494) + (0.56 \times -0.84) + (0.56 \times -0.529) + (-0.84 \times -0.685) + (-0.84 \times 0.56) + (-0.84 \times -0.529) + (-0.529 \times 0.56) + (-0.529 \times -0.84)] = 10.9
\]

The sum of the weights is 10 (count the 1s in the \(W\) matrix), so the denominator in the equation is \(4 \times 10 = 40\). The Moran’s \(I\) value is therefore \(10.9/40 = 0.273\). This suggests that there is a modest amount of spatial autocorrelation among these crime statistics.

The solutions for spatial autocorrelation are similar in logic to the solutions for serial correlation. First, we may add a variable or set of variables to the model that explains the autocorrelation. It is often difficult to find these types of variables, though. Second, we may use a regression model – known as \textit{geographically weighted regression} – that weights the analysis by a distance measure. Large weights apply to units that are close together; small weights apply to those that are farther apart. The regression coefficients are estimated in an iterative fashion after finding the optimal weight. Third, there are spatial regression models designed specifically to address spatial data and autocorrelation. The research area of \textit{Geographical Information Systems} (GIS) includes a host of spatial regression approaches. Rogerson, \textit{op. cit.}, provides a relatively painless overview of spatial statistics and regression models.

\textbf{Chapter Summary}

We now know about the risks of heteroscedasticity and autocorrelation. Both of these problems are often symptoms of specification error in linear regression models, either because we use the wrong functional form or because we’ve left
HETEROSCEDASTICITY AND AUTOCORRELATION

important variables out of the model. The main consequence of heteroscedasticity and autocorrelation is biased standard errors. Since we normally wish to obtain results that allow inferences to the population, obtaining unbiased standard errors is important. However, since specification error often accompanies heteroscedasticity or autocorrelation, the slopes from linear regression models may also be incorrect. Therefore, it is always essential that you think about the nature of your data and variables to evaluate the likelihood of specification error, lack of independent observations, heteroscedasticity, and autocorrelation.

There are several diagnostic tools and many solutions available for heteroscedasticity and autocorrelation. Here is a summary of several of these that are discussed in this chapter.

1. It is always a good idea to plot the studentized residuals by the standardized predicted values after estimating a linear regression model. A fan-shaped pattern to the residuals, one that either spreads out or narrows down at higher values of the predicted values, is indicative of heteroscedasticity. A cubic or similar snaking pattern to the residuals is indicative of serial correlation.

2. There are also several statistical tests for heteroscedasticity and autocorrelation. White’s test and Glejser’s test offer simple ways to use a regression model to test for heteroscedasticity. The Durbin Watson $d$ statistic is the most common numeric test for serial correlation.

3. Solutions to heteroscedasticity and serial correlation are legion. If you know the variable that is inducing heteroscedasticity or serial correlation, then you should consider a WLS regression analysis to adjust the model. If you don’t know the source, or if you are genuinely worried about heteroscedasticity or serial correlation in your linear regression models, then use a model specifically designed to correct these problems. The Newey-West or White-Huber corrections for standard errors under heteroscedasticity work well. Prais-Winsten regression provides better estimates than OLS regression in the presence of serially correlated residuals.

4. For longitudinal data, a frequently used approach is to employ one of the GEE models. They provide a flexible way to model longitudinal associations
among variables.

5. Data collected over spatial units present the same conceptual problems as data collected over time, although the techniques for diagnosing spatial autocorrelation and for adjusting the models are different. Moran’s $I$ is a widely used diagnostic test for spatial autocorrelation. And there are many regression routines designed specifically for spatial data.
12 Influential Observations:
Leverage Points and Outliers

We’ve seen many examples of scatterplots in previous chapters. These scatterplots show a variety of patterns to the relationships among variables. We’ve now seen linear patterns, nonlinear patterns, heteroscedastic patterns, and autocorrelated patterns. Yet nonlinear patterns, in particular, force us to question the use of a linear model to assess the association between two variables. It is not uncommon for two variables to be associated in a curvilinear or nonlinear fashion (see Chapter 10). However, this does not exhaust the ways that data points diverge from a linear pattern. It is perhaps just as common to find small portions of the variables, perhaps even a single point, that diverge from the straight line. Even these small departures from linearity can have large consequences for the linear regression model.

In this chapter we’ll discuss influential observations, or those data points that affect linear regression equations and results to an inordinate degree. The degree of influence is not always clear, but can be quite substantial. There are two types of influential observations that affect a regression model: Outliers and leverage points. Outliers are observations that pull the linear regression line in one direction or another. These are typically relatively extreme values of the $y$ variable. In other words, they (there can be one or more) are far away from the other $y$ values. For instance, suppose we have as a dependent variable a measure of SAT scores from five high school seniors. We wish to predict these scores based on junior-year grades in math, science, and English (of course, we’d want more that five observations to estimate such a model, but we’ll ignore this for now). Their SAT scores (using the old scoring system) are $\{980, 900, 1,010, 940, 1,600\}$. The last score constitutes an outlier; it is extreme on the $y$ variable.

Leverage points are observations that are discrepant or distant from the other values of the $x$ variable. They may or may not also be outliers. For example, if we wish to use SAT scores to predict college GPA, then the score of 1,600 would
constitute a leverage point. The observation might also be an outlier if, say, the student who scored 1,600 also obtained a 1.5 GPA in college whilst the other students obtained GPAs close to 3.0.

The main effect of leverage points that are not outliers is to affect the standard error of the slopes. Recall that the formula for the standard error of a slope in a multiple linear regression model (OLS) is

\[
\text{se}(\hat{\beta}_i) = \frac{\sqrt{\sum (y_i - \hat{y}_i)^2}}{\sum (x_i - \bar{x})^2 (1 - R_i^2)(n-k-1)}
\]

The key component of this equation that is affected by leverage points involves the \(x_i\) values in the denominator. Suppose that, all else being equal, we place a leverage point in the equation. What effect will it have? Well, it will result in a relatively large squared value in the denominator. Hence, the denominator will increase and the standard error will decrease. Thus, leverage points tend to reduce standard errors of slopes in linear regression models. (What affect do outliers have on standard errors?)

The following three graphs show an outlier, a leverage point, and a combination of the two. They also include estimated linear regression slopes.
Note that the outlier pulls what should be a relatively steep, positive regression line upwards near the low end of the $x$ distribution. Therefore, the slope will be smaller than if the outlier were not part of the set of observations. The leverage point falls on the regression line, so the slope is not influenced but, as mentioned earlier, the standard error of the slope is affected. The third graph shows a common situation: The observation is extreme on both the $x$ and $y$ variables. It has a relatively strong influence on the slope and the standard error. In brief, it has a large effect on the regression model.

Influential observations result from a number of sources. Perhaps the most common is coding error. When entering numeric data, it is easy to hit the wrong entry key or to forget a numeral or a decimal place. Therefore, it is always a good idea to explore your data to determine if there are any coding errors. Oftentimes, though, influential observations occur as a routine part of data collection exercises: Some people do earn a lot more money than others; some adolescents do drink alcohol or smoke marijuana much more often than other youth. If the nature of a variable often leads to extreme values, then a common solution is to simply pull in these values by taking the square root or natural logarithm of the variable (see Chapter 10). Before estimating a model, it is a good idea to use exploratory data analysis techniques, such as $q-q$ plots or box-and-whisker plots, to visualize the distributions and check for leverage points or outliers.

In a multiple linear regression context, there are several diagnostic techniques that are useful for unearthing influential observations. These are covered in detail in John Fox (1991), *Regression Diagnostics: An Introduction*, Newbury Park, CA: Sage; and David A. Belsley, Edwin Kuh, and Roy E. Welsch (2004), *Regression...*
Diagnostics: Identifying Influential Data and Sources of Collinearity, New York: Wiley-Interscience. Here we’ll see a brief overview of some of the more common techniques.

Detecting leverage points requires us to consider the spatial patterning among the sets of \( x \) values (assuming we have multiple independent variables). If one observation is considerably discrepant from the others, it should show up as far from the others in a measure of spatial distribution. Recall that we may represent multiple linear regression models using matrix notation (see Chapter 3). Suppose we depict the \( x \) variables as a matrix with the entries representing particular observations:

\[
\begin{bmatrix}
  x_{11} & x_{12} & \ldots & x_{1p} \\
  x_{21} & x_{22} & \ldots & x_{2p} \\
  \vdots & \vdots & \ddots & \vdots \\
  x_{n1} & x_{n2} & \ldots & x_{np}
\end{bmatrix}
\]

The hat matrix based on the \( X \) matrix is defined as:

\[
H = X(X'X)^{-1}X' = (h_{ij})
\]

Recall that the matrix \( X' \) is the transpose of \( X \), whereas \( X^{-1} \) is the inverse of \( X \) (see Chapter 3). This is known as the hat matrix because of the predicted values of \( y \) may be computed using the vector of \( y \) values premultiplied by \( H \). The diagonals of the hat matrix \( (h_{ii}) \) lie between zero and one. We may compute the mean of these values to come up with an overall mean for the set of \( x \) values used in the multiple linear regression model. The points off the diagonals represent distance measures from this joint mean.

Many statistical software packages, including SPSS, provide centered leverage values based on the hat matrix. Larger values are more likely to influence the results of the regression model. In SPSS these values range from zero to \( (n - 1)/n \), which gets closer and closer to one as the sample size increases. A standard rule-of-thumb states that any centered leverage value that is equal to or exceeds \( 2(k+1)/n \) should
be scrutinized as an influential observation. For example, in a model with three independent variables and a sample size of 50, leverage points of $2(3+1)/50 = 0.16$ or more should be evaluated further.

The most common method for detecting outliers is through the use of *deleted studentized residuals* (also known as *jackknife residuals* or simply as studentized residuals by some researchers). These are similar to the studentized residuals we used in earlier chapters, but they are estimated using the root mean squared error ($S_e$; see Chapter 4) from the linear regression model after deleting the specific observation. The formula for deleted studentized residuals is

$$t_i = \frac{e_i}{S_{E(-i)}\sqrt{1-h_i}}$$

The $e_i$ value is the unstandardized residual and the $h_i$ value is from the hat matrix described earlier. This is a standardized measure of the residuals, with a mean of zero, so we would expect, if the residuals are normally distributed, that about five percent of them would fall two units away from the mean. To detect outliers, look for deleted studentized residuals that are substantially greater than two or substantially less than negative two.

As suggested earlier, it is not uncommon to find observations that are both leverage points and outliers. Hence, statisticians have developed a number of summary measures of influential observations. The two most prominent of these measures are Cook’s D or Distance (named for its developer, the statistician R.D. Cook) and DFFITS. Both of these measures combine information from leverage values and deleted studentized residuals. Cook’s D is computed as

$$D_i = \frac{t_i^2}{k+1} \times \frac{h_i}{1-h_i}$$

Larger values of Cook’s D indicate more influence on the regression results. There are various rules-of-thumb for Cook’s D. One is to look for values greater than or equal to one. A more common – and less stringent – rule is to consider Cook’s D values greater than $4/(n-k-1)$ as indicative of influential observations. For
example, in a model with three independent variables and a sample size of 75, any Cook’s D value greater than 4/(75 – 3 – 1) = 0.056 should be scrutinized.

The other general diagnostic measure of influential observations, DFFITS, is quite similar to Cook’s D. It is computed as

$$DFFITS_i = t_i \sqrt{\frac{h_i}{1 - h_i}}$$

A general rule-of-thumb is to consider any DFFITS with an absolute value greater than or equal to two as indicative of an influential observation. (Note that DFFITS, like deleted studentized residuals, can be positive or negative.) However, a rule-of-thumb that considers the sample size is $2\sqrt{(k + 1)/n}$; hence, if we have a model with a three independent variables and a sample size of 75, any DFFITS values greater than $2\sqrt{(3 + 1)/75} = |0.46|$ should be evaluated as an influential observation.

Once these measures are estimated, it is a simple matter to compute the cut-off points for their respective rules-of-thumb and then identify influential observations. Another useful technique is to construct a scatterplot with the leverage values on one axis and the deleted studentized residuals on the other axis. Then you may determine if one or more observations are both outliers and high leverage points. A scatterplot of the Cook’s D values (or the absolute value of the DFFITS) by the leverage values or deleted studentized residuals may also be useful for determining the type of influential observations that are affecting the model.

An Example of Diagnostic Tests for Influential Observations

Let’s return to an example we used in earlier chapters: A multiple linear regression model with violent crimes per 100,000 as the dependent variable. We’ll use the following independent variables: The unemployment rate, gross state product, and the population density. As you set up the model, open the Save window and check the boxes next to Studentized deleted, Cook’s, Leverage values, and Standardized DfFit (the DfFit option in SPSS provides a different measure). The results of the model are in the table on the next page.
The table that follows this one in the SPSS output is labeled *Residual Statistics*. It includes the minimum and maximum values of several diagnostic measures, including the deleted studentized residuals (labeled Stud. Deleted Residual), Cook’s D, and the centered leverage values. We are particularly interested in the maximum values of these measures. First, however, it is useful to compute the cut-off values for these measures. Given our model, the Cook’s D rule-of-thumb cut-off is $4/(50 - 3 - 1) = 0.087$ and the leverage value rule-of-thumb cut-off is $2(3+1)/50 = 0.16$. Although SPSS didn’t print the characteristics of the DFFITS, the cut-off point in this model is $2\sqrt{(3+1)/50} = 0.566$.

According to the SPSS output, the maximum deleted studentized residual value is 2.197, the maximum leverage value is 0.467, and the maximum Cook’s D value is 0.461. The deleted studentized residuals are no cause for concern; it appears that most are within two units of the mean of zero. However, there is at least one observation that exceeds the preferred maximum leverage value and at least one that exceeds the preferred maximum Cook’s D value. We should explore these further.

A convenient way in SPSS to consider these values in more detail is by using exploratory analysis. In SPSS a set of exploratory techniques appears under the *Analyze-Descriptive Statistics-Explore* menu. Placing the newly created variables (labeled SDR_1, COO_1, LEV_1, and DFF_1 by SPSS) in the Dependent List and asking for Outliers in the Statistics window result in substantial information
about potential influential observations. It also allows us to identify which cases comprise the maximum values for these diagnostic variables. For example, according to the SPSS output table labeled *Extreme Values* the largest Cook’s D value (0.461) is case number 5; the next largest is case number 48 (Cook’s D = 0.12). According to the cut-off value, both of these cases are influential observations, but there are no others. The largest leverage value (0.467) is also case number 5, as is the largest DFFITS (absolute value) (1.37). According to the rule-of-thumb for the leverage values (0.16), there are four high leverage points (case numbers 5, 30, 39, and 8). Using the rule-of-thumb for DFFITS (0.566) there are three influential observations (case numbers 5, 20, and 48).

Since case number 5 appears in all three diagnostic measures as an influential observation, it is important that we evaluate it further. The first step is to determine which state is represented by case number 5. Returning to the raw data, it is perhaps unsurprising that California, the state with the largest population, is case number 5. The box-and-whisker plot for the leverage values, Cook’s D, and DFFITS shows how far this case is from the others. The next step is to determine why California appears as an influential observation. Given that it is not an outlier, perhaps it is extreme on one or more of the independent variables. Considering this possibility, a good analyst will explore the data further to understand more about the influence that California’s data have on the regression model.

### Solutions for Influential Observations

We’ve already discussed one of the best solutions for influential observations: Make sure the variables are coded correctly. Oftentimes, influential observations are due to coding errors or non-normally distributed variables. Carefully exploring the variables before estimating a regression model through the various techniques discussed in earlier chapters will go a long way toward alleviating problems with influential observations.

Though not recommended in virtually any situation, another solution is to delete the influential observation. For example, if we cannot figure out why California is affecting our regression model of violent crimes, then we might simply remove it from consideration. For example, the following table provides the results of the regression model with California omitted. The results change quite
INFLUENTIAL OBSERVATIONS

dramatically. In particular, the unstandardized coefficient for the gross state product increases by about 15 units or by about 38%. Nonetheless, this shift in the coefficient should not be used as an excuse to remove the influential observation.

As an alternative, we could simply ignore the influential observations. If our sample is large enough, this may be a reasonable alternative (although the solutions discussed later are preferred). For smaller samples, though, influential observations can have a large impact on the results. We’ve seen one example of this already.

Coefficients(a) – California omitted

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>8.477</td>
<td>141.711</td>
</tr>
<tr>
<td>Unemployment Rate</td>
<td>78.462</td>
<td>26.377</td>
</tr>
<tr>
<td>Gross State Product in $100,000s</td>
<td>90.997</td>
<td>26.683</td>
</tr>
<tr>
<td>Population Density</td>
<td>.001</td>
<td>.142</td>
</tr>
</tbody>
</table>

a Dependent Variable: Violent crimes per 100,000

The best solution is to try to understand influential observations. For example, why do California’s data affect the results so dramatically? Let’s explore this a little further by considering the independent variable – gross state product – that seems most affected by including California in the model. Recall that this variable assesses a state’s economic productivity. California is the U.S.’s most populous state, one of the major producers of agricultural products, the setting for many large shipping ports, and home to some of the most valuable real estate in the world. Even if you are unfamiliar with its economy, it should not surprise you to know that its gross state product exceeds the gross national product (a measure of a nation’s economic productivity) of most nations of the world. In other words, California’s economy is huge. Try running exploratory analyses on the gross state product variable. California easily outstrips the other states, with a gross state product in $100,000s of 9.13. The next largest is New York at 5.88. Next, try running a q-q
plot of gross state product. Notice the extreme value; it is California. It is partially causing a highly skewed variable that appears as if it would benefit from taking its square root or natural logarithm (you may recall that we checked its distribution in Chapter 10 and found it to be highly skewed).

After computing the natural logarithm of the gross state product, re-estimate the multiple linear regression model (with this logged variable) and assess the influential observation measures. You should find that the largest Cook’s D is now 0.135 and the largest DFFITS (absolute value) is 0.76, which are much closer to the cut-off points for both measures. (Which state now has the maximum value on the Cook’s D and the DFFITS?)

Other popular solutions to the problem of influential observations involve the use of regression techniques that are not so sensitive to extreme values. For instance, you may recall that the median is known as a robust measure of central tendency because it is less sensitive than the mean to extreme values (see Chapter 1). Since our multiple linear regression model using least-squares is based on means, perhaps an alternative regression technique based on medians would be influenced less by extreme values. This reasoning is the basis for two very similar regression models: Median regression and least-median-squares (LMS) regression. Rather than minimizing the sum of squared errors, these techniques minimize the sum of absolute residuals or the median squared residual. An example of a median regression model estimated with the statistical software package Stata is provided in the table on the next page.

Another alternative is to use the Huber-White sandwich estimators (see Chapter 11) to adjust the standard errors for influential observations. Recall that this estimation technique is useful for heteroscedastic errors. Often, influential observations mimic heteroscedasticity, so adjusting the standard errors with this widely-used approach yields more robust results. The second table on the next page shows the results of an OLS model with Huber-White estimators of the standard errors using Stata.
A related approach is known as *robust regression* (there are actually several types, including median regression). One form of robust regression begins with an OLS model and then calculates weights based on the absolute value of the residuals. It goes through this process repeatedly (iteratively) until the change in the weights drops below a certain point. In essence it down-weights observations that have the most influence on the model. SPSS does not include a robust regression technique as part of its regular suite of modeling routines. Therefore, the following table presents the results of the regression model using Stata’s *rreg* (robust regression) routine.
Robust Regression (a)

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(Constant)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>49.214</td>
<td>148.139</td>
</tr>
<tr>
<td></td>
<td>73.811</td>
<td>28.197</td>
</tr>
<tr>
<td></td>
<td>64.262</td>
<td>21.522</td>
</tr>
<tr>
<td></td>
<td>.023</td>
<td>.151</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

a Dependent Variable: Violent crimes per 100,000

The results of these various models indicate that there are consistent positive associations between the gross state product or the unemployment rate and the number of violent crimes per 100,000 across states in the U.S. Moreover, even though the evidence is clear that there are influential observations, a number of robust techniques suggest that the effects of these influential observations would not force us to change our general conclusions. For example, here is a comparison of the unstandardized coefficients for the unemployment rate and the gross state product across the various models:

Unstandardized coefficients by type of regression model

<table>
<thead>
<tr>
<th>Variable</th>
<th>OLS</th>
<th>OLS - California omitted</th>
<th>OLS - Huber-White Estimators</th>
<th>Median regression</th>
<th>Robust regression - reweighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployment rate</td>
<td>72.062</td>
<td>78.462</td>
<td>72.062</td>
<td>89.791</td>
<td>73.811</td>
</tr>
<tr>
<td>Gross state product</td>
<td>65.867</td>
<td>90.997</td>
<td>65.867</td>
<td>57.916</td>
<td>64.262</td>
</tr>
</tbody>
</table>

It appears that the worst thing we can do is omit California from the model. This option has the most dramatic effect on the association between the gross state product and the number of violent crimes per 100,000. In fact, we would
overestimate by a substantial margin the association between these two variables if we simply removed California from consideration. Yet, the OLS model without adjustment for influential observations offers results consistent with the other, more robust regression models. Note also that the unemployment rate coefficient shifts considerably in the median regression model. You may wish to study this technique further to understand why this shift occurs.

We have only touched the surface of the abundant information about influential observations and the many robust regression techniques that are available. Much more information is provided in Peter J. Rousseeuw and Annick M. Leroy (2003), *Robust Regression and Outlier Detection*, New York: Wiley-Interscience. John Fox (1997), *op. cit.*., also provides an overview of several robust regression techniques. The main point is that influential observations are a common part of regression models. Although they are often caused by coding errors or reflect highly skewed variables, they are also a routine part of data collection and variable construction. Exploring your data carefully before estimating the model and assessing influence measures after estimating the model should therefore be part of any regression exercise.
13 A Brief Introduction to Logistic Regression

Although this presentation has emphasized dependent variables that are continuous and either normally distributed or transformable to normality, it is actually the rare research endeavor that uses only these types of variables. It is much more common to find dependent variables that are measured as categories, such as low, medium, and high; or yes and no. The topic of categorical dependent variables requires a full book in order to examine the many modeling possibilities (see, e.g., Hoffmann, op. cit., for an overview). Therefore, in this chapter we’ll briefly introduce and discuss perhaps the most popular of the categorical variable regression techniques: Logistic regression. To be precise, this model is identified as binary logistic regression because it is concerned with binary dependent variables.

There are many examples of binary variables (also known as dichotomous variables) in the social, behavioral, and health sciences. We’ve seen some examples when discussing dummy variables: Gender, race/ethnicity, religious preferences, and many others may be conceptualized as one or more two-category variables. It is not uncommon to find binary variables as dependent variables. For example, some binary dependent variables that appear in the research literature include support for the death penalty, graduation from college, whether companies in a sample test for illegal drugs, whether people vote, whether Republican or Democratic Presidential nominees win a state, whether adolescents drink alcohol, and whether a medicine cures an illness. All of these variables lend themselves, some better than others, to yes or no responses.

It should be clear that this type of variable violates at least two of the key assumptions of multiple linear regression models. In Chapter 3 we learned that one of these assumptions states: The error term is normally distributed with a mean of zero and constant variance. Another assumption declares: The mean value of Y for each specific combination of the X’s is a linear function of the X’s. Now, imagine a binary dependent variable in a model; can it satisfy either of these assumptions?
The quick answer is no. In order to gather evidence to this effect, consider one of the variables that appears in the data set `lifesat.sav`. The variable, also labeled `lifesat`, measures respondents’ self-reported satisfaction with life using two categories, low (coded zero) and high (coded 1). Of course, it would be preferable to have a continuous measure of life satisfaction, but suppose this is all we have. Can we assess predictors of this variable in a regression context? For instance, does age predict life satisfaction?

As an initial step, let’s ask SPSS to construct a scatterplot with life satisfaction (`lifesat`) on the y-axis and age in years (`age`) on the x-axis. Place a linear fit line in the scatterplot. You should find a plot with a fit line that looks like the following graph (the age and life satisfaction scales have been revised from the SPSS output):

![Scatterplot with linear fit line](image)

Simply looking at the fit line, we would conclude that there is a negative association between age and life satisfaction. However, notice how peculiar the scatterplot looks. All the life satisfaction scores fall in line with the values of zero or one. This, in itself, is not peculiar since respondents can take on scores of only zero or one. However, try to imagine lines running from the points to the fit line. Notice that there will be a systematic pattern to these residuals. Is this indicative of any problems we’ve discussed in previous chapters? Of course it is: There is a systematic pattern that indicates heteroscedasticity.

As a next step, we’ll use SPSS to estimate a linear regression model with life satisfaction as the dependent variable and age as the independent variable. When
setting up the model, ask for a normal probability plot of the residuals (using the Plots window). The results (see the table below) provide an unstandardized coefficient of –0.016. Hence, the interpretation is that each one year increase in age is associated with a 0.016 decrease in life satisfaction. There is also a standardized coefficient of –0.15: Each one standard deviation increase in age is associated with a 0.15 standard deviation decrease in life satisfaction. Do these interpretations make sense? Can life satisfaction decrease by 0.016 units or by 0.15 standard deviations? Recall that we also had problems interpreting dummy variables using standardized coefficients (see Chapter 6). This illustrates a key problem with using a linear regression model to predict a binary dependent variable. Look at the normal probability plot of the residuals. Is it even possible to claim that the residuals approach a normal distribution? This would constitute a leap in interpretation no sincere analyst would be willing to make. Another problem is that, even though these types of dependent variables can take values of only zero or one, it is possible to obtain predicted values from a linear regression equation that exceed one or are less than zero. For instance, we might predict that a person’s life satisfaction score is a meaningless –0.3 or a 1.75.

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>(Constant)</td>
<td>1.094</td>
<td>.406</td>
<td>2.698</td>
</tr>
<tr>
<td></td>
<td>Age in years</td>
<td>-.016</td>
<td>.011</td>
<td>-.149</td>
</tr>
</tbody>
</table>

a Dependent Variable: Life satisfaction

An Alternative to the Linear Model: Logistic Regression

One assumption that has appeared in the statistical literature is that, as long as the binary dependent variable is distributed somewhere close to 50% in one category and 50% in the other (some claim that anything between 30% and 70% is acceptable), then the linear regression model – which is labeled a linear probability model when it includes a binary dependent variable – yields acceptable results. But there is no longer any reason to rely on this model since there are several widely
available alternatives. We’ll discuss the most popular alternative model used in the social, behavior, and health sciences: Binary logistic regression. As background, consider that, when interested in a non-normal dependent variable, one of the tools we used was a transformation. A variable with a long right-tail, for example, is typically transformed using the natural logarithm or square root function. Although not as apparent from a mathematical point of view, we may also transform a binary dependent variable using a specific function. Before seeing this function, it is important to recall some material from elementary statistics.

Remember when you were asked in some secondary school mathematics course to flip a coin and estimate elementary probabilities? You were probably told that, given a fair coin, the probability of a heads was 0.50 and the probability of a tails was also 0.50. That is, from a frequentist view of statistics, we expect that if we flip a coin numerous times about half the flips will come up as heads and the other half will come up as tails. This is usually represented as \( P(H) = 0.50 \), with \( P \) shorthand notation for probability (see Chapter 1). The development of logistic regression – as well as other techniques for binary dependent variables – was predicated on the notion that a binary variable could be represented as a probability. Returning to our earlier example, we might ask the probability of a person reporting high or low life satisfaction. Using our shorthand method, this is depicted as \( P(\text{high life satisfaction}) = 0.49 \). In practical terms, this means that about 49% of the sample respondents report high life satisfaction. A fundamental rule of probabilities is that they must add up to one. Hence, if there are only two possibilities, we know that the probability of low life satisfaction \( P(\text{low life satisfaction}) \) is 1 – 0.49 = 0.51.

The simplest way to determine these probabilities in SPSS is to ask for Frequencies (using the Analyze-Descriptives drop-down menu) for the lifesat variable.

Once we shift our interest from estimating the mean of the dependent variable to estimating the probability that it takes on one of the two possible choices, we may utilize a regression model that is designed to predict probabilities rather than means. This is what a binary logistic regression model is designed to do: It estimates the probability that a binary variable takes on the value of one rather than zero (notice that, like dummy variables, we assume that the dependent variable is coded \{0, 1\}). In our example, we may use independent variables – which may be
LOGISTIC REGRESSION

221

continuous or dummy variables – to estimate the probability that life satisfaction is
high.

Binary logistic regression accomplishes this feat by transforming the linear
regression equation through the use of the following logistic function:

\[
P(Y = 1) = \frac{1}{1 + \exp[-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)]}
\]

The part of the denominator in parentheses is similar to the linear regression model
that we’ve used in earlier chapters. But note that it is transformed in a very specific
way. The advantage of this function is that it guarantees that the predicted values
range from zero to one, just as probabilities are supposed to do. For example, the
following table shows several negative and positive predicted values (i.e., from the
linear regression equation in the denominator) and their values after running them
through the function:

<table>
<thead>
<tr>
<th>Initial value</th>
<th>Transformed value</th>
</tr>
</thead>
<tbody>
<tr>
<td>–10.0</td>
<td>0.000045</td>
</tr>
<tr>
<td>–5.0</td>
<td>0.006693</td>
</tr>
<tr>
<td>–0.5</td>
<td>0.487503</td>
</tr>
<tr>
<td>0.0</td>
<td>0.500000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.622459</td>
</tr>
<tr>
<td>5.0</td>
<td>0.993307</td>
</tr>
<tr>
<td>10.0</td>
<td>0.999955</td>
</tr>
</tbody>
</table>

If we placed more extreme values in the function it would return numbers closer to
zero or one, but they would never fall outside the boundary of [0, 1].

Logistic regression models do not use OLS or some similar estimation routine.
Rather they use a widely-used estimation procedure known as Maximum Likelihood
(ML). ML is a common statistical procedure, but its particulars require more
statistical knowledge than we currently have at our disposal. Suffice it to say that
ML estimates the most likely value of a statistic, whether the value of interest is a
mean, a regression coefficient, or a standard error, given a particular set of data.
Detailed information on ML estimation is available in most books on statistical

Fortunately or unfortunately (you’ll be able to choose one of these after a couple more pages), logistic regression models are not often used to come up with predicted probabilities. Although there is no good reason why this occurs, most researchers who use logistic regression employ _odds_ and _odds ratios_ to understand the results of the model.

Odds are used frequently in games of chance. For example, slot machines and roulette wheels are usually “handicapped” by being accompanied by the odds of winning any particular attempt. Horse races are also supplemented by odds: “The odds that Felix the Thoroughbred will win the Apple Downey Half-Miler are four to one.” In general, odds are simply probabilities that have been transformed using the following equation:

\[
\frac{P}{1 - P}
\]

Say that our thoroughbred has been in ten half-mile races and won two of them. We may then say that the probability that she wins a half-miler is 0.20. This translates into an odds of winning of \(0.20/(1 - 0.20) = 0.25\). Another way of saying this is that the odds of winning are four to one, or for each race she wins she loses four of them.

An example closer to the interests of the research community is the following. Suppose we conduct a survey of adolescents and find that 25% report marijuana use and 75% report no marijuana use in the past year. Hence, the probability of marijuana use \(P(\text{marijuana use})\) among adolescents in the sample is 0.25. What are the odds of marijuana use?

\[
\frac{P(\text{marijuana use})}{1 - P(\text{marijuana use})} = \frac{0.25}{1 - 0.25} = 0.33 \text{ or } 1/3
\]

This implies that for every three adolescents who do not use marijuana, we expect one adolescent to use marijuana. In other words, three times as many adolescents have not used marijuana as have used marijuana in the past year.
This is not yet very interesting since we have restricted our attention so far to only one variable. Let’s extend our attention to two variables. We’ll treat adolescent past-year marijuana use as the dependent variable. To keep it simple, we’ll use a dummy variable that measures gender (coded 0=female; 1=male). Now we have up to four outcomes, with males and females who have used or have not used marijuana. If interested in probabilities, we may compare the two: \( P(\text{male marijuana use}) \) and \( P(\text{female marijuana use}) \). However, in keeping with our interest in odds, we’ll compare the odds of marijuana use among male and female adolescents.

Let’s assume that our survey data indicate that 40% of male adolescents and 20% of female adolescents report past-year marijuana use. Hence, the odds of marijuana use among males is \( \frac{0.40}{0.60} = 0.667 \) or \( \frac{2}{3} \) and the odds of marijuana use among females is \( \frac{0.20}{0.80} = 0.25 \) or \( \frac{1}{4} \). Keep in mind that these are not probabilities; odds are different. Another way of saying this that for every two males who used marijuana, three did not; and for every one female who used marijuana, four did not.

An odds ratio is just what it sounds like: the ratio of two odds. Let’s continue our adolescent marijuana use example to see the utility of this measure. Assume we wish to compare marijuana use among males and females. What is the odds ratio for these two groups? First, we have to decide on which is the focal group and which is the comparison group. Let’s use males as the focal group since they are more likely to report use. The odds ratio (OR) of males to females is

\[
\text{OR}_{\text{males vs. females}} = \frac{\text{Odds}(\text{males})}{\text{Odds}(\text{females})} = \frac{P(\text{males})}{1 - P(\text{males})} / \frac{P(\text{females})}{1 - P(\text{females})} = \frac{0.667}{0.25} = 2.67
\]

An odds ratio of 2.67 gives rise to the following interpretation: The odds of past-year marijuana use among males are 2.67 times the odds of past-year marijuana use among females. Some analysts use a shorthand phrase that males are 2.67 times as likely as females to use marijuana. However, this can mislead some observers into thinking that we are dealing with probabilities rather than odds. The ratio of the probabilities is \( \frac{0.40}{0.20} = 2.0 \); clearly not the same as the odds ratio.
Odds ratios are simply two odds that are compared to determine whether one group has higher or lower odds than another group on some binary outcome. A number greater than one implies a positive association between an independent and dependent variable (but keep in mind that the coding of the variables drives the interpretation) and a number between zero and one indicates a negative association. If there is no difference in the odds, then the odds ratio is one. For instance, if males and females are equally likely to report past-year marijuana use, then the odds ratio is 1.0.

A logistic regression model is one way to compute predicted odds and odds ratios for binary dependent variables. A useful equation that lends itself to these computations is

\[ \text{logit}[P(y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \]

where the logit is defined as

\[ \text{logit}[P(y = 1)] = \log_e \left( \frac{P(y = 1)}{1 - P(y = 1)} \right) \]

The quantity \( \log_e \left( \frac{P(y = 1)}{1 - P(y = 1)} \right) \) is also known as the log-odds. The term log-odds indicates the natural logarithm. Recall the logistic function that was specified earlier. It is another version of this equation. Moreover, the term log-odds should bring to mind a possible solution to computing the odds: Take the exponential of this quantity to transform it from log-odds to odds (or, as we shall see, an odds ratio).

This can be a confusing exercise, so let’s return to a real example of a binary dependent variable. In the data set we used earlier, lifesat, there is a variable gender that is coded as 0=female and 1=male. We’ll treat gender as the independent variable and life satisfaction as the dependent variable. The results of an SPSS cross-tabulation for these two variables are shown on the next page.

What are some of the relevant probabilities and odds from this table? First, the probability of high life satisfaction overall is 50/103 = 0.485. The overall odds of high life satisfaction are 50/53 = 0.94 (notice that we do not need to compute probabilities to compute odds). These overall probabilities and odds may be interesting, but they’re not very informative. Rather, we may also consider specific
measures among males and females. The probabilities of high life satisfaction among males and females are $37/85 = 0.435$ and $13/18 = 0.722$. The respective odds are $37/48 = 0.771$ among females and $13/5 = 2.6$ among males. We can easily see that males are more likely than females to report high life satisfaction.

### Life Satisfaction

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>48</td>
<td>37</td>
<td>85</td>
</tr>
<tr>
<td>Male</td>
<td>5</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>53</td>
<td>50</td>
<td>103</td>
</tr>
</tbody>
</table>

The odds ratio provides a summary measure of the association between gender and life satisfaction. Since we have the two odds, computing the odds ratio is simple: $\text{OR}_{\text{males vs. females}} = 2.6/0.771 = 3.37$. An interpretation is that the odds of high life satisfaction among males are 3.37 times the odds of high life satisfaction among females. Recall that an odds ratio greater than 1.0 indicates a positive association. Since males comprise the higher-coded category, we claim a positive association between gender and life satisfaction (without, of course, making any qualitative judgments about gender and life satisfaction).

It appears that there is little advantage to going beyond this simple exercise to estimate the association between gender and life satisfaction. However, keep in mind that we might (1) wish to determine whether there is a statistically significant association between gender and life satisfaction (assuming we have a good sample from a population) and (2) want to add additional independent variables to determine whether the association is spurious. This is where binary logistic regression comes in handy: It estimates standard errors and $p$-values; and it allows additional independent variables in the model.

First, we’ll run the simple logistic model with one independent variable. In SPSS, this model is requested with the **Analyze-Regression-Binary Logistic** drop-down menu. There is a box for the dependent variable (place `lifesat` there) and a box for covariates (which is another name for the independent variables) (place `gender` there). Then click OK. Scrolling down the screen, the following table should appear in the output window:
The appearance of the results is similar to that obtained using a linear regression model, with a couple of important exceptions. First, there is a column labeled \textit{Wald}. This is the Wald statistic (for the statistician A. Wald), which is similar to a $t$-value in linear regression, except it is the coefficient divided by its standard error squared. For example, the Wald value for the gender row is $(1.216/.57)^2 = 4.551$. This is used to calculate the $p$-value listed in the column Sig. It is used in the same manner as in the linear regression model, with, for instance, a value below 0.05 typically recognized as designating a statistically significant result.

The column labeled \textit{Exp(B)} is of particular interest. The value in the gender row should look familiar: It is the same value we obtained for the odds ratio from the cross-tabulation (but keep in mind that this works only because we coded female as zero and male as one). In fact, it is the odds ratio for the gender–life satisfaction association. However, assuming we wish to infer something about the population from which the sample was drawn, we now may determine whether this odds ratio is significantly different from 1.0 (why 1.0?). According to the $p$-value of 0.033, we may claim that it is significantly different from 1.0 at the $p<0.05$ level and conclude (tentatively, in the absence of much more information) that males have significantly higher odds of reporting high life satisfaction. Using your knowledge of significance tests in regression models (see Chapter 2), how would you interpret a $p$-value of 0.033?

The interpretation of the odds ratio is the same as it was earlier: The odds of high life satisfaction among males are expected to be 3.37 times the odds of high life satisfaction among females. Notice that we’ve added the term “expected to be” rather than something more determinate. This is to remind the reader that we are inferring from a sample to a population. We can never be certain, given a single sample, that the odds among males \textit{are} 3.37 times higher; only that we expect or infer them to be this much higher given the model results.
LOGISTIC REGRESSION

Suppose we wish to figure out the odds of high life satisfaction among males and females rather than just the summary odds ratio. Just like in linear regression analysis, we may estimate predicted values. Hence, the predicted value for females is \([-0.26 + (1.216 \times 0)]\) = -0.26 and the predicted value for males is \([-0.26 + (1.216 \times 1)]\) = 0.956. Recall the equation presented earlier in the chapter:

\[
\text{logit}[P(y = 1)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,
\]

where the logit \([P(y = 1)]\) = \(\log\left(\frac{P(y = 1)}{1 - P(y = 1)}\right)\)

Notice that the values we just predicted are therefore known as logit-values or, more commonly, log-odds. To transform them into odds, we simply take the exponential (inverse of the natural logarithm) of these values. The predicted odds for females are thus \(\exp(-0.26) = 0.771\) and for males \(\exp(0.956) = 2.601\). Compare these to the odds we computed using the cross-tabulation of gender and life satisfaction.

There are some researchers who are uncomfortable with odds and odds ratios and prefer probabilities. The coefficients from a logistic regression model may be transformed into probabilities by utilizing the logistic function:

\[
P(y = 1) = \frac{1}{1 + \exp[-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)]}
\]

We simply substitute the part of the denominator in parentheses with the logistic regression coefficients. However, we must choose specific groups to represent. With the previous model, there are only two groups, females (coded 0) and males (coded 1), and we may predict their probabilities by including the coefficients in the equation:

\[
P(\text{females}) = \frac{1}{1 + \exp[-(-0.26 + (1.216 \times 0))] } = 0.435
\]

\[
P(\text{males}) = \frac{1}{1 + \exp[-(-0.26 + (1.216 \times 1))] } = 0.722
\]
Notice that these predicted probabilities are identical to the probabilities we computed with the cross-tabulation of gender by life satisfaction. In SPSS (as well as most other statistical software packages) you may also save the predicted probabilities and then assess their values for particular groups that were included in the model. In SPSS the simplest way to assess predicted probabilities is to first save them (open up the Save window in the Binary Logistic menu and check the appropriate box) from the model and then use the Analyze-Compare Means-Mean drop-down menu with the predicted probabilities (labeled PRE_1) in the Dependent List box and gender in the Independent List box. However, note that this works best with dummy variables or independent variables with only a few categories. There are also options for residuals and influence statistics that we will ignore in this presentation but are worth exploring and understanding if you wish to realize the capabilities and limitations of specific logistic regression models.

**Extending the Logistic Regression Model**

We have not reached the limit of the binary logistic regression model. Rather, as with the linear regression model, we may estimate models with multiple independent variables – whether dummy or continuous – and interaction terms. Let’s throw a multiple logistic regression model together. We’ll continue to use life satisfaction as the dependent variable and include the following independent variables: gender, age, and intelligence scores (intell). We’ll treat age and intelligence scores as continuous variables. The results of this model are in the table on the next page.

The model indicates that, controlling for the effects of age and intelligence scores, the odds of high life satisfaction among males is expected to be 3.39 times the odds of high life satisfaction among females. This is also known as an *adjusted odds ratio* since we have adjusted for the presumed effects of other variables in the model. Note, though, that controlling for the effects of these other variables has little effect on the association between gender and life satisfaction. Moreover, we may interpret the Exp(B) for age or intelligence scores. Let’s focus on age since it is at least close to the normal threshold ($p < .05$) for statistical significance. Given that it is a continuous variable measured in year increments, the interpretation is
Controlling for the effects of gender and intelligence scores, each one year increase in age is associated with a 0.91 times difference in the odds of high life satisfaction.

Multiple Binary Logistic Regression Model

<table>
<thead>
<tr>
<th>Step 1(a)</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>1.222</td>
<td>.579</td>
<td>4.454</td>
<td>1</td>
<td>.035</td>
<td>3.393</td>
</tr>
<tr>
<td>age</td>
<td>-.095</td>
<td>.051</td>
<td>3.481</td>
<td>1</td>
<td>.062</td>
<td>.910</td>
</tr>
<tr>
<td>intell</td>
<td>-.054</td>
<td>.046</td>
<td>1.352</td>
<td>1</td>
<td>.245</td>
<td>.948</td>
</tr>
<tr>
<td>Constant</td>
<td>8.247</td>
<td>5.391</td>
<td>2.340</td>
<td>1</td>
<td>.126</td>
<td>3814.915</td>
</tr>
</tbody>
</table>

a Variable(s) entered on step 1: gender, age, intell.

This type of interpretation is not very satisfactory to most analysts. It is difficult, without significant experience, to interpret this as a negative association between age and life satisfaction. However, there is a nice property of regression models that transform dependent variables using natural logarithms that is useful in this situation. That is, we may use the following percentage change formula to interpret coefficients from logistic regression models:

\[
\{\exp(\beta) - 1\} \times 100
\]

This formula uses the untransformed coefficient (labeled B in the SPSS output) to determine the percent difference or change in the odds associated with a one unit difference or change in the independent variable. Using the age coefficient and this formula, we find \{[\exp(-0.095)] - 1 \times 100\} = -9.06%. Thus, the interpretation of the coefficient is

Controlling for the effects of gender and intelligence scores, each one year increase in age is associated with a 9.06% decrease in the odds of high life satisfaction.

Multiple logistic regression models may also be used to estimate predicted odds or probabilities. This is complicated by the presence of continuous variables,
however. As with multiple linear regression models (see, e.g., Chapter 6), we should choose particular values of the independent variables to compare. Some researchers prefer to use minimize and maximum values, but any plausible values, as long as they are represented by a sufficient number of cases in the data set, work reasonably well. For example, we may use the logistic function to compare the probabilities for males and females at the modal categories of age (41) and intelligence score (92):

\[
P(\text{females}) = \frac{1}{1 + \exp[-(8.25 + [1.22 \times 0] + [-0.095 \times 41] + [-0.054 \times 92])]} = 0.351
\]

\[
P(\text{males}) = \frac{1}{1 + \exp[-(8.25 + [1.22 \times 1] + [-0.095 \times 41] + [-0.054 \times 92])]} = 0.647
\]

Another way of thinking about these predicted probabilities is to infer that approximately 35% of females in the modal categories of age and intelligence are expected to report high life satisfaction; whereas about 65% of males in these categories are expected to report high life satisfaction.

Finally, there are tests of fit for logistic regression models that are analogous to fit statistics in linear regression models. Nonetheless, these are not computed in the same way, nor should they be interpreted in the same way as \( F \)-values, \( S_e \) values, or \( R^2 \) values. They may, however, be used to compare models. For example, SPSS provides a –2 log likelihood statistic (also referred to as a Deviance statistic; see the Model Summary table) that may be used to compare nested models (but not non-nested models; see Chapter 4). Values closer to zero tend to indicate better fitting models, although additional information is required before one may reach this conclusion. As an example, the original model with only gender as an independent variable has a Deviance of 137.678, whereas the model with gender, age, and intelligence scores has a Deviance of 133.817. To compare these models, subtract one from the other. The result is distributed \( \chi^2 \) with degrees of freedom equal to the difference in the number of independent variables across the two models. Therefore, one takes the difference between the two Deviances – in this case 137.678 – 133.817 = 3.861 – and compares it to a \( \chi^2 \) value with two degrees of freedom (since we added two independent variables in the second model). The \( p- \)
The value associated with this $\chi^2$ value is 0.145, which suggests that the difference between the two models is not statistically significant. In other words, we do not improve the model’s predictive capabilities by adding age and intelligence score. In the spirit of parsimony, we should conclude that the simpler model provides a better fit to the data.

The other two fit statistics provided in the SPSS output are labeled as R square values. However, these are not measures of explained variance (see Chapter 4). They should be used with caution since they measure fit by comparing the estimated logistic regression model with a model with no independent variables. They use a transformed version of the Deviance values to compare models, but there is no acceptable standard of interpretation for either measure.

Two fit statistics that are widely used are Akaike’s Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Both use the Deviance value from the estimated models. The standard formulas for these statistics (there are several variations) are

$$\text{AIC} = \frac{\text{Deviance} + 2(k + 1)}{n}$$

$$\text{BIC} = \text{Deviance} - \{df_k \times \ln(n)\}$$

The $df_k$ in the BIC formula is $(n - k - 1)$. The rule-of-thumb for both statistics is that smaller values indicate better fitting models. However, the BIC is a typically a negative number (a positive number indicates a very poor model), so one should be careful when evaluating it. Smaller implies that a BIC farther from zero in the negative direction indicates a better fitting model.

The AICs and BICs for the two logistic regression models of life satisfaction are provided in the table on the following page. Both statistics indicate that the logistic regression model with gender only provides a better fit to the data than the model with gender, age, and intelligence score. More information about these and other fit statistics is provided in Walter Zucchini (2000), “An Introduction to Model Selection,” Journal of Mathematical Psychology 44: 41-61.
LINEAR REGRESSION ANALYSIS

<table>
<thead>
<tr>
<th></th>
<th>Independent Variables</th>
<th>Gender only</th>
<th>Gender, age, and intelligence scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>1.376</td>
<td>1.377</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>−330.429</td>
<td>−325.021</td>
<td></td>
</tr>
</tbody>
</table>

Binary logistic regression provides a powerful tool for evaluating dependent variables that take on only two values. These types of variables are not uncommon in the various scientific disciplines that utilize statistical analyses. Rather, they are of broad interest and provide important information about many relevant phenomena. We have only presented a brief introduction to this valuable regression technique. Moreover, we have not addressed a twin model: Probit regression. Probit regression provides similar results in general (although the coefficients are scaled differently), but is based on the assumption that underlying a binary variable is a continuous variable that is represented along a continuum from a probability of zero to a probability of one. Substantially more information on logistic and probit regression is provided in the numerous books that address categorical dependent variables. An especially lucid guide to logistic models is provided in David G. Kleinbaum and Mitchel Klein (2002), *Logistic Regression: A Self-Learning Text, Second Edition*, New York: Springer.
14 Conclusions

The linear regression model offers a powerful tool for analyzing the association between one or more independent variables and a single dependent variable. Some novice researchers wish to move quickly beyond this model and learn to use more sophisticated models because they get discouraged about its limitations and believe that other regression models are more appropriate for their analysis needs. Although there are many situations where alternative regression models are needed (see, e.g., Chapter 13), the linear model should not be overlooked or underutilized. Even when one or more of the assumptions described in the previous chapters is violated, linear regression will frequently provide accurate results. Although saying this model is robust is ill-advised because robustness has a very specific statistical meaning, it is not disingenuous to claim that the model works well under a variety of circumstances. There are myriad practical uses in the social, behavioral, and health sciences for linear regression analysis.

The key to using linear regression analysis appropriately is to understand its strengths and weaknesses. It works quite nicely with dummy and continuous independent variables. As long as the dependent variable can be transformed to something close to normality, the coefficients and standard errors are within a reasonable distance to the true population parameters (assuming, of course, that we have a good sample). And, with adjustments to the standard errors that minimize the negative effects of heteroscedasticity and influential observations, the results can move even closer to the precision most of us desire. In particular, Huber-White sandwich estimators are especially useful when heteroscedasticity or influential observations show up or are suspected (see Chapters 11 and 12).

However, linear regression models should not be used when the dependent variable is binary or takes on only a few values. How many is “a few” has been a point of contention in the research literature. Some claim less than nine, others less than seven. They key, though, is whether the variable is distributed normally, not precisely how many categories it has. Nevertheless, given the availability of the many regression models designed for categorical dependent variables, there is little
need to rely on linear regression analysis in this situation. The binary logistic regression model, for example, provides a solid alternative when the dependent variable is binary. Other models are described in the many books on categorical analysis and generalized linear models (see, e.g., Hoffmann, op. cit.).

There are many other topics that we have not had time to address in this presentation. For example, the issue of non-independence was mentioned only briefly when discussing the assumptions of linear regression analysis (Chapter 2) and autocorrelation (Chapter 11). Non-independence occurs when we have longitudinal data or spatial data, but also in other contexts. Recall that standard errors are biased in the presence of non-independent observations.

The general issue of non-independence of observations is particularly germane to the more specific topic of survey sampling. It is rare to find large surveys that do not use some form of sampling that leads to non-independence of observations. These sampling schemes are conducted for efficiency and cost-effectiveness. It is prohibitively expensive to gather a simple random sample of people in the United States. Hence, most national surveys (and many state surveys) use some form of stratified or cluster sampling. This means that units, such as metropolitan areas, are sampled first, followed by smaller units such as census tracts, block groups, and households. There are usually several stages to the sampling. Telephone surveys typically use groups of exchanges that also result in a certain degree of clustering. When using most national data sets, it is therefore important to consider the effects of stratification and clustering on regression estimates. As mentioned earlier, the main effect is on the standard errors, so they should be adjusted. Many statistical software packages now include techniques for survey data that adjust standard errors for non-independence. (For example, SPSS has recently introduced a new module called Complex Samples that allows analysts to consider non-independent observations from surveys; SAS and Stata also include a set of statistical techniques for survey data that include non-independent observations.)

Another issue related to sampling involves the use of weights. Sampling weights are employed to designate the number of people (or other units) in the population each sample member represents. Recall that we want our sample to represent some well-understood population. Therefore, each observation in our sample should represent some group from the population. For example, in a nationally repre-
sentative sample a 40-year-old Caucasian man may represent several thousand 40-
year-old Caucasian men from some part of the U.S. Survey data often include
sample weights that may be used in analyses (e.g., our 40-year-old man has a
weight of 12,340 since he represents 12,430 40-year-old men). In SPSS, the drop-
don menu Data-Weight Cases is used to weight analyses by sampling weights.
Yet, if these weights are whole numbers designed to indicate the actual number of
people each observation represents, the presumed sample size can be inflated
tremendously. What effect do larger samples have on standard errors? All else being
equal, they make them smaller. So, experience shows that almost any regression
coefficient is statistically significant – whether or not it is important – if the sample
is weighted to represent a large population. This may be seen as an advantage, but
it can also be misleading and dubious results. One solution is to compute normed
sampling weights for the analysis. These are computed as

$$\bar{w}_i = \frac{w_i}{\bar{w}}$$

In words, take each observed weight and divide it by the overall mean of the
weights.

Let’s say our national sample has a mean weight of 15,000 (on average, each
observation represents 15,000 people). Our 40-year-old man will then have a
normed sampling weight of 12,430/15,000 = 0.829. If another observation
represents more people, say 17,500, then its weight is 17,500/15,000 = 1.167.
Taking the normed weights not only makes the weights smaller and more
manageable, but also preserves their utility since some observations still have more
influence than others. It also results in more reasonable standard errors.
Substantially more information on survey data and sampling weights may be found
in Paul S. Levy and Stanley Lemeshow (1999), Sampling of Populations: Methods

There are many other issues we could discuss, but space is limited. The
information provided in the previous chapters should provide the groundwork for
further coursework on and self-study of statistical modeling. Thus, perhaps it is
best to ask interested readers to pursue addition topics on their own through the
many excellent books and courses on linear regression and related statistical techniques. Topics such as sample selection bias, nonlinear models, bootstrapping methods, multilevel linear models, generalized linear models, simultaneous equations, methods of moments (MM) estimators, generalized additive models (GAMs), marginal models, event history analysis, and log-linear models (including transition models) have all been used to complement, supplement, or replace linear regression models in one way or another.

Summary

To summarize what we’ve learned, here is a list of steps to consider as you pursue linear regression analyses on your own. It is not necessarily a comprehensive list since each analysis and model has its own idiosyncrasies that make it unique. Nonetheless, it prescribes the most common steps that should normally be taken as you estimate linear regression models.

How to Conduct a Linear Regression Analysis: A Summary

Know the background literature on your subject. Does previous literature indicate that your dependent variable is not distributed normally or is not continuous? Has previous literature found interactions between independent variables? What control variables are important for your model? Do you anticipate measurement error? What might you do to correct for it? What theoretical frameworks or conceptual models have been used in the past? Use your imagination/intuition about the social or behavioral processes of interest.

Establish hypotheses and decision rules. Which theory or conceptual model guides your selection of variables or the ways you think they are associated? Why do you think the variables are associated? Will you use directional or non-directional hypotheses? Will you use one- or two-tailed significance tests to decide whether coefficients are statistically distinct from zero? Set up the hypotheses.

Know the data set. Are your data longitudinal? If they are, consider using a regression technique – such as a GEE – that is designed for longitudinal data. Do
your data come from a survey? Have you read the codebook and documentation for information about how the data were collected? How are the variables of interest coded? Are there missing values? How are they coded? How will you handle missing data? You will probably need to recode and construct new variables. Are there sample weights in the data set? Will you use a split-sample approach to verify the model?

**Check the distributions of your variables.** Use q-q plots and box-and-whisker plots. Run descriptive statistics. Is your dependent variable distributed normally? What about the independent variables? If your variables are not distributed normally, consider using a transformation such as the natural logarithm, square root, or quadratic (depending on the direction and degree of skewness). If your dependent variable is binary, use logistic regression. Are there outliers or other influential observations that you can see? Consider their source. Do you need to compute dummy variables and include them in your model? Check to make sure you’ve taken care of missing values – they can throw everything off if they are not adjusted for in the model!

**Assess the bivariate associations in your data.** Use scatterplots for continuous variables. Plot linear and nonlinear lines to determine the bivariate associations. Compute bivariate correlations for continuous variables. Look for outliers and potential collinearity problems.

**Run your regression model.** Avoid automated variable selection procedures unless your goal is simply to find the best prediction model. Assess the results. Are there unusual coefficients (overly large or small; negative when they should be positive)? Save the collinearity diagnostics; save the influential observation diagnostics (deleted studentized residuals, leverage values, Cook’s D, and DFFITS); run a scatterplot of deleted studentized residuals (or studentized residuals) by standardized predicted values; estimate partial residual plots; ask for a normal probability plot of the residuals; ask for the Durbin-Watson statistic, if appropriate; compute Moran’s $I$, if using spatial data and it is available. Assess the goodness-of-fit statistics (adjusted $R^2$; $F$-value and its accompanying $p$-value; $S_E$).
Run nested models if appropriate and use nested $F$-tests to compare them. These are particularly useful for assessing potential specification errors.

**Check the diagnostics.** Are there any collinearity problems? If yes, you might need to combine variables, collect more data, or, as last resort, drop variables. If the collinearity problem involves interaction or quadratic terms, use centered values such as $z$-scores to recompute them. Are the residuals normally distributed? If not, consider a transformation. Do the partial residual plots provide evidence of nonlinear associations? Is there evidence of heteroscedasticity? (If the plot is inconclusive visually, try White’s test or Glejser’s test.) If yes, consider transforming a variable, weighted least squares regression, or using Huber-White sandwich estimators. Is there evidence of autocorrelation? Consider the source and try to correct for it. If you have spatial data, a spatial regression model may be needed. Use Prais-Winsten regression or time-series techniques for data collected over time, if appropriate. Are there influential observations? If yes, consider their source. Are there coding errors? Will a transformation help? If not, use a robust regression technique to adjust for influential observations.

**Interpret and present the results.** Interpret the unstandardized slopes and $p$-values. What do the goodness-of-fits statistics tell you about the model? Compare the results to the guiding hypotheses. Given the decision rules, are the hypotheses or the conceptual model supported by the analysis? Consider presenting predicted values, especially from models that include interaction terms. Consider graphical presentations of coefficients, nonlinear associations, and interactions.
Appendix A

Data Management

In order to become a proficient user of statistical models – whether or not they are regression-based – it is important to always remember the key role of data management. This term is used to indicate a vital concern with the proper handling and coding of variables in a data set. It is unusual to be provided with a data set that includes variables that are immediately ready to use in a regression model or in any type of statistical analysis. (The data sets used in the previous chapters notwithstanding.) Rather, variables are often constructed and coded in a variety of ways, missing values are included without specific treatment, and categorical variables are not accompanied by dummy variables. Thus, this appendix offers some suggestions for managing data and variables. It is a good idea, as you are starting out with SPSS or any other statistical software, to read the documentation carefully and obtain one of the many useful guides to using the software (e.g., Eric L. Einspruch (2005), An Introductory Guide to SPSS for Windows®, Thousand Oaks, CA: Sage; Lawrence C. Hamilton (2004), Statistics with Stata, Belmont, CA: Brooks/Cole; Lora D. Delwiche and Susan J. Slaughter (2003), The Little SAS® Primer Book, Third Edition, Cary, NC: SAS Institute). Moreover, data sets should be accompanied by codebooks that describe the variables and their coding schemes. Browsing the codebook and fully understanding the variables of interest is a crucial prerequisite to any applied statistical exercise.

An elementary, yet important, issue to consider is the nature of data sets. Most data sets appear in spreadsheet format with the variables in columns and the observations in rows. For example, consider the table on the next page. It provides a revised and limited portion of the data set usdata.sav (as it appears in SPSS’s data view window). Notice that the variables – Robbrate, Larcrate, and so forth – appear as columns and the states – which constitute the observations in the data set – appear as rows. It is always a good idea to scan at least a portion of the data set (some are huge) to determine if there are any unusual values or patterns to the data.
or variables. In addition, evaluate carefully those variables that you plan to analyze. Are dummy variables or binary dependent variables coded as \{0, 1\}? Are there categorical variables that need to be transformed into sets of dummy variables? How are continuous variables coded? (e.g., Is income measured in dollars or thousands of dollars? What is the base number for rates?) Are there unusual values that don’t appear tenable (see Chapter 12)? Remember to always conduct exploratory analyses of your variables to check for outliers, non-normal distributions, and unusual codes. At the very least, ask the software for frequencies and the distributional properties of the variables (e.g., means, medians, standard deviations, skewness) you plan to analyze and evaluate each variable carefully.

<table>
<thead>
<tr>
<th>State</th>
<th>Robbrate</th>
<th>Larcrate</th>
<th>Asssrate</th>
<th>Fips</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>185.75</td>
<td>2844.27</td>
<td>403.69</td>
<td>1</td>
</tr>
<tr>
<td>Alaska</td>
<td>155.13</td>
<td>3624.34</td>
<td>526.32</td>
<td>2</td>
</tr>
<tr>
<td>Arizona</td>
<td>173.76</td>
<td>4925.63</td>
<td>495.71</td>
<td>4</td>
</tr>
<tr>
<td>Arkansas</td>
<td>125.68</td>
<td>2815.42</td>
<td>379.83</td>
<td>5</td>
</tr>
<tr>
<td>California</td>
<td>331.16</td>
<td>2856.87</td>
<td>590.26</td>
<td>6</td>
</tr>
<tr>
<td>Colorado</td>
<td>.</td>
<td>3634.48</td>
<td>298.75</td>
<td>8</td>
</tr>
<tr>
<td>Connecticut</td>
<td>163.21</td>
<td>2668.73</td>
<td>214.41</td>
<td>9</td>
</tr>
<tr>
<td>Delaware</td>
<td>198.74</td>
<td>3114.23</td>
<td>442.54</td>
<td>10</td>
</tr>
<tr>
<td>Florida</td>
<td>299.91</td>
<td>4322.41</td>
<td>715.11</td>
<td>12</td>
</tr>
<tr>
<td>Georgia</td>
<td>205.21</td>
<td>3678.27</td>
<td>407.14</td>
<td>13</td>
</tr>
<tr>
<td>Hawaii</td>
<td>130.83</td>
<td>5046.93</td>
<td>.</td>
<td>15</td>
</tr>
</tbody>
</table>

**Missing Data**

The portion of the *usdata.sav* data set that appears in the table has been revised by placing dots in two of the cells. These dots are a commonly used representation of missing values. But oftentimes missing values are coded with untenable values such as \(-99\) for age or a 9 for race/ethnicity when the other choices are 1 = Caucasian, 2 = African-American, 3 = Latino, and 4 = other race/ethnicity. There are many other possibilities; check the codebook accompanying the data for information about how missing values are treated.

Regardless of the coding scheme for missing values, before using a variable it is important to decide how you are going to handle missing values. As you do this, though, first consider their source. Why do missing values appear in your data set? The reasons will likely vary depending on the nature of the variable. An important
issue to consider is whether the data entry person forgot to place a value in a specific cell of the data set. Perhaps, in this example, we need to go back to the data source and find the robbery number for Colorado or the assault number for Hawaii. Yet, these data may not exist.

A common source of missing values in survey data involves skip patterns. Skip patterns in surveys are a convenient way to make responding more efficient. For instance, suppose we ask a sample of adolescents to fill out a survey about their use of cigarettes. Our questions ask details such as how often they smoke, how many cigarettes they smoke in a day or a week, from where do they obtain cigarettes, and so forth. A majority of adolescents don’t smoke at all, so it is common to find initial questions such as

Q.1. Have you ever smoked cigarettes? (circle one answer only)
(a) No [Go to question Q.16]
(b) Yes [Go to question Q.2]

Q.2. When was the last time you smoked a cigarette? (circle one answer only)
(a) Today
(b) In the last week
(c) More than a week ago but less than a month ago
(d) A month ago or longer

In contemporary surveys these types of questions are often programmed into a computer so that those responding “no” to the first question are automatically asked question 16 next. Suppose we wish to analyze the second question either as a dependent or an independent variable (perhaps by creating dummy variables). Would we only want to analyze data from those who actually answered the question; or would we want to also include those who never smoked in the statistical analysis? This is an essential issue to address since those who answered “no” to the first question will be missing on the second question. If we decide to include those who never smoked in the analysis, then it is a simple matter to create a new variable that includes a code for “never smoked” along with others for those who smoked in the last day, week, etc.
It is also not uncommon to find missing data due to refusals or “don’t know” responses (Question: “When was the last time you smoked a cigarette?” Response: “I don’t know”). For example, many people don’t like to answer questions about personal or family income. A perusal of various survey data shows that missing values on income are frequent, with usually more than 10% of respondents refusing to answer. Yet many researchers are interested in the association between income and a host of potential independent or dependent variables. Is it worth it to lose 10-20% of your sample so that you can include income in the analysis? Are people who refuse to answer income questions (or other types of questions) different from those who do answer? If they are systematically different, then the sample used in the analysis may no longer be a representative sample from the population.

Making decisions about missing values is a crucial part of most statistical exercises. There are several widely used solutions to missing data problems. First, some researchers simply ignore them by omitting them from the analysis. They argue that a few missing cases will not bias the results of their model too much. However, be very careful not to leave them with codes (such as –9 or 999) that will be included in the analysis. Imagine, for instance, if you left a missing code of 999 in an analysis of years of education! Now that would be an influential observation.

The most commonly used technique for “throwing out” missing cases is known as listwise deletion (this is the SPSS default). Suppose we analyze the association between robbery and assault in the revised usdata.sav data set shown earlier with a correlation or a linear regression model. In this situation, listwise deletion removes both Colorado and Hawaii from the analysis. Now imagine if we estimate a multiple linear regression model with, say, 5 independent variables. Suppose further that 10 different observations were each missing only one value, but from 5 different independent variables. Looking over the frequencies, we’d find 48 observations for each independent variable, but once we place all of them in the regression model and use listwise deletion, our sample size goes from 50 to 40; we lose one-fifth of the sample. This can be costly.

A problem that is not infrequent involves the use of multiple nested regression models with different patterns of missing data. Say we estimate a model with robberies per 100,000 predicted by per capita income. If there are no missing data
on either variable, the sample size is 50. We then add migrations per 100,000, which is missing two observations. A third model adds population density and the unemployment rate, each of which has four missing observations. We therefore have up to 10 missing observations. It is common to claim that the first two models are nested within the third, thus allowing statistical comparisons (see Chapter 5). However, they are not truly nested because they have different numbers of observations (n’s): 50 in the first model, 48 in the second model, and 40 – 44 in the third model. Not only do nested F-tests break down when this occurs, but the regression coefficients should not be compared across models. It is therefore imperative that missing data be handled before running any of the regression models so that each has the same number of observations.

There are a variety of techniques for handling missing values in addition to listwise deletion. These include mean substitution, regression substitution, raking techniques, hot deck techniques, creating a code for missing values, and multiple imputation. Mean substitution replaces missing values of a variable with the mean of that variable. For example, assume that the mean number of assaults per 100,000 is 342.4. We could then replace the missing value in SPSS using either the *Compute-Recode* drop-down menu or the following commands in a syntax file:

```
RECODE
  asrate (MISSING=342.4) .
EXECUTE .
```

A problem with mean substitution is that it can lead to biased slope coefficients and standard errors, especially as a higher proportion of observations is missing. Unfortunately, there is no rule-of-thumb about the proportion of missing values that lead to biased regression results when using mean substitution.

Regression substitution replaces missing values with values that are predicted from a regression equation. For instance, suppose that a small number of values are missing from the personal income variable that appears in the *gss96.sav* data set. Yet, there is complete data for education, occupational prestige, and age. If a regression equation does a good job of predicting the valid income values with
these three variables, then the missing values may be predicted by using the coefficients from the following linear regression model

\[
\text{income}_i = \alpha + \beta_1(\text{education}_i) + \beta_2(\text{occupation}_i) + \beta_3(\text{age}_i)
\]

An easy technique is to save the predicted values from the model and ask the program to substitute the predicted values for income when income is missing. An example of this is provided in the following SPSS syntax commands:

```
REGRESSION
/DEPENDENT pincome
/METHOD=ENTER educate occprest age
/SAVE PRED.
EXECUTE.

IF MISSING(pincome) pincome=PRE_1.
EXECUTE.
```

The syntax predicts personal income using a linear regression model, saves the predicted values (note that these are the unstandardized predicted values), then uses the IF command to replace the missing values with the predicted values from the model. However, if the variables are not good predictors of the variable with the missing value problem, then biased regression coefficients and standard errors will result.

Raking is an iterative procedure that is based on using values of adjacent observations to come up with an estimated value for the observation that is missing. Hot deck involves partitioning the data into groups that share similar characteristics and then randomly assigning a value based on the values of this group. Both require specialized software to implement.

Creating a code for a missing value is a common procedure. Suppose that we have a problem with personal income: 10% of the observations are missing. We could create a new dummy variable that is coded as 0=non-missing on income and 1=missing on income, and include this variable in the regression model, along
with other income dummy variables. This is only useful if the variable is transformed into a set of dummy variables, though. So income would have to be categorized into a number of discrete groups (e.g., 0 - $10,000; $10,001 - $20,000; etc.) that would then be the basis for a set of dummy variables.

It is rarely, if ever, a good idea to use these techniques to replace missing values for dependent variables. Some analysts argue that raking and hot-deck techniques are useful if there are only a few missing values in the dependent variable. However, if this is the case, multiple imputation is a better approach. Moreover, it has become a widely used technique throughout the regression world when there are missing values for independent variables. It involves three steps:

1. Impute, or fill in, the missing values of the data set, not once, but $q$ times (usually 3-5 times). The imputed values are drawn from a particular probability distribution, which depends on the nature of the variable (e.g., continuous, binary). This step results in $q$ complete data sets.
2. Analyze each of the $q$ data sets. This results in $q$ analyses.
3. Combine the results of the $q$ analyses into a final result. There are rules that are followed to combine them, usually by taking some average of the coefficients and standard errors if a regression model is used.

Specialized software or programs are used for multiple imputation since the imputation stage requires sophisticated algorithms. SPSS, SAS, and Stata have special routines for multiple imputation. An excellent overview of this technique is provided in Joe L. Schafer (1999), “Multiple Imputation: A Primer.” Statistical Methods in Medical Research 8: 3-15. For a general review of procedures for missing data, see Paul D. Allison (2001), Missing Data, Thousand Oaks, CA: Sage.

**Variable Creation, Coding, and Recoding**

It is highly likely that you will have to create new variables from those in a data set at some point. Many social and behavioral surveys include sets of questions designed to measure some phenomenon of interest. For example, there are sets of questions that measure self-esteem, symptoms of depression, illicit drug use, happiness, anxiety, self-efficacy, job satisfaction, trust, racial prejudice,
authoritarianism, religiosity, and many other interesting phenomena. Suppose you are studying adolescent depression and the survey questions are drawn from the CES-D scale, a widely used depression symptom inventory. You plan to estimate a linear regression model with depressive symptoms as the independent variable (along with gender, family closeness, and other measures) and self-efficacy as the dependent variable (we’ll ignore possible endogeneity problems; see Chapter 7). It is unlikely that you will wish to enter the variables representing each depression question separately. Rather, you will probably attempt to combine the questions in some way to come up with a global measure of depressive symptoms.

In Chapter 8, we learned that one way to combine variables is through factor analysis and structural equation modeling. Another way that generally ignores measurement error is to combine variables either through adding up their values or taking the mean of their values. When combining variables in this way, however, it is important that they are coded in the same direction and in the same way. For instance, if we are measuring symptoms of depression, all the items should be coded so that increasing numbers indicate either more or less symptoms. You may need to reverse code some items so that all are in the same direction. Furthermore, if one variable is coded 1 – 4 and another is coded 1 – 10, then adding them up or taking the mean of the items will be influenced more by the variable with more response categories. As an alternative, some researchers recommend first standardizing all the variables (taking their $z$-scores) and then adding them up or taking their mean. The reasoning is that even if the number of response categories differs, standardizing will “normalize” them so they are on the same scale.

There is a risk to simply adding up the variables, however. Suppose, for instance, that we use the following SPSS syntax to compute a new variable, Depress, from five variables in a depression inventory:

```
COMPUTE Depress = dep1 + dep2 + dep3 + dep4 + dep5 .
EXECUTE .
```

An alternative is to use the SUM function to add up the items:

```
COMPUTE Depress = SUM(dep1, dep2, dep3, dep4, dep5) .
EXECUTE .
```
But say there are different patterns of missing values for each of the variables. For some observations, the new variable will depend not just on their responses to the questions, but also on whether they are missing on one or more items. Respondents with response patterns of \{2, 2, 2, 2, 2\} will have the same depression score those with response patterns of \{4, 6, .., .\}, even though their level of depressive symptoms differs substantially.

Some researchers prefer to take the mean of the items. In SPSS this appears in a syntax file as

```plaintext
COMPUTE Depress = MEAN(dep1, dep2, dep3, dep4, dep5) .
EXECUTE .
```

Hence, respondents with the patterns listed earlier will have depression scores of 2 and \((4+6)/2 = 5\), which reflects more accurately their level of depressive symptoms.

Coding and recoding variables is a crucial part of any data management exercise, whether or not they are needed for statistical analyses. We’ve already learned about coding dummy variables (see Chapter 6). It is important to remember to create dummy variables that are mutually exclusive if they measure aspects of a particular variable (e.g., marital status). It is a good idea to run cross-tabulations on dummy variables – along with the variable from which they are created – to ensure that they are coded correctly.

Making sure continuous variables are coded properly is also essential. It is best to use coding strategies that are easily understandable and widely accepted, if possible. Measuring income in dollars, education in years, or age in years is commonplace. Yet, for various reasons, sometimes it is better to code income in $1,000s. Some data sets have education grouped into categories (e.g., 0 – 11 years, 12 years, 13 – 15 years, 16 years, etc.) rather than years. Keeping track of coding strategies is therefore of utmost importance. Constructing your own codebook or keeping track of codes and recodes by keeping copies of syntax files is useful and will allow you to come back later and efficiently recall the steps you took before analyzing the data.
There are numerous other coding issues that we will not address. Experience is the best teacher. Moreover, keeping careful records and back-up files of data sets that include before-and-after recodes of key variables will be helpful. Remember that once you save an updated SPSS file using the same name the old file is gone (unless you’ve made a back-up copy). So make sure you are satisfied with the variables before you overwrite an old data file.

Once you have completed taking care of missing values, creating new variables, and recoding old variables – as well as transforming non-normal variables – it is always a good idea to ask for frequencies (for categorical or dummy variables) and descriptive statistics (for continuous variables) for all of the variables that you plan to use in the analysis. SPSS’s Analyze-Reports-Case Summaries-Summarize Cases menu and LIST VARIABLES syntax command are also handy for assessing whether what you intended to do with the variables actually occurred.

As an example, let’s say we use the delinq.sav data set to estimate a linear regression model. First, though, we ask a colleague to transform the variable stress using the natural logarithm because previous research indicates that variables measuring stressful life events are positive skewed. Our colleague uses the following SPSS syntax command to transform the variable:

\[
\text{COMPUTE newstress = LN(stress) .} \\
\text{EXECUTE .}
\]

After asking for descriptive statistics on the newstress variable we find that it has more than 400 missing cases. Why did this occur? If we didn’t pay attention to the original variable or did not look carefully at the syntax file, we might simply think that more than 400 adolescents did not respond to this question during survey administration. Perhaps we’d chalk it up to the inexactness of data collection and ignore these cases. Or we might use a substitution or imputation procedure to replace the missing values. However, if we looked at the original stress variable – which we should always do – we’d find no missing cases. Something must have happened during the transformation.

Recall that taking the natural logarithm of zero is not possible; the result is undefined. But notice that the original stress variable has legitimate zero values
(some adolescents report no events), so the syntax needs to be revised to take into account the zeros in the variable. A simple solution is to revise the syntax to read

```
COMPUTE newstress = LN(stress + 1) .
EXECUTE .
```

Now, those with zero scores on the original variable will have zero (not missing) scores on the new variable. Yet we might not have found this out if we didn’t check the descriptive statistics for both the old and the new stress variables. Knowing your data and variables, and always checking distributions and frequencies following variable creation and recoding will go a long way toward preventing modeling headaches later.
Appendix B
Formulas

1. Mean of a variable
\[ E[X] = \bar{x} = \frac{\sum x_i}{n} \]

2. Sum of Squares of a variable
\[ SS[X] = \sum (x_i - \bar{x})^2 \]

3. Variance of a variable
\[ \text{Var}[X] = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \]

4. Standard Deviation of a variable
\[ \text{SD}[X] = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \]

5. Coefficient of Variation (CV) of a variable
\[ CV = \frac{s}{\bar{x}} \]
6. Standard Error of the Mean
\[ \text{se}(\bar{x}) = \frac{s}{\sqrt{n}} \]

7. Computing z-scores
\[ z - \text{score} = \frac{(x_i - \bar{x})}{s} \]

8. Covariance of two variables
\[ \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \]

9. Pearson’s Correlation of two variables
\[ \text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}[x] \times \text{Var}[y]}} \]
\[ \text{Corr}(x, y) = \frac{\sum (z_x)(z_y)}{n - 1} \]
10. $T$-Test comparing two means (equal variances)

$$t = \frac{\bar{x} - \bar{y}}{s \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $s = \sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 + n_2) - 2}}$

11. $T$-Test comparing two means (unequal variances)

$$t' = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\text{Var}[x]}{n_1} + \frac{\text{Var}[y]}{n_2}}}$$

12. General formula for a Confidence Interval (CI)

Point estimate $\pm [\text{(confidence level)} \times \text{(standard error)}]$ 

(Note: The point estimate can be a mean or a regression coefficient)

13. Sum of Squared Errors (SSE) for the linear regression model

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

14. Regression Sum of Squares (RSS) for the linear regression model

$$\text{RSS} = \sum (\hat{y}_i - \bar{y})^2$$
15. Residuals from a linear regression model
   \[ \text{resid}_i = (y_i - \hat{y}_i) \]

16. Slope in the simple linear regression model
   \[
   \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}
   \]

17. Intercept in the simple linear regression model
   \[ \hat{\alpha} = \bar{y} - (\hat{\beta}_1 \times \bar{x}) \]

18. Standard Error for a simple linear regression slope
   \[
   \text{se}(\hat{\beta}_1) = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2 / n - 2}{\sum (x_i - \bar{x})^2}} = \sqrt{\frac{\text{SSE}/n - 2}{\text{SS}[x]}}
   \]

19. T-Value for a linear regression slope
   \[ t - \text{value} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} \]

20. Standardized Linear Regression Coefficient (Beta-Weight)
   \[ \hat{\beta}^* = \hat{\beta} \left( \frac{s_x}{s_y} \right) \]
21. Multiple Linear Regression Slopes (matrix form)

\[ \hat{\beta} = (X'X)^{-1} X'Y \]

22. Variance-Covariance Matrix – multiple linear regression slopes

\[ V = (X'X)^{-1} \sigma^2 \quad \text{where} \quad \sigma^2 = \frac{1}{n-k} \times \hat{\epsilon}^T \hat{\epsilon} \]

23. Standard Error for multiple linear regression slopes

\[ se(\hat{\beta}_i) = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{\sum (x_i - \bar{x})^2 (1 - R_i^2)(n-k-1)}} \]

24. Coefficient of Determination for the linear regression model

\[ R^2 = \frac{RSS}{TSS} = 1 - \frac{SSE}{TSS} \]

25. Adjusted Coefficient of Determination for the linear regression model

\[ \bar{R}^2 = \left( R^2 - \frac{k}{n-1} \right) \left( \frac{n-1}{n-k-1} \right) \]

26. Mean Squared Error for the linear regression model

\[ MSE = \frac{SSE}{n-k-1} \quad \sqrt{MSE} = S_e \]
27. Prediction Intervals for the linear regression model

\[
\text{PI} = \bar{y} + \hat{\beta}_1(x_0 - \bar{x}) \pm (t_{n-2}S_e)(\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\text{Var}(x)}})
\]

28. Partial F-Test for comparing linear regression models

\[
F = \frac{\text{RSS (full)} - \text{RSS (restricted)}}{\text{MSE (full)}}
df = 1, n - k - 1 \text{ (full model)}
\]

29. Multiple Partial F-Test for comparing linear regression models

\[
F = \frac{[\text{RSS (full)} - \text{RSS (restricted)}]/q}{\text{MSE (full)}}
df = q, n - k - 1 \text{ (full model)}
\]

30. Mallow’s C_p for comparing linear regression models

\[
C_p = \frac{\text{SSE}(p)}{\text{MSE}(k)} - [n - 2(p + 1)]
\]

31. Variance Inflation Factors (VIF) in linear regression models

\[
\text{VIF} = \frac{1}{[1 - R^2_j]}
\]

32. Reliability of a measure

\[
r_{xx}^* = \frac{x(\text{true score})}{x(\text{true score}) + \nu(\text{error})} = \frac{s^2_x}{s^2_x + \nu}
\]
33. Weighted Least Squares (WLS) equations

\[ \text{SSE} = \sum \left( \frac{1}{s_i^2} \right) (y_i - \hat{y}_i)^2 \]

\[ \hat{\beta}_1 = \frac{\sum \frac{1}{s_i^2} (x_i - \bar{x}')(y_i - \bar{y}')}{\sum \frac{1}{s_i^2} (x_i - \bar{x}')^2} \]

where \( \bar{y}' = \frac{\sum y_i}{\sum s_i^2} \) and \( \bar{x}' = \frac{\sum x_i}{\sum s_i^2} \)

34. Durbin-Watson Statistic to assess Serial Correlation of residuals

\[ d = \frac{\sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n} \hat{e}_i^2} \]

35. Moran’s I for assessing Spatial Autocorrelation

\[ I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \]

36. Hat Matrix for estimating Leverage values

\[ H = X(X'X)^{-1} X' = (h_y) \]
37. Deleted Studentized Residuals
\[ t_i = \frac{e_i}{S_{E(-i)} \sqrt{1 - h_i}} \]

38. Cook’s D for identifying Influential Observations
\[ D_i = \frac{t_i^2}{k + 1} \times \frac{h_i}{1 - h_i} \]

39. DFFITS for identifying Influential Observations
\[ \text{DFFITS}_i = t_i \sqrt{\frac{h_i}{1-h_i}} \]

40. Logistic Function
\[ P(Y = 1) = \frac{1}{1 + \exp[-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)]} \]

41. Odds of an event
\[ \frac{P}{1 - P} \]

42. Odds Ratio (OR)
\[ \frac{P_1}{1 - P_1} / \frac{P_2}{1 - P_2} \]
43. Akaike’s Information Criterion (AIC)

\[ AIC = \frac{\text{Deviance} + 2(k + 1)}{n} \]

44. Bayesian Information Criterion (BIC)

\[ \text{BIC} = \text{Deviance} - \{df_i \times \ln(n)\} \]

45. Normed Sampling Weights

\[ \overline{w}_i = \frac{w_i}{\overline{w}} \]