Applied Linear Regression Notes set 1

Jamie DeCoster

Department of Psychology
University of Alabama
348 Gordon Palmer Hall
Box 870348
Tuscaloosa, AL 35487-0348

Phone: (205) 348-4431
Fax: (205) 348-8648

November 13, 2007

Textbook references refer to Cohen, Cohen, West, & Aiken's (2003) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. I would like to thank Angie Maitner and Anne-Marie Leistico for comments made on earlier versions of these notes. If you wish to cite the contents of this document, the APA reference for them would be:

DeCoster, J. (2007). *Applied Linear Regression Notes set 1.* Retrieved (month, day, and year you downloaded this file, without the parentheses) from http://www.stat-help.com/notes.html

For future versions of these notes or help with data analysis visit
`http://www.stat-help.com`

Contents

Chapter 1

Introduction and Review

1.1 Data, Data Sources, and Data Sets

- Most generally, *data* can be defined as a list of numbers with meaningful relations. We are interested in data because understanding the relations among the numbers can help us understand the relations among the things that the numbers measure.
- The numbers that you collect from an experiment, survey, or archival source is known as a *data source*. Before you can learn anything from a data source, however, you must first translate it into a *data set*. A data set is a representation of a data source, defining a set of "variables" that are measured on a set of "cases."
    - A *variable* is simply a feature of an object that can be categorized or measured by a number. A variable takes on different *values* to reflect the particular nature of the object being observed. The values that a variable takes will change when measurements are made on different objects at different times. A data set will typically contain measurements on several different variables.
    - Each time that we record information about an object we create a *case* in the data set. Cases are also sometimes referred to as *observations*. Like variables, a data set will typically contain multiple cases. The cases should all be derived from observations of the same type of object, with each case representing a different example of that type. The object "type" that defines your cases is called your *unit of analysis*. Sometimes the unit of analysis in a data set will be very small and specific, such as the individual responses on a questionnaire. Sometimes it will be very large, such as companies or nations. The most common unit of analysis in social science research is the participant or subject.
    - Data sets are traditionally organized in a table where each column represents a different variable and each row represents a different case. The value in a particular cell of the table represents the value of the variable corresponding to the column for the case corresponding to the row. For example, if you give a survey to a bunch of different people, you might choose to organize your data set so that each variable represents an item on the survey and each row represents a different person who answered your survey. A cell inside the data set would hold the value that the person represented by the row gave to the item represented by the column.
- The distinction between a data source and a data set is important because you can create a number of different data sets from a single data source by choosing different definitions for your variables or cases. However, all of those data sets would all still reflect the same relations among the numbers found in the data source.
- If the value of a variable has actual numerical meaning (so that it measures the amount of something that a case has), it is called a *continuous* variable. If the value of a variable is used to indicate a group the case is in, it is called a *categorical* variable.

  In terms of the traditional categorizations given to scales, a continuous variable would have either an interval, or ratio scale, while a categorical variable would have a nominal scale. Ordinal scales sort of fall in between. Bentler and Chou (1987) argue that ordinal scales can be reasonably treated as

continuous as long as they have four or more categories. However, if an ordinal variable only has one or two categories you are probably better off either treating it as categorical, or else using procedures specifically designed to handle ordinal data.

- When describing a data set you should name your variables and state whether they are continuous or categorical. For each continuous variable you should state its units of measurement, while for each categorical variable you should state how the different values correspond to different groups. You should also report the unit of analysis of the data set. You typically would not list the specific cases, although you might describe from where the data came.

1.2 Order of operations

- For many mathematical and statistical formulas, you can often get different results depending on the order in which you decide to perform operations inside the formula. Consider equation 1.1.

$$X = 12 + 6 \div 3 \tag{1.1}$$

If we decide to do the addition first ($12 + 6 = 18; 18 \div 3 = 6$) we get a different result than if we decide to do the division first ($6 \div 3 = 2; 12 + 2 = 14$). Mathematicians have therefore developed a specific order in which you are supposed to perform the calculations found in an equation. According to this, you should perform your operations in the following order.

1. Resolve anything within parentheses.
2. Resolve any functions, such as squares or square roots.
3. Resolve any multiplication or division operations.
4. Resolve any addition or subtraction operations.

This would mean that in equation 1.1 above we would perform the division before we would do the addition, giving us a result of 14.

- One of the most notable things is that addition and subtraction are at the bottom of the list. This also includes the application of the summation function ($\sum$). For example, in the equation

$$X = \sum Y_i Z_i \tag{1.2}$$

we must first multiply the values of Y and Z together for each case before we add them up in the summation.

- It is also important to note that terms in the two parts of a fraction are assumed to have parentheses around them. For example, in the equation

$$a = \frac{5 + 1}{7 - 4} \tag{1.3}$$

we must first resolve the numerator ($5 + 1 = 6$) and the denominator ($7 - 4 = 3$) before we do the actual division ($6 \div 3 = 2$).

1.3 General information about statistical inference

- Most of the time that we collect data from a group of subjects, we are interested in making inferences about some larger group of which our subjects were a part. We typically refer to the group we want to generalize our results to as the *theoretical population*. The group of all the people that could potentially be recruited to be in the study is the *accessible population*, which ideally is a representative subset of the theoretical population. The group of people that actually participate in our study is our *sample*, which ideally is a representative subset of the accessible population.

- Very often we will calculate the value of an *estimate* in our sample and use that to draw conclusions about the value of a corresponding *parameter* in the underlying population. Populations are usually too large for us to measure their parameters directly, so we often calculate an estimate from a sample drawn from the population to give us information about the likely value of the parameter. The procedure of generalizing from data collected in a sample to the characteristics of a population is called *statistical inference*.

  For example, let us say that you wanted to know the average height of 5th grade boys. What you might do is take a sample of 30 5th grade boys from a local school and measure their heights. You could use the average height of those 30 boys (a sample estimate) to make a guess about the average height of all 5th grade boys (a population parameter).

- We typically use Greek letters when referring to population parameters and normal letters when referring to sample statistics. For example, the symbol μ is commonly used to represent the mean value of a population, while the symbol $\overline{X}$ is commonly used to represent the mean value of a sample.

- One type of statistical inference you can make is called a *hypothesis test*. A hypothesis test uses the data from a sample to decide between a *null hypothesis* and an *alternative hypothesis* concerning the value of a parameter in the population. The null hypothesis usually makes a specific claim about the parameters (like saying that the average height of 5th grade boys is 60 inches), while the alternative hypothesis says that the null hypothesis is false. Sometimes the alternative hypothesis simply says that the null hypothesis is wrong (like saying that the average height of 5th grade boys is *not* 60 inches). Other times the alternative hypothesis says the null hypothesis is wrong in a particular way (like saying that the average height of 5th grade boys is *less than* 60 inches).

  For example, you might have a null hypothesis stating that the average height of 5th grade boys is equal to the average height of 5th grade girls, and an alternative hypothesis stating that the two heights are not equal. You could represent these hypotheses using the following notation.

  $H_0 : \mu\{\text{boys}\} = \mu\{\text{girls}\}$
  $H_a : \mu\{\text{boys}\} \neq \mu\{\text{girls}\}$

  where H_0 refers to the null hypothesis and H_a refers to the alternative hypothesis.

- To actually perform a hypothesis test, you must collect data from a sample drawn from the population of interest that will allow you to discriminate your hypotheses. For example, if your hypotheses involve the value of a population parameter, then your data should provide a sample estimate corresponding to that parameter. Next you calculate a *test statistic* that reflects how different the data in your sample is from what you would expect if the null hypothesis were true. You then calculate a *p-value* for the statistic, which is the probability that you would get a test statistic as extreme as you did if the null hypothesis was actually true. To get the p-value we must know the distribution from which our test statistic was drawn. Typically this involves knowing the general form of the distribution (t, F, etc.) and the degrees of freedom associated with your test statistic.

  If the p-value is low enough, you reject the null hypothesis in favor of the alternative hypothesis. If the probability is high, you fail to reject the null hypothesis. The breakpoint at which you decide whether to accept or reject the null hypothesis is called the *significance level*, and is often indicated by the symbol α. You typically establish α before you actually calculate the p-value of your statistic. Many fields use a standard $\alpha = .05$.

- When reporting p-values you should use the following guidelines.

  - Use exact p-values (e.g., p = .03) instead of confidence levels (e.g., p < .05) whenever possible.
  - If the p-value is greater than or equal to .10, use two significant digits. For example, p = .12.
  - If the p-value is less than .10, use one significant digit. For example, p = .003.
  - Never report that your p-value is equal to 0. There is always some probability that the results are due to chance alone. Some software packages will say that your p-value is equal to 0 if it is below a specific threshold. In this case you should report that the p-value is less than the threshold. SPSS will report that your p-value is equal to .000 if it is less than .001. If you see this, then you should simply report that p < .001.

- To summarize the steps to a hypothesis test

1. Determine your null and alternative hypotheses.
2. Draw a sample from the population of interest.
3. Collect data that can be used to discriminate the null and alternative hypotheses.
4. Calculate a test statistic based on the data.
5. Determine the probability that you would get the observed statistic if the null hypothesis is true.
6. State whether you reject or fail to reject the null hypothesis.

1.4 Hypothesis tests of a point estimate

- People often test hypotheses that compare a population parameter to a specified, concrete value. For example, you might want to see if the average weight of male students at a certain university is greater than 180 pounds. These are referred to as *tests of a point estimate*. You can use the following formula to test any point estimate that follows a normal distribution.

$$t = \frac{\text{estimate} - \text{null value}}{\text{standard error of estimate}} \tag{1.4}$$

  The test statistic follows a t distribution, although the degrees of freedom will depend on the nature of your estimate.

- When calculating the p-value for your statistic you will need to take into consideration whether H_a is a *one-tailed* or a *two-tailed* hypothesis. A one-tailed hypothesis claims that the value of the population parameter is *either* greater than or less than the null value. A two-tailed hypothesis simply claims that the value of the population is *different* from the null value, so that you can reject the null hypothesis if the observed value is extreme in either direction. This influences the p-value that you calculate for your statistic. If the observed value is in the direction predicted under the alternative hypothesis in a one-tailed test, the observed p-value will be equal to the p-value to test a two-tailed hypothesis divided by two, because you only need to be concerned with errors in a single direction. You automatically fail to reject the null hypothesis whenever the point estimate is on the wrong side of the value under the null in a one-tailed test. Most tables present one-tailed p-values, so you will need to double them when performing a two-tailed test. However, most statistics programs (including SPSS) provide two-tailed p-values, so you will need to divide them by two when performing a one-tailed test.

- When testing a point estimate you should report the following.

  1. The null and alternative hypotheses. The null hypothesis should say that a population parameter is equal to a constant, and the alternative hypothesis should say that the parameter is either greater than, less than, or simply not equal to the same constant.
  2. The estimate calculated from the sample.
  3. The standard error of estimate.
  4. The degrees of freedom.
  5. The value of the test statistic
  6. The p-value for the t statistic.
  7. The conclusion you draw based on the test.

- You also have the option of calculating a *confidence interval* when making inferences about a point estimate. Instead of just determining whether or not the population parameter is equal to a specified value, a confidence interval allows us to determine the entire range of values that could be reasonable values for the parameter, based on our data.

  The first thing you must do to construct a confidence interval is to determine the *confidence level* you want to use. The confidence level is the probability that the true value in the population will actually fall within the bounds of the interval you create. Next you compute the actual interval using the formula

$$CI = \text{estimate} \pm t_{\text{crit}} * \text{standard errror of estimate}, \tag{1.5}$$

where t_{crit} is equal to the value from the t distribution that corresponds to a p-value of (1 - your confidence level) ÷2.

- Confidence intervals are typically preferred to hypothesis tests because they indicate exactly what parameter values are consistent and which are inconsistent with the observed data. In fact, a confidence interval can always double as a hypothesis test. You would reject the null hypothesis if the value under the null is outside of the confidence interval, whereas you would fail to reject the null hypothesis if the value is inside the confidence interval.

1.5 General information about regression

- Regression is a statistical tool that allows you to predict the value of one continuous variable from one or more other variables.
- The variable that we are trying to predict is called the *dependent variable* (DV) while the variables on which we are basing our prediction are called the *independent variables* (IVs).
- When you perform a regression analysis, you create a *regression equation* that predicts the values of your DV using the values of your IVs. Each IV is associated with specific coefficients in the equation that summarizes the relation between that IV and the DV.
- Once we estimate a set of coefficients in a regression equation, we can use hypothesis tests and confidence intervals to make inferences about the corresponding parameters in the population. You can also use the regression equation to predict the value of the DV for a new observation based on its values on the IVs.
- We typically think of using regression when we have continuous IVs, while we think of using *analysis of variance* (ANOVA) when we have categorical IVs. However, both regression and ANOVA are actually based on the same set of statistical principles, called the *general linear model* (GLM). You can therefore use regression to examine the ability of both continuous and categorical IVs to predict the values of a continuous DV.

1.6 Working with SPSS

- Most users typically open up an SPSS data file in the data editor, and then select items from the menus to manipulate the data or to perform statistical analyses. This is referred to as the *interactive mode*, because your relation with the program is very much like a personal interaction, with the program providing a response each time you make a selection. If you request a transformation, the data set is immediately updated. If you select an analysis, the results immediately appear in the output window.
- It is also possible to work with SPSS in *syntax mode*, where the user types code in a syntax window. Once the full program is written, it is then submitted to SPSS to get the results. Working with syntax is more difficult than working with the menus because you must learn how to write the programming code to produce the data transformations and analyses that you want. However, certain procedures are only available through the use of syntax. You can also save the programs you write in syntax. This can be very useful if you expect to perform the same or similar analyses multiple times, since you can just reload your old program and run it on your new data (or your old data if you want to recheck your old analyses). If you would like more general information about writing SPSS syntax, you should examine the *SPSS Base Syntax Reference Guide*. Assuming that the guide is installed with your version of SPSS, you can access it by selecting **Help → Syntax Guide → Base**.
- Whether you should work in interactive or syntax mode depends on several things. Interactive mode is easier and generally quicker if you only need to perform a few simple procedures. You should therefore probably work interactively unless you have a specific reason to use syntax. Some reasons to choose syntax would be:
    - You need to use options or procedures that are not available using interactive mode.
    - You expect that you will perform the same manipulations or analyses on several different data sets and want to save a copy of the program code so that it can easily be re-run.

- You are performing a very complicated set of manipulations or analyses, such that it would be useful to document all of the steps leading to your results.

- There are several ways for people who use the interactive mode to get the exact syntax that is associated with their menu selections. This enables people who are not familiar with the details of SPSS syntax to take advantage of some of its benefits.
  - Whenever you make selections in interactive mode, SPSS actually writes down syntax code reflecting the menu choices you made in a "journal file." The name of this file can be found (or changed) by selecting **Edit** → **Options** and then selecting the **General** tab. If you ever want to see or use the code in the journal file, you can edit the journal file in a syntax window.
  - SPSS also provides an easy way to see the code corresponding to a particular menu function. Most selections include a **Paste** button that will open up a syntax window containing the code for the function, including the details required for any specific options that you have chosen in the menus.
  - You can have SPSS include the corresponding syntax in the output whenever it runs a statistical analysis. To enable this
    - Choose **Edit** → **Options**.
    - Select the **Viewer** tab.
    - Check the box next to **Display commands in the log**.
    - Click the **OK** button.

- SPSS was originally designed to work from an ANOVA rather than a regression viewpoint, so it uses some unusual conventions when talking about variables. Specifically, it refers to categorical variables as *factors* and continuous variables as *covariates* in many of its menus. This can be confusing because traditionally, covariates are defined as continuous variables that you include in your analyses only to control for their effects on the IVs of interest. However, SPSS still allows you to perform all the analyses you might want on continuous variables in these cases, so you can just ignore the odd name.

1.7 Working with Microsoft Excel

- There are several reasons that you might want to take information from a Microsoft Excel spreadsheet and bring it into SPSS for analysis.
  - You might choose to enter the data from a study you conduct into Excel instead of directly into SPSS. Excel is designed to store information in a tabular format, which is perfect for creating a data set for analysis. It also contains a number of functions to make data entry easier and more accurate.
  - Excel is a widely available program, so if you obtain a data set from a nonacademic source (such as a government archive or the Internet) there is a decent chance that it will be stored as an Excel file.

- It is actually very easy to import a data set from Excel into SPSS if you first save a copy of your spreadsheet in Excel version 4.0. This is the highest version of Excel that does not allow you to store several different spreadsheets in the same file. It is considerably more difficult to read data in from the higher versions, since you will need to specify the worksheet and the cells you want to import.

- Within Excel, the following steps allow you to save a copy of the current spreadsheet in Excel version 4.0.
  - Choose **File** → **Save as**.
  - Click the button on the right-hand side of the **Save as type** box.
  - Choose **Microsoft Excel 4.0 Worksheet**.
  - Click the **Save** button.

- Within SPSS, the following steps allow you to load an Excel version 4.0 worksheet.

- Choose **File → Open → Data**.
- Click the button on the right-hand side of the **Files of type** box
- Choose **Excel**.
- Either select the file you want to open using the explorer window or type the name directly into the **File name** box.
- Click the **Open** button.
- Click the check box next to **Read variable names** if the first row of your spreadsheet contains the variable names.
- Click the **OK** button.

- If you have a data set stored as a word-processing or text file, one of easiest ways to get it into SPSS is to first import the data into Excel, and then transport the Excel data set to SPSS.

  In order for you to read in data from a text file, you must organize the document so that Excel knows what values should be placed in what columns in the spreadsheet. There are two different ways to do this. In a *delimited* file, data that will go into different columns in the spreadsheet are separated by a specific character, such as a tab, space, or comma. You must be careful to choose an appropriate delimiter. Specifically, your delimiter should not be a character that would actually appear in any of the entries, or else Excel will not be able to tell when the character should be considered part of a cell value and when the character should be treated as a delimiter. Usually the tab character is a good choice.

  In a *fixed-width* file, data that will go into different columns in the spreadsheet are placed in specific columns in the text file. You should be careful when entering in data in a fixed-width format, because many fonts don't make all of the letters the same width. In this case you could have two words that appear to be in the same location but which are actually in different columns. It is therefore always a good idea to use *nonproportional font* (such as Courier) when working with a fixed-width file in a word processor so that you can see how things actually line up.

  Excel can read both delimited and fixed-width files. Each provides you with a clear way of knowing what characters from the text file should be placed into each column in the spreadsheet. However, most people find delimited files easier to work with than fixed-width files.

  If you are working with a word-processing document you must save it as a text file before you can try to load it into Excel. A text file contains all the words and numbers from the document, but does not have any special formatting. To save a Microsoft Word document as a text file

  - Choose **File → Save as**
  - Click the button on the right-hand side of the **Save as type** box.
  - Choose **Text Only**.
  - Click the **Save** button.

- To read a text file into Excel

  - Choose **File → Open**.
  - Click the button on the right-hand side of the **Files of type** box.
  - Choose **Text Files**.
  - Either select the file you want to open using the explorer window or type the name directly into the **File name** box.
  - Click the **Open** button.
  - Click the radio button next to either **Delimited** or **Fixed Width** to indicate how the text file is formatted.
  - Click the **Next** button.
  - *For delimited data files*
    - At the top of the next window, Excel will ask you to indicate what characters you used as delimiters. Make sure that the boxes next to your delimiters are checked and those that are not your delimiters are unchecked.

        - Excel will now show you how it thinks you want to divide the contents of the file into different spreadsheet cells. This should be correct if you applied the formatting properly.
        - Click the **Next** button.
    - *For fixed-width data files*
        - Excel will now show you how it thinks you want to divide the contents of the file into different spreadsheet cells. Excel does its best to guess where the divisions should be, but you may need to make changes to import the file properly. You can move existing divisions by clicking on the dividing line and dragging it to a new location. You can completely remove a divider by dragging its line outside the box. You can add a new division by clicking on a location where there currently is no line.
        - Click the **Next** button.
    - Excel now provides you with the opportunity to define the variable type for each spreadsheet column. By default, Excel assumes that each column will be of the **General** type, which automatically converts cells with numbers to numeric format and cells with letters to text format. There is usually no need to change this.
    - Click the **Finish** button. You will now have a version of the data set loaded into Excel. If you want to bring the file into SPSS you should then save it as an Excel 4.0 file and import it using the procedures described above.

- If you ever have an SPSS data set that you want to save as an Excel file, you can do so by taking the following steps.
    - Open the data set you want to convert
    - Choose **File** → **Save as**.
    - Click the button on the right-hand side of the **Save as type** box.
    - Choose **Excel**.
    - Click the **Save** button.

Chapter 2

Bivariate Correlation and Regression

2.1 Bivariate relations

- Two variables are said to have a *relation* if knowing the value of one variable gives you information about the likely value of the second variable.
- There are several different ways to categorize relations. One way is based on the precision of the relation.
    - *Functional relations* are those where knowing the value of the one variable tells you exactly what the value of the second variable will be. For example, if I know your height in inches (variable 1) I can precisely calculate what your height will be in feet (variable 2).
    - *Statistical relations* are those where knowing the value of one variable can only be used to approximate the value of the second variable. For example, if I know your height in inches (variable 1) I can probably make a pretty good guess about your weight in pounds (variable 2). However, since different people have different builds, there will be error associated with my approximation. The essence of statistical inference is prediction with error.

  In the social sciences we are primarily interested in discovering statistical relations.
- We can also categorize relations as to whether or not they imply that changes in one variable actually cause changes in the second variable.
    - *Causal relations* are those where we believe that one of the variables directly influences the value of the other variable. In a causal relation, the IV is often referred to as the *causal variable* and the DV is referred to as the *effect variable*. The presence of a causal relation implies that if we change the value of the causal variable, we would expect to see changes in the value of the effect variable. For example, if I were to change the amount of time that I gave you to study for an exam (variable 1), I would expect to find a change in your exam performance (variable 2).
    - *Noncausal* or *correlational relations* are those where there is no expectation that either one of the variables directly influences each other. This would imply that changing the value of the one variable shouldn't have any direct influence on the value of the second variable. For example, the total number of books you have in your house could be related to a person's performance on an exam, since people who are more intelligent tend to read more. However, we would not expect that taking books out of someone's home would directly influence their intelligence or their performance on an exam.
- Identifying that there is any type of relation between two variables allows you to predict the value of one variable from the value of the other variable. It can also help you understand how and why the values of the variables change across cases or over time.
- One of the easiest ways to see if there is a relation between a pair of variables is to create a *scatterplot*. A scatterplot is a graph where you place the values of one variable on the horizontal axis and the values of the second variable on the vertical axis. You then plot a point for each case in your data set at the location corresponding to its values on the two variables. You have a relation between your

two variables if you can see a reliable pattern in the scatterplot. Most commonly researchers look for patterns that follow a straight line, but any pattern that allows you to predict the value of one variable from the other indicates a relation between the variables.

To create a scatterplot in SPSS

- Choose **Graphs** → **Scatter** → **Simple**.
- Click the **Define** button.
- SPSS will now display a variable selection screen. You should move the variable you want on the X-axis into the box labeled **X-axis** and the variable you want on the Y-axis to the box labeled **Y-axis**.
- Click the **OK** button.

2.2 The correlation coefficient

- While a scatterplot is a good tool for determining whether there is a relation between a pair of variables, it does not provide a quantitative measure of the strength of that relation.
- The simplest measure of the strength of the relation between two continuous variables is the *Pearson correlation coefficient*. The correlation coefficient is typically denoted using the symbol r, and provides an estimate of how well a straight line would fit the graph inside of a scatterplot between the two variables. The value of the correlation provides information about both the nature and the strength of the relation. To interpret a correlation you should consider the following.
  - Correlations range between -1.0 and 1.0.
  - The sign of the correlation describes the direction of the relation. A positive sign indicates that as one variable gets larger the other also tends to get larger, while a negative sign indicates that as one variable gets larger the other tends to get smaller.
  - The magnitude of the correlation describes the strength of the relation. The further that a correlation is from zero, the stronger the relation is between the two variables. A zero correlation would indicate that the two variables aren't related to each other at all.
- Note that correlations only measure the strength of the *linear* relation between the two variables. Sometimes you have a relation that would be better measured by a curve of some sort rather than a straight line. In this case the correlation coefficient would not provide a very accurate measure of the strength of the relation.

  If the relation between your two variables is accurately described by a line, your ability to predict the value of one variable from the value of the other is directly related to the correlation between them. When the points in your scatterplot are all clustered closely about a line your correlation will be large and the accuracy of the predictions will be high. If the points tend to be widely spread your correlation will be small and the accuracy of your predictions will be low.
- Cohen (1992) established a set of guidelines for interpreting the strength of correlations. He claimed that a correlation of .1 is a "small" effect, a correlation of .3 is a "medium" effect, and that a correlation of .5 is a "large" effect. Cohen established the medium effect size to be one that was large enough so that people would naturally recognize it in everyday life, the small effect size to be one that was noticeably smaller but not trivial, and the large effect size to be the same distance above the medium effect size as small was below it.
- You should be very cautious about interpreting correlations between series that increment over time. Statisticians have noted that any two "random walks" (a process where the value at one time point is equal to the value at the last time point plus a random change in one direction or another) are bound to have a relatively high correlation even when they are not affected by any of the same processes (Granger & Newbold, 1974). A classic example is that during the 18th century, there was a very strong correlation between the number of ministers in Boston and the amount of rum sold in Jamaica. The only true link between them is that both of these values changed over time. The appropriate way to try to test for relations in this case is to use a procedure called *cointegration* (Enders, 2004).

- There are a number of different formulas available to calculate the correlation coefficient. We will consider three different formulas, each serving a different purpose.
  - One way to calculate a correlation is by using the formula

$$r_{xy} = \frac{\sum Z\{X_i\}Z\{Y_i\}}{n-1} \tag{2.1}$$

    where r_{xy} stands for the correlation between variable X and variable Y, $Z\{X_i\}$ refers to the standardized score of X_i, $Z\{Y_i\}$ refers to the standardized score of Y_i, and n is the number of cases in the data set.

    Equation 2.1 is called the *definitional formula* because it can be easily used to gain a conceptual understanding of the correlation coefficient. Let's consider how the data from a single case affects the calculation of the correlation. If a case has values that are either both above the mean or both below the mean, the product of the two Z scores will be positive, making the resulting correlation more positive. If a case has a value above the mean on one variable and a value below the mean on the other variable, the product of the two Z scores will be negative, making the resulting correlation more negative. We can see that if two variables consistently have standardized scores in the same direction, they will end up with a positive correlation. If two variables consistently have standardized scores in opposite directions, they will end up with a negative correlation. If two variables sometimes have scores in the same direction and sometimes have scores in the opposite direction, the positives will cancel with the negatives and they will end up with a correlation near zero.

    This equation also lets us see that the correlation between a pair of variables will be independent of the actual scales in which those two variables are measured. Since the correlation is purely a function of the standardized scores, changing the scale on which a variable is measured will actually have no effect on the correlations between that variable and anything else. An additional bonus of this is that you can make direct comparisons between different correlations, even when those correlations were computed using entirely different variables.
  - Another formula for the correlation is

$$r_{xy} = \frac{s_{xy}}{s_x s_y}, \tag{2.2}$$

    where s_{xy} is the covariance between X and Y, s_x is the standard deviation of X, and s_y is the standard deviation of Y.

    Equation 2.2 is called the *covariance formula* because it expresses the correlation in terms of the variability in X and Y. From this formula we can see that the correlation can be thought of as the degree to which X and Y vary together relative to the degree to which they vary separately.

    Specifically, the square of the correlation (r^2) is called the *coefficient of determination* and is equal to the proportion of the variability in one variable that can be explained by the other. When comparing the strengths of different relations you should therefore focus on r^2 rather than r. For example, a pair of variables with a correlation of .4 ($r^2 = .16$) actually have four times the amount of shared variability compared to a pair of variables with a correlation of .2 ($r^2 = .04$).
  - The final formula for the correlation that we will consider is

$$r_{xy} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum {X_i}^2 - \left(\sum X_i\right)^2}\sqrt{n\sum {Y_i}^2 - \left(\sum Y_i\right)^2}}, \tag{2.3}$$

    Equation 2.3 is called the *computational formula*, because it is the equation from which it is easiest to calculate the correlation by hand. However, it does not provide much insight into the meaning of the correlation.

- If you want to compare an observed correlation to a specified value you can do so using a hypothesis test of a point estimate. However, you have to take some special steps because the higher a correlation is, the more difficult it is to increase it. The difference between correlations of .80 and .85 is actually considerably larger than the difference between correlations of .10 and .15 (remember the rule about focusing on r^2 instead of r). This means that the distribution is not normal, so we can't use the standard test of a point estimate. Luckily, Fisher (1928) developed a way to transform correlations to

Z-scores so that they have a normal distribution, allowing us to perform the standard test of a point estimate on the transformed value.

To transform correlations to Z-scores you use the r-to-Z formula

$$Z_r = \frac{\ln(1+r) - \ln(1-r)}{2}, \tag{2.4}$$

where ln() stands for the natural log function. The standard error of this Z-score can be calculated using the formula

$$s\{Z_r\} = \frac{1}{\sqrt{n-3}}. \tag{2.5}$$

To determine if the correlation is different from a specified null value you must first transform both the observed correlation and the null value to Z scores using formula 2.4. You can then perform a hypothesis test of a point estimate with the following characteristics.

- $H_0 : Z_\rho = Z_\rho\{\text{null}\}$
  $H_a : Z_\rho \neq Z_\rho\{\text{null}\}$
- Estimate = $Z_r\{\text{observed}\}$ = Z transformation of the observed correlation
- Standard error of estimate = $s\{Z_r\} = \frac{1}{\sqrt{n-3}}$
- $Z = \frac{Z_r\{\text{observed}\} - Z_\rho\{\text{null}\}}{s\{Z_r\}}$

Note that in this circumstance, the test produces a Z statistic instead of the standard t statistic. We therefore do not need to worry about degrees of freedom when testing the converted Z-scores. You can just take the p-value directly from the standard normal distribution.

- Fisher's r-to-Z formula can also be used if you want to average a set of correlations. In this case the appropriate procedure would be to transform each of your correlations to Zr using equation 2.4, compute the mean Zr score, and then transform the mean Zr back to a correlation using equation 2.6.
- There is a simpler test you can use when the expected value under the null hypothesis is zero. In this case you actually do not need to transform your correlation at all. To test whether a correlation is equal to zero you would perform a hypothesis test of a point estimate with the following characteristics.
  - $H_0 : \rho = 0$
    $H_a : \rho \neq 0$
  - Estimate = r = sample correlation
  - Standard error of estimate = $s_r = \sqrt{\frac{1-r^2}{n-2}}$
  - Degrees of freedom = $n - 2$
  - $t = \frac{r-0}{s_r} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

  This is a commonly performed test because it lets you determine if there is a significant relation between your two variables.
- To calculate a confidence interval around a correlation you must first transform your correlation to a Z-score using equation 2.4. Next you calculate the standard error of the transformed score using equation 2.5. Then you use the Z-score, its standard error, and the value of the standard normal distribution associated with your desired confidence level to create an interval around Z_r (you use the value from the standard normal distribution in the place where you normally use the value from the t distribution). Finally, you translate the boundaries of your confidence interval around Z_r back into correlations using Fisher's (1928) Z-to-r formula

$$r = \frac{e^{2Z_r} - 1}{e^{2Z_r} + 1}, \tag{2.6}$$

where e is the base of the natural logarithm (approximately 2.71828). You would then report these final correlations as your confidence interval.

- To obtain the correlation between two variables in SPSS
  - Choose **Analyze** → **Correlate** → **Bivariate**.
  - Select the variables you want to correlate and move them into the box labeled **variables**.
  - Check the box next to **Pearson**.
  - Click the **OK** button.

  If you have more than two continuous variables and would like to obtain the simple Pearson correlation between all pairs of the variables, perform the analysis described above but move all of the variables of interest to the **Variables** box.

- The SPSS output from a correlation analysis only includes a single section.
  - **Correlations**. This section contains a *correlation matrix* for all of the variables you selected. Each variable is represented in both a row and a column in the table. The values inside a given cell of the table contains the correlation between the variable at the top of its column with the variable to the left of its row. The table also provides the p-value for the correlation and the sample size on which the calculation is based.

- If you want to compare the correlations found in two independent samples you will need to first transform both correlations into Z scores using equation 2.4. You can then perform a hypothesis test of a point estimate with the following characteristics.
  - $H_0 : Z_{\rho 1} - Z_{\rho 2} = 0$
    $H_a : Z_{\rho 1} - Z_{\rho 2} \neq 0$
  - Estimate $= Z_{r1} - Z_{r2}$
  - Standard error of estimate $= s\{\text{zdiff } 1\} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$
  - $Z = \frac{Z_{r1} - Z_{r2}}{s\{\text{zdiff } 1\}}$

  Again notice that this test produces a Z statistic instead of the standard t statistic.

- You can also calculate a confidence interval around the difference between the correlations obtained from two different samples. You must first transform both correlations to Z-scores using equation 2.4. Next you calculate the standard error of the difference score $s\{\text{zdiff } 1\}$ using the formula provided above. Then you use the difference between the Z-scores, the standard error of the difference, and the value of the standard normal distribution associated with your error level to create an interval around the difference (you use the value from the standard normal distribution in the place where you normally use the value from the t distribution). Finally, you translate the boundaries of your confidence interval back into correlations using equation 2.6. You then report these final correlations as your confidence interval.

- The above procedure will produce a confidence interval that is not centered on the actual difference score. This is perfectly appropriate because correlations themselves have an asymmetrical distribution. There is, however, a method to obtain a symmetric interval. Olkin and Finn (1995) propose that you take the actual difference between the correlations as your estimate, and then calculate the standard error of this difference using the formula

$$s_{\text{rdiff}} = \sqrt{\frac{1 - {r_1}^2}{n_1} + \frac{1 + {r_2}^2}{n_2}}. \tag{2.7}$$

  Then you use the difference between the correlations, the standard error of the difference, and the value of the standard normal distribution associated with your error level to create an interval around the difference (you use the value from the standard normal distribution in the place where you normally use the value from the t distribution).

- The formulas we have discussed so far only work if you are examining correlations taken from independent samples. Steiger (1980) provides several methods of comparing correlations measured in the same sample. To compare the correlation between one pair of variables with the correlation between two other variables (so that neither of the variables in the first correlation are the same as those in the second correlation), you would first transform both correlations into Z scores using equation 2.4. You can then perform a hypothesis test of a point estimate with the following characteristics.
  - $H_0 : Z_{\rho\{jk\}} - Z_{\rho\{hm\}} = 0$
    $H_a : Z_{\rho\{jk\}} - Z_{\rho\{hm\}} \neq 0$
  - Estimate $= Z_{r\{jk\}} - Z_{r\{hm\}}$
  - Standard error of estimate $= s\{\text{zdiff } 2\} = \sqrt{\frac{2-2(\text{COV}_{jk,hm})}{n-3}}$
  - $Z = \frac{Z_{r\{jk\}} - Z_{r\{hm\}}}{s\{\text{zdiff } 2\}}$

  In the equation for the standard error, $\text{cov}_{jk,hm}$ refers to the covariance between the two correlations, which can be computed using the formula

$$\text{cov}_{jk,hm} = \frac{\begin{array}{c}(r_{jh} - r_{jk}r_{kh})(r_{km} - r_{kh}r_{hm}) + (r_{jm} - r_{jh}r_{hm})(r_{kh} - r_{kj}r_{jh}) \\ +(r_{jh} - r_{jm}r_{mh})(r_{km} - r_{kj}r_{jm}) + (r_{jm} - r_{jk}r_{km})(r_{kh} - r_{km}r_{mh})\end{array}}{2(1 - r_{jk}^2)(1 - r_{jh}^2)}. \tag{2.8}$$

  To compare how well one variable correlates with two other variables (so one of the variables in the first correlation is also involved in the second correlation), you would first transform both correlations into Z scores using equation 2.4. You can then perform a hypothesis test of a point estimate with the following characteristics.

  - $H_0 : Z_{\rho\{jk\}} - Z_{\rho\{jh\}} = 0$
    $H_a : Z_{\rho\{jk\}} - Z_{\rho\{jh\}} \neq 0$
  - Estimate $= Z_{r\{jk\}} - Z_{r\{jh\}}$
  - Standard error of estimate $= s\{\text{zdiff } 3\} = \sqrt{\frac{2-2(\text{COV}_{jk,jh})}{n-3}}$
  - $Z = \frac{Z_{r\{jk\}} - Z_{r\{jh\}}}{s\{\text{zdiff } 3\}}$

  In the equation for the standard error, $\text{cov}_{jk,jh}$ refers to the covariance between the two correlations, which can be computed using the formula

$$\text{cov}_{jk,jh} = \frac{r_{kh}(1 - r_{jk}^2 - r_{jh}^2) - \frac{1}{2}(r_{jk}r_{jh})(1 - r_{jk}^2 - r_{jh}^2 - r_{kh}^2)}{(1 - r_{jk}^2)(1 - r_{jh}^2)}. \tag{2.9}$$

- You can also compute a confidence interval around the differences between two correlations measured in the same sample. You would first transform both correlations into Z scores using equation 2.4. Then you use the difference between the Z-scores, the standard error of the difference (using the appropriate formula), and the value of the standard normal distribution associated with your error level to create an interval around the difference. Finally, you translate the boundaries of your confidence interval back into correlations using equation 2.6.

2.3 Simple linear regression

- The correlation coefficient is just one way to examine the relation between a pair of variables. *Simple linear regression* provides a second way, in which we try to predict the values of one variable (Y) from the value of the other variable (X) using the equation

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i. \tag{2.10}$$

In this equation, Y_i and X_i are the measured values of our variables, β_0 and β_1 are parameters that summarize the linear relation between our variables, and ϵ_i represents the difference between the predicted and the actual values (i.e., the error in prediction). The values of ϵ_i are assumed to have a normal distribution with a mean of 0 and a variance of σ^2.

- Predicting the values of Y from the values of X is referred to as *regressing Y on X*. When analyzing data from a study you will typically want to regress the values of the DV on the values of the IV. This makes sense since you want to use the IV to explain variability in the DV.
- Creating an equation to predict the value of a variable is also called *building a statistical model.* In simple linear regression you would say that you are trying to model the DV using the IV.
- You may remember from geometry that equation 2.10 is equivalent to a straight line. This is no accident, since the purpose of simple linear regression is to define the line that represents the relation between our two variables. β_0 is the intercept of the line, indicating the expected value of Y when $X = 0$. β_1 is the slope of the line, indicating how much we expect Y will change when we increase X by a single unit. Using this line, we can input different values of X and find out what value of Y we should be most likely to see.
- Notice that equation 2.10 is written in terms of population parameters. That indicates that our goal is to determine the relation between the two variables in the population as a whole. We typically do this by taking a sample and then performing calculations to obtain the estimated regression equation

$$Y_i = b_0 + b_1 X_i \tag{2.11}$$

that best fits the observed data.

- We typically calculate b_0 and b_1 using the methods of *least squares*. This chooses estimates that minimize the sum of squared errors between the values of the estimated regression line and the actual observed values. The least squares estimate of the slope can be calculated as

$$b_1 = r_{xy}\left(\frac{s_y}{s_x}\right) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum {X_i}^2 - \frac{(\sum X_i)^2}{n}}, \tag{2.12}$$

with the variance

$$s^2\{b_1\} = \left(\frac{s_y^2}{s_x^2}\right)\left(\frac{1 - r_{xy}^2}{n-2}\right). \tag{2.13}$$

The estimate of the intercept can be calculated as

$$b_0 = \bar{Y} - b_1\bar{X} = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right). \tag{2.14}$$

with the variance

$$s^2\{b_0\} = \text{MSE}\left[\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}\right], \tag{2.15}$$

where MSE is the *mean squared error*, defined below in formula 2.18.

- When performing linear regression, we typically make the following assumptions about the error terms ϵ_i.
  - The errors have a normal distribution.
  - The same amount of error in the model is found at each level of X.
  - The errors in the model are all independent.
- Under the conditions of the regression model given above, the least squares estimates are unbiased and have minimum variance among all unbiased linear estimators. This means that the estimates get us as close to the true unknown parameter values as we can get.

- The least squares regression line always passes through the point $(\bar{X}, \bar{Y})$.
- We can use our estimated regression line to predict values of Y from the values of X. We can obtain a *predicted value* for each case in our data set using the formula

$$\hat{Y}_i = b_0 + b_1 X_i. \tag{2.16}$$

- The difference between the predicted value and the value that is actually observed is called the *residual*, calculated using the formula

$$e_i = Y_i - \hat{Y}_i. \tag{2.17}$$

  The residuals will always have a mean of zero since they are centered around the regression line. The variance of the residuals is called the *mean squared error* (MSE) and can be calculated using the formula

$$\text{MSE} = s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}. \tag{2.18}$$

  In addition to using equations 2.13 and 2.15 above, you can also calculate variances of b_1 and b_0 directly from the MSE using the formulas

$$s^2\{b_1\} = \frac{\text{MSE}}{\sum (X_i - \bar{X})^2} \tag{2.19}$$

  and

$$s^2\{b_0\} = \text{MSE}\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right). \tag{2.20}$$

- If the slope of your regression line is small, then we know that changes in the X variable have very little influence on the Y variable. If the slope is large, however, then we can expect that the values of Y will change with even a small change in X. In fact, the value of the slope can be used to measure the strength of the relation between X and Y. Researchers will therefore often test hypotheses about b_1.

  If we want to compare b_1 to a specific value, we can perform a hypothesis test of a point estimate with the following characteristics.

  - $H_0 : \beta_1 = \beta_{\text{null}}$
    $H_a : \beta_1 \neq \beta_{\text{null}}$
  - Estimate $= b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$
  - Standard error of estimate $= s\{b_1\} = \frac{s_y}{s_x}\sqrt{\frac{1 - r_{xy}^2}{n-2}}$
  - Degrees of freedom $= n - 2$
  - $t = \frac{b_1 - \beta_{\text{null}}}{s\{b_1\}}$

- The intercept from the regression model is not often used in statistical inference because it just tells you the expected value of Y when $X = 0$. However, you could examine it using a test of a point estimate with the following characteristics.

  - $H_0 : \beta_0 = \beta_{\text{null}}$
    $H_a : \beta_0 \neq \beta_{\text{null}}$
  - Estimate $= b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right)$
  - Standard error of estimate $= s\{b_0\} = \sqrt{\text{MSE}\left[\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_x^2}\right]}$

- Degrees of freedom $= n - 2$
- $t = \frac{b_0 - \beta_{\text{null}}}{s\{b_0\}}$

- You can calculate a confidence interval around β_0 or β_1 using the estimate, its standard error, and the value of the t distribution with $n - 2$ degrees of freedom that is associated a p-value equal to (1 - your desired confidence level) $\div$ 2.

- To perform simple linear regression in SPSS
  - Choose **Analyze** $\rightarrow$ **Regression** $\rightarrow$ **Linear**.
  - Place the DV (Y) in the **Dependent** box.
  - Place the IV (X) in the **Independent(s)** box.
  - Click the **OK** button.

- The SPSS output from a simple linear regression analysis contains the following sections.
  - **Variables Entered/Removed**. This section is only used in model building (discussed in Chapter 5) and contains no useful information in simple linear regression.
  - **Model Summary**. The value listed below **R** is the correlation between your variables. The value listed below **R Square** is the proportion of variance in your DV that can be accounted for by your IV. The value in the **Adjusted R Square** column is a measure of model fit, adjusting for the number of IVs in the model. The value listed below **Std. Error of the Estimate** is the standard deviation of the residuals.
  - **ANOVA**. Here you will see an ANOVA table, which provides an F test of the relation between your IV and your DV. If the F test is significant, it indicates that there is a relation.
  - **Coefficients**. This section contains a table where each row corresponds to a single coefficient in your model. The row labeled **Constant** refers to the intercept, while the row containing the name of your IV refers to the slope. Inside the table, the column labeled **B** contains the estimates of the parameters and the column labeled **Std. Error** contains the standard error of those parameters. The column labeled **t** contains the value of the t-statistic testing whether the value of each parameter is equal to zero. The p-value of this test is found in the column labeled **Sig**. If the test for the IV is significant, then there is a relation between the IV and the DV. Note that the square of the t statistic is equal to the F statistic in the ANOVA table and that the p-values of the two tests are equal. This is because both of these are testing whether there is a significant linear relation between your variables.

    The column labeled **Beta** contains the *standardized regression coefficient*, which is the parameter estimate that you would get if you standardized both the IV and the DV by subtracting off their mean and dividing by their standard deviations. Standardized regression coefficients are sometimes used in multiple regression (discussed in Chapter 3) to compare the relative importance of different IVs when predicting the DV. In simple linear regression, the standardized regression coefficient will always be equal to the correlation between the IV and the DV.

- Pages 46-47 of the textbook presents a method of directly comparing the regression coefficients you get when you test the same regression model on two independent groups. However, the method they present has a problem: It is unclear exactly what distribution the difference actually follows. It is actually much better to test for a interaction between the X variable and a categorical variable representing the different groups, which is discussed in Chapter 9.

- Once we have calculated a regression equation we can predict the value of Y for a new observation at X_h using the formula

$$\hat{Y}_h = b_0 + b_1 X_h. \tag{2.21}$$

  You must be careful when predicting new values if X_h is outside the range of X's used to build the regression equation. This is called extrapolation, and can lead to incorrect prediction.

- We can make inferences about the value of $\hat{Y}_h$. To compare the mean expected value of $\hat{Y}_h$ to a specific constant, you can perform a hypothesis test of a point estimate with the following characteristics.

- $H_0 : Y_h = Y_0$
  $H_a : Y_h \neq Y_0$
- Estimate $= \hat{Y}_h = b_0 + b_1 X_h$
- Standard error of estimate $= s\{\text{mean } \hat{Y}_h\} = \sqrt{\text{MSE}\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}$
- Degrees of freedom $= n - 2$
- $t = \frac{\hat{Y}_h - Y_0}{s\{\text{mean } \hat{Y}_h\}}$

- You can actually calculate two different types of intervals around $\hat{Y}_h$. First, you can calculate a confidence interval around the expected value of Y using the estimate and standard error provided above, along with the value of the t distribution with $n - 2$ degrees of freedom that is associated with a p-value equal to (1 - your desired confidence level) ÷ 2.

  The confidence interval provides information about the population mean of all new values. Specifically, your confidence level is the probability that the population mean of all new values fall within your confidence interval. You can also calculate a *prediction interval* that would encompass a specified percent of all the new values. The boundaries of the prediction interval would be

$$PI = \hat{Y}_h \pm t_{\text{crit}} * s\{\text{pred}\}, \tag{2.22}$$

  where t_{crit} is the value from the t distribution with $n - 2$ degrees of freedom associated with your confidence level and $s\{\text{pred}\}$ is calculated using the formula

$$s\{\text{pred}\} = \sqrt{\text{MSE}\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)}. \tag{2.23}$$

  In the case of a prediction interval, the confidence level is the percent of all new observations that you would expect to fall within the interval. The confidence interval tells you how sure you are about the true mean value of Y_h, while the prediction interval tells you how much variability in Y_h you can expect to see when you look at different cases that have the value X_h.

- SPSS can provide you with confidence intervals and prediction intervals around each of the observations in your data set if you take the following steps.
  - Choose **Analyze** → **Regression** → **Linear**.
  - Define your regression model as described above, but do not click the **OK** button.
  - Click the **Save** button.
  - In the **Prediction Intervals** section, check the box next to **Mean** if you want a confidence interval, and check the box next to **Individual** if you want a prediction interval.
  - Enter your confidence level in the **Confidence Interval** box.
  - Click the **Continue** button.
  - Click the **OK** button.

- After performing this analysis, SPSS will add variables to your data set that contain the lower and upper bounds of the intervals you selected for each of your observations. If you asked for a confidence interval, its lower bound will be contained in a variable starting with **lmci** and its upper bound will be contained in a variable starting with **umci**. If you asked for a prediction interval, its lower bound will be contained in a variable starting with **lici** and its upper bound will be contained in a variable starting with **uici**.

  You can actually use this same procedure to obtain confidence and prediction intervals around new observations. All that you need to do is create a new case in the data set that has value of X_h as your IV and nothing as your DV. This case will not influence the estimated regression equation, but SPSS will still provide the appropriate values for the confidence and prediction intervals in the variables listed above.

- There are several strong parallels between correlation and simple linear regression. In fact, you can directly calculate the slope parameter from the correlation coefficient using the formula

$$b_1 = r\left(\frac{s_y}{s_x}\right). \tag{2.24}$$

  Just like the correlation coefficient, the slope of the regression line provides an estimate of the linear relation between your two variables. However, least squares regression specifically builds a line that minimizes the deviations in the Y direction without worrying about the X direction. The equation you get when you regress Y on X is therefore *not* the same equation you get when you regress X on Y. Correlation, on the other hand, treats the two variables equally. The correlation between X and Y is always exactly the same as the correlation between Y and X. The correlation simply measures the amount of shared variability between two variables.

- After you create a scatterplot between two variables you can add in a least-squares regression line by taking the following steps.
  - Double click the chart in the output file. This will open up the SPSS Chart Editor.
  - Choose **Chart** → **Options** in the Chart Editor.
  - Under **Fit Line** check the box next to **Total**.
  - Click the **OK** button.
  - Close the Chart Editor box.

- As an alternative to drawing a regression line on a scatterpolot, you can have SPSS display a graph that contains your predicted values overlayed on top of the real values. When running your analysis you must ask SPSS to save the predicted values to your data set by clicking the **Save** button and checking the box next to **Unstandardized Predicted Values**. After you run the analysis, you can then have SPSS produce an overlayed graph using the following steps.
  - Choose **Graphs** → **Scatter** → **Overlay**.
  - Click the **Define** button.
  - Click the variable representing your DV and the variable representing your IV.
  - Move that pair into the box labeled **X-Y pairs**.
  - Click the variable representing your predicted values and the variable representing your IV.
  - Move that pair into the box labeled **X-Y pairs**.
  - Make sure that your IV is on the right hand side of both pairs. If a pair is incorrect you should select it and then click the **Swap Pair** button.
  - Click the **OK** button.

  The main advantage of using this method is that you can extract the SPSS syntax for this procedure using the **Paste** button if you want to save your work. The Chart Editor (used to apply actual lines to your graphs) does not use the SPSS base system and so does not have any syntax associated with it.

2.4 Factors affecting the observed strength of a correlation

- The correlation coefficient is designed to summarize the linear relation between a pair of continuous variables that have approximately normal distributions. You can use a correlation to summarize the relations between variables with different distributions, but this makes it more difficult to detect a relation. This is particularly a problem when one or both of your variables are dichotomous.

  Sometimes you expect that the underlying nature of a variable is continuous even though it is measured in a categorical fashion. For example, people will often perform a median split on a variable in order to use it in an ANOVA design. In this case you can actually correct any observed correlations with that variable for the influence of the dichotomy using the formula

$$r\{\text{dichotomy corrected}\} = r\{\text{observed}\}\frac{\sqrt{PQ}}{h}, \tag{2.25}$$

where P and Q are the proportions of observations falling into the two categories of the dichotomous variable (so $Q = 1-P$), and h is the height of the normal distribution at the point where the probability to the left of the Z is equal to either P or Q (the heights will be the same at both points). Values of h can be obtained from Appendix C of Cohen, et al. (2003), or can be directly computed using the equation

$$h = \frac{\exp\left(-\frac{Z^2}{2}\right)}{\sqrt{2\pi}}, \tag{2.26}$$

where "exp" refers to the exponential function and Z is the value of the standard normal distribution where the probability to the left is equal to P.

- Correlation coefficients can be reduced if you have random error in the measurement of either variable. The *reliability* of a measure is defined as the proportion of the variability in the observed scores that can be attributed to systematic elements of the measure. Reliability ranges from 0 to 1, where higher values indicate more reliable measures. The maximum correlation that you can obtain with a measure is equal to the square root of the reliability. You can correct the observed correlation to determine what the relation would have been if the study had used a perfectly reliable measurement using the formula

$$r\{\text{reliability corrected}\} = \frac{r\{\text{observed}\}}{\sqrt{r_{xx}r_{yy}}}, \tag{2.27}$$

where r_{xx} and r_{yy} are the reliabilities of your two variables.

- Correlations can also be reduced if you have a *restriction of range* in either of your variables. You can get a better estimate of the relation between a pair of variables when you examine them across a broader range. If you only have data from a limited range of one of your variables it can reduce your observed correlation. You can correct the observed correlation to determine what the relationship would have been if the study did not suffer from a restriction of range using the formula

$$r\{\text{range corrected}\} = \frac{r\{\text{observed}\}\left(\frac{s\{\text{full}\}}{s\{\text{observed}\}}\right)}{\sqrt{1 + r^2\{\text{observed}\}\left[\left(\frac{s^2\{\text{full}\}}{s^2\{\text{observed}\}}\right) - 1\right]}}, \tag{2.28}$$

where s_{full} is the expected standard deviation of the variable when it does not suffer from restriction of range, and s_{observed} is the observed standard deviation of the variable suffering from restriction of range.

- You should always be very cautious when interpreting a correlation where one of the variables is a composite of other variables. Some examples of composites would be difference scores or ratios. If any of the individual components making up the composite are related to the other measure in the correlation, you will observe a correlation between the entire composite and the measure.

Chapter 3

Multiple Correlation and Regression

3.1 Introduction to multiple correlation and regression

- While correlation and simple linear regression can tell us about the relation between two variables, they cannot tell us about the relations among three or more variables. However, both of these procedures have extensions that allow us to determine how multiple IVs relate to a single, continuous DV. The extension of correlation is called *multiple correlation*, while the extension of regression is called *multiple regression*.
- Multiple regression allows us to build an equation predicting the value of the DV from the values of two or more IVs. The parameters of this equation can be used to relate the variability in our DV to the variability in specific IVs. Multiple correlation can tell us the correlation between the dependent variable and an optimally weighted sum of the independent variables. The optimal weights are actually those found through multiple regression.
- Sometimes people use the term *multivariate regression* to refer to multiple regression, but most statisticians do not use "multiple" and "multivariate" as synonyms. Instead, they use the term "multiple" to describe analyses that examine the effect of two or more IVs on a single DV, while they reserve the term "multivariate" to describe analyses that examine the effect of any number of IVs on two or more DVs.

3.2 Relating two IVs to a single DV

- We are going to start our discussion of multiple correlation and regression by examining the situation where you want to determine the effect of two different IVs on a single DV. This will allow us to see the general effects of including multiple IVs in a correlation or regression analysis without requiring very complicated equations.
- The general model for multiple regression with two predictors X_1 and X_2 is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \tag{3.1}$$

  where $i = 1 \ldots n$ and the ϵ_i's are normally distributed with a mean of 0 and a variance of σ^2.
- We interpret the parameters in this model as follows:
  - β_0 is the expected mean response when $X_1 = 0$ and $X_2 = 0$.
  - β_1 is the change in the expected mean response per unit increase in X_1, when X_2 is held constant.
  - β_2 is the change in the expected mean response per unit increase in X_2, when X_1 is held constant.
- β_1 and β_2 are called *partial regression coefficients*. Least squares estimates for these parameters can be computed using the formulas

$$b_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right) \tag{3.2}$$

and

$$b_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right), \tag{3.3}$$

where s_y is the standard deviation of Y, s_1 is the standard deviation of X_1, s_2 is the standard deviation of X_2, r_{y1} is the correlation between Y and X_1, r_{y2} is the correlation between Y and X_2, and r_{12} is the correlation between X_1 and X_2.

We can examine these equations to understand the factors that influence our parameter estimates in multiple regression. The first term scales our parameters in terms of the standard deviation of the DV and the corresponding IV. Looking at the second term, we can see that our coefficients get larger as the correlation between the DV and the corresponding IV gets larger, and they get smaller as the correlation between the DV and the other IV gets larger. We also see that the value of the coefficient is dependent on the correlation between the two IVs, although that relation is somewhat complex because the term falls in both the numerator and the denominator of the second term.

- The important thing to notice is that the coefficient for an IV is only partly determined by its relation to the DV. It is also affected by its relation to other IVs, as well as their relations to the DV. Specifically, the value of a multiple regression coefficient represents the relation between the part of the corresponding IV that is unrelated to the other IVs with the part of the DV that is unrelated to the other IVs. It therefore represents the unique ability of the IV to account for variability in the DV.

- One implication of the way coefficients are determined is that your parameter estimates become very difficult to interpret if there are large correlations among your IVs. The effect of these relations on multiple regression coefficients is called *multicollinearity*. This changes the values of your coefficients and greatly increases their variance. It can cause you to find that none of your coefficients are significantly different from zero, even when the overall model does a good job predicting the value of the DV.

  The typical effect of multicollinearity is to reduce the size of your parameter estimates. Since the value of the coefficient is based on the unique ability for an IV to account for variability in a DV, if there is a portion of variability that is accounted for by multiple IVs, all of their coefficients will be reduced.

  Under certain circumstances multicollinearity can also create a *suppression effect*. If you have one IV that has a high correlation with another IV but a low correlation with the DV, you can find that the multiple regression coefficient for the second IV from a model including both variables can be *larger* (or even opposite in direction!) compared to the coefficient from a model that doesn't include the first IV. This happens when the part of the second IV that is independent of the first IV has a different relation with the DV than does the part that is related to the first IV. It is called a suppression effect because the relation that appears in multiple regression is suppressed when you just look at the second variable by itself.

- The *partial correlation* can be used to measure the ability of the part of one IV that is independent of the second IV to predict the part of the DV that is independent of the second IV. When you have two IVs, you can calculate the partial correlation coefficients using the formulas

$$pr_1 = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{y2}^2}\sqrt{1 - r_{12}^2}} \tag{3.4}$$

and

$$pr_2 = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{y1}^2}\sqrt{1 - r_{12}^2}}. \tag{3.5}$$

The square of a partial correlation is the proportion of the variance in your DV not associated with the other IV that can be explained by an IV. This is conceptually the same as what your regression coefficients represent.

- You can also calculate the *semipartial correlation* (also called the *part correlation*), which measures the relation between the part of one IV that is independent of the second IV with the entire DV. You might prefer this to the partial correlation if you want to have an estimate of the amount of the total variability in your DV that is related to each IV. When you have two IVs, you can calculate the semipartial correlation coefficients using the formulas

$$sr_1 = \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}} \tag{3.6}$$

and

$$sr_2 = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{1 - r_{12}^2}}. \tag{3.7}$$

The square of a semipartial correlation is the proportion of the total variance in your DV that is uniquely explained by an IV.

Notice the similarity of the formulas for the parital and semipartial correlations. The only difference is that the partial correlation includes an additional term in the denominator representing the portion of variance in the DV that is associated with the other IV.

- The standard correlation between an IV and the DV is often referred to as the *zero-order correlation* to distinguish it from the partial and semipartial correlations.

- You can summarize the joint ability of your IVs to predict the DV using the multiple correlation coefficient, calculated using the formula

$$R_{y.12} = \sqrt{\frac{r_{y1}^2 + r_{y2}^2 - 2r_{y1}r_{y2}r_{12}}{1 - r_{12}^2}}. \tag{3.8}$$

Here we use a capital "R" to represent the correlation to indicate that it is a multiple correlation instead of a bivariate correlation. This coefficient represents the maximum correlation between a linear combination of the variables to the left of the period (Y) with a linear combination of the variables to the right of the period (X_1 and X_2). It just so happens that in our circumstance, the optimal correlation can be found if we multiply the values of our IVs by their least squares regression coefficients.

The square of the multiple correlation is the coefficient of determination, which is the proportion of variance in Y that can be jointly explained by X_1 and X_2.

- It is important to note that while multicollinearity affects the estimation of your individual parameters, it does not affect inferences regarding the full model. Multicollinearity will therefore never affect the multiple correlation or the coefficient of determination.

3.3 Relating more than two IVs to a single DV

- The general form of the multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i, \tag{3.9}$$

where $i = 1 \ldots n$, and ϵ_i follows a normal distribution with mean 0 and variance σ^2.

- This is known as the *general linear model* (GLM), and can actually be used for many different types of regression. The main restriction of this model is that it must be linear in the *parameters*. This means that you cannot have a parameter that is raised to a power. However, there are not restrictions on the terms multiplied by the parameters. Using different types of terms, the GLM can be used to

  - Perform polynomial regression, where your response variable is related to some power of your predictor. You could then explain quadratic, cubic, and other such relations. This will be discussed in detail in Chapter 6.

- Test categorical variables, where the value codes the observation as being in a particular group. You can then test how this categorization predicts some response variable. This will be discussed in detail in Chapter 8.
- Consider interactions between variables. In this case one of your predictors is actually the product of two other variables in your model. This lets you see if the effect of one predictor is dependent on the level of another. This will be discussed in detail in Chapters 7 and 9.

- The basic ideas discussed in multiple regression with two IVs can all be extended to multiple regression with more than two IVs.
  - β_0 is the mean response where all of your IVs = 0.
  - β_i is the expected change in Y per unit increase in X_i when all other IVs are held constant.
  - The values for β_i are based on the relation between the independent part of each X_i with the independent part of Y.

- You can obtain least squares estimates for the regression coefficients in equation 3.9, but these are too complicated to present in a standard formula. In matrix notation, you can calculate your vector of parameter estimates using the formula

$$\mathbf{b} = (\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X}'\,\mathbf{Y}\,. \tag{3.10}$$

  For more information on the matrix approach to multiple regression, see Chapters 5 and 6 of Neter, Kutner, Nachtsheim, and Wasserman (1996). The important thing to understand is that these estimates are based on the same three factors discussed above in section 3.2.

  1. The relation between the IV corresponding to the parameter being estimated and the DV.
  2. The relations between other IVs in the model and the DV.
  3. The relations between the corresponding IV and the other IVs in the model.

- The standard errors of the parameter estimates can be calculated using the formula

$$s\{b_i\} = \left(\frac{s_y}{s_i}\right)\sqrt{\frac{1}{1-R_i^2}}\sqrt{\frac{1-R_y^2}{n-k-1}}, \tag{3.11}$$

  where s_y is the standard deviation of the DV, s_i is the standard deviation of the X_i, R_y^2 is the multiple correlation between the DV and all of the IVs, and R_i^2 is the multiple correlation between X_i and all of the other IVs.

- If we want to compare one of the estimated regression coefficients to a specific value, we can perform a hypothesis test of a point estimate with the following characteristics.
  - $H_0 : \beta_i = \beta_{\text{null}}$
    $H_a : \beta_i \neq \beta_{\text{null}}$.
  - Estimate = b_i
  - Standard error of estimate = $s\{b_i\}$
  - Degrees of freedom = $n - k - 1$, where k is the number of IVs
  - $t = \frac{b_i - \beta_{\text{null}}}{s\{b_i\}}$

  This test is commonly performed using $\beta_{\text{null}} = 0$, which tells you whether a given IV can independently account for a signifcant amount of the variability in the DV.

- You can calculate a confidence interval around a coefficient using the estimate, its standard error, and the value of the t distribution with $n-k-1$ degrees of freedom that is associated with your confidence level.

- For each case in your data set, you can calculate its predicted value by simply substituting the observed values of your X_i's in the equation

$$\hat{Y_i} = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_k X_{ik}. \tag{3.12}$$

  You can also determine the residual for each case by subtracting the predicted value from the observed value, as in the equation

$$e_i = Y_i - \hat{Y_i}. \tag{3.13}$$

- To perform multiple regression in SPSS
  - Choose **Analyze** → **Regression** → **Linear**
  - Move the the DV to the box labeled **Dependent**.
  - Move all of the IVs to the box labeled **Independent(s)**.
  - Click the **OK** button.
- The SPSS output from a multiple regression analysis contains the following sections.
  - **Variables Entered/Removed**. This section is only used in model building (discussed in Chapter 5) and contains no useful information in standard multiple regression.
  - **Model Summary**. The value listed below **R** is the multiple correlation between your IVs and your DV. The value listed below **R square** is the proportion of variance in your DV that can be accounted for by the entire collection of your IVs. The value in the **Adjusted R Square** column is a measure of model fit, adjusting for the number of IVs in the model. It is sometimes used to compare models containing different numbers of IVs. The value listed below **Std. Error of the Estimate** is the standard deviation of the residuals.
  - **ANOVA**. This section provides the F test for your overall model. If this F is significant, it indicates that the model as a whole (that is, all IVs combined) predicts significantly more variability in the DV compared to a null model that only has an intercept parameter. Notice that this test is affected by the number of IVs in the model being tested.
  - **Coefficients**. This section contains a table where each row corresponds to a single coefficient in your model. The row labeled **Constant** refers to the intercept, while the other rows refer to the regression coefficients for your IVs. Inside the table, the column labeled **B** contains the regression coefficients and the column labeled **Std. Error** contains the standard error of those coefficients. The column labeled **Beta** contains the standardized regression coefficients. The column labeled **t** contains the value of the t-statistic testing whether each coefficient is significantly different from zero. The p-value of this test is found in the column labeled **Sig**. A significant t-test indicates that the IV is able to account for a significant amount of variability in the DV, independent of the other IVs in your regression model.
- As in simple linear regression, we can make inferences about a predicted value based on our regression equation. Given a set of values for our IVs $X_{h1}, X_{h2} \ldots X_{hk}$ we can calculate an predicted value using the formula

$$\hat{Y_h} = b_0 + b_1 X_{h1} + b_2 X_{h2} + \cdots + b_k X_{hk} \tag{3.14}$$

  To compare the mean expected value of $\hat{Y_h}$ to a specific number, you can perform a hypothesis test of a point estimate with the following characteristics.
  - $H_0 : Y_h = Y_0$
    $H_a : Y_h \neq Y_0$
  - Estimate = $\hat{Y_h}$
  - Standard error of estimate = $s\{\text{mean } \hat{Y_h}\}$ (see below)
  - Degrees of freedom = $n - k - 1$ where k is the number of IVs

- $t = \frac{\hat{Y_h} - Y_0}{s\{\text{mean } \hat{Y_h}\}}$

The formula for the standard error of this estimate is very difficult to represent without using matrix notation. It is presented on page 95 of the textbook as formula 3.8.1. In matrix notation, the formula would be written as

$$s\{\text{mean } \hat{Y_h}\} = \sqrt{\text{MSE}\,(\mathbf{X_h}'(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X_h})}, \tag{3.15}$$

where $\mathbf{X}$ is a matrix containing the observed values of your IVs in your data set and $\mathbf{X_h}$ is a vector containing the values of the IV for which you want to make your prediction.

- You can calculate a standard confidence interval around the expected value of Y_h using the estimate and standard error provided above, along with the value of the t distribution with $n - k - 1$ degrees of freedom that is associated with your confidence level.

  You can calculate a prediction interval using the estimate $\hat{Y_h}$, the prediction error, and the value of the t distribution with $n - k - 1$ degrees of freedom that is associated with your confidence level. The formula for the prediction error using matrix notation would be

$$s\{\text{pred}\} = \sqrt{\text{MSE}\,(1 + \mathbf{X_h}'(\mathbf{X}'\,\mathbf{X})^{-1}\,\mathbf{X_h})}, \tag{3.16}$$

  where $\mathbf{X}$ is a matrix containing the observed values of your IVs in your data set and $\mathbf{X_h}$ is a vector containing the values of the IV for which you want to make your prediction. Unfortunately this formula is too complicated to present easily using scalar notation.

- You can have SPSS produce confidence intervals and prediction intervals for cases in multiple regression in exactly the same way discussed in section 2.3 for simple linear regression. The only difference would be that you would include multiple IVs in the model you define.

- It can be difficult to directly compare the coefficients for your two IVs if they have substantially different standard deviations. In this case you might want to calculate *standardized regression coefficients* which will be on the same scale. These coefficients can be calculated from the regular regression coefficients using the formula

$$\beta_i = b_i \left(\frac{s_i}{s_y} \right), \tag{3.17}$$

  where b_i is your regression coefficient, s_i is the standard deviation of the IV corresponding to the coefficient, and s_y is the standard deviation of the DV. You can also obtain the standardized regression coefficients by simply standardizing all of your variables and then performing the typical regression analysis.

  It is important to note that people conventionally use the symbol β to represent a standardized regression coefficient. It is unfortunate that this is the same symbol that is used to represent the population parameter being represented by a regression coefficient. Whenever you see this symbol you should take extra care to make sure you know which one it refers to. If you are reading it off the output of a statistical analysis, it will almost always indicate a standardized regression coefficient.

- If you know the correlations among all of your IVs and the standardized regression coefficients between each IV and the DV, you can compute the value of the zero-order correlation between a given X_i and Y using the formula

$$r_{yi} = \sum_{j=1}^{k} r_{ij}\beta_j, \tag{3.18}$$

  where r_{yi} is the correlation between X_i and Y, k is the number of IVs in the model, r_{ij} is the correlation between X_i and X_j, and β_j is the standardized regression coefficient for X_j. Note that when $i = j$, r_{ij} will be equal to 1.

- You can compute the partial and semipartial correlation between each IV and the DV within multiple regression using the formulas

$$pr_i = \frac{\beta_i \sqrt{1 - R_i}}{\sqrt{1 - R^2_{y.(i)}}} \tag{3.19}$$

and

$$sr_i = \beta_i \sqrt{1 - R_i}, \tag{3.20}$$

where β_i is the standardized regression coefficient for X_i, R_i is the multiple correlation between X_i and all of the other IVs, and ${R^2}_{y.(i)}$ is the multiple correlation between the DV and all of the IVs *except* X_i. The significance values of both the partial and semipartial correlations will always be the same as the p-value associated with the corresponding regression coefficient.

- You can have SPSS produce both the partial and semipartial correlations for all of the IVs in a regression equation by taking the following steps.
  - Choose **Analyze** → **Regression** → **Linear**.
  - Define your IVs and DVs as you would in a normal multiple regression.
  - Click the **Statistics** button.
  - Check the box next to **Part and Partial Correlations**.
  - Click the **Continue** button.
  - Click the **OK** button.

  SPSS will then produce the standard output for a regression analysis, but with three extra columns in the **Coefficients** table. The first, labeled **Zero-order**, contains the raw bivariate correlation between the IV listed to the left of the row and the DV. The column labeled **Partial** contains the partial correlation while the column labeled **Part** contains the semipartial correlation of the IV listed to the left of the row with the DV, controlling for all the other IVs in the model.

- The *total sums of squares* (SST) measures the total amount of variability found in your DV, and can be calculated using the formula

$$\text{SST} = \sum (Y_i - \bar{Y})^2. \tag{3.21}$$

  It is often useful to partition the SST into the part that is explained by your regression equation and the part that is not. The part of the variability that can be explained is called the *regression sums of squares* (SSR), and can be calculated using the formula

$$\text{SSR} = \sum (\hat{Y}_i - \bar{Y})^2. \tag{3.22}$$

  The part of the variability that cannot be explained is called the *sum of squared errors* (SSE), and can be calculated using the formula

$$\text{SSE} = \sum (Y_i - \hat{Y}_i)^2. \tag{3.23}$$

  The SSR and the SSE will always sum to be equal to the SST.

- If you divide the SSR by its degrees of freedom (equal to k, the number of IVs) you can obtain the *regression mean square* (MSR). Similarly, if you divide the SSE by its degrees of freedom (equal to $n - k - 1$) you can obtain the *mean squared error* (MSE). We discussed the MSE in Chapter 2, where we defined it as the variance of the residuals. Both methods of calculating the MSE will provide the same result.

- You can calculate the *coefficient of multiple determination* using the formula

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \sum \beta_i r_{yi}, \tag{3.24}$$

  where β_i is the standardized regression coefficient for X_i and r_{yi} is the bivariate correlation between the DV and X_i. This is the proportion of variance in the DV that can be explained using all of your IVs.

- You can test whether the model R^2 is significantly different from zero to determine if your collection of IVs can account for a significant portion of the variance in the DV. This tests the hypotheses

  $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$
  H_a : At least one $\beta \neq 0$.

  To perform this test you calculate the statistic

$$F = \frac{\text{MSR}}{\text{MSE}} = \frac{\left(\frac{\sum(\hat{Y}_i - \bar{Y})^2}{k}\right)}{\left(\frac{\sum(Y_i - \hat{Y}_i)^2}{n-k-1}\right)} = \frac{R^2(n-k-1)}{(1-R^2)k} \tag{3.25}$$

  which follows an F distribution with k numerator and $n-k-1$ denominator degrees of freedom. Typically you only look at the tests of individual coefficients if the F test for the overall model is significant. This acts as a control to prevent the inflation of your experimentwide α from performing separate tests for each IV.

Chapter 4

Regression Assumptions and Basic Diagnostics

4.1 How assumptions affect inferences

- In Chapters 2 and 3 we discussed how to both determine the least-squares estimates for linear regression and how to make inferences based on those estimates. However, these estimates and inferences are only valid for certain combinations of IVs and for certain types of data. The features that your analysis must have that are required for valid inferences from a statistical analysis are its *assumptions*.
- The assumptions for an analysis are determined when mathematicians originally develop the formulas for estimates and define the distributions of the test statistics. When the assumptions for an analytic procedure are violated, the formulas for parameter estimates may no longer produce values that are actually centered on the population parameter. It is also possible that the values of any test statistics calculated as part of the analysis will not have the expected distribution, so the p-values you calculate from those statistics will not actually represent the true probability of obtaining your observed data when the null hypothesis is true.
- Your data will almost never meet the assumptions of your analyses perfectly. The influence of violating a particular assumption depends on the extent of the violation and the degree to which the analysis is *robust* to violations of that assumption. Sometimes the results of an analysis are basically accurate even when an assumption is violated. In this case we would say that that analysis is robust to violations of that assumption. Other times the results of the analysis will be invalid with even small violations of an assumption, in which case we would say that the analysis is not robust to violations of that assumption.
- You should perform a given analysis only when your data and statistical model do not substantially violate the assumptions, or when the analysis is robust to the particular types of violations that you have. If you have a substantial violation of the assumptions you can sometimes change your data or your statistical model so that the assumptions are then met. Other times you will need to choose a different type of analysis that better matches the characteristics of your data.

4.2 Assumptions of linear regression

- As we mentioned earlier, the mathematical basis for regression analysis is based on the GLM (general linear model). The assumptions for regression analysis are therefore taken from the assumptions of the GLM. Both regression and ANOVA are derived from this model, so you will notice that the assumptions of regression are basically the same as the assumptions for ANOVA.
- Recall the general form of the GLM

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \tag{4.1}$$

  where $i = 1 \ldots n$ and ϵ_i follows a normal distribution with mean 0 and variance σ^2.

- Regression has three assumptions regarding the error terms (ϵ_i) from the above model. Specifically, it assumes that
  - The errors have a normal distribution.
  - The same amount of error is found at each level of X. This means that the average difference between the regression line and the observed values is constant across all values of X. Having equal variance at each level of X is called *homoscedasticity*, while having unequal variance is called *heteroscedasticity*.
  - The errors are all independent.

  It is important to note that these assumptions are all made regarding the *error terms* and not the actual IVs or DVs. For example, it is perfectly acceptable for your DV to have a non-normal distribution as long as the error terms are normally distributed.
- Regression makes two assumptions about the statistical model you use for analysis. Specifically, it assumes that
  - Your model includes all of the IVs that have a significant influence on the DV.
  - The relation between the IVs and the DV is accurately represented in your model. This means that your model contains all of the higher-order (like quadratic or cubic) and interaction terms that are significantly related to your DV.
- Finally, regression assumes that your IVs all have perfect reliability. This means that the IVs are all measured without error. Note, however, that it is perfectly acceptable to have error in your measurements of the DV.

4.3 Determining if the assumptions are met

- You typically test the assumptions of regression by examining plots of your residuals. This requires that you store the values of your residuals in your data set as a variable. To have SPSS save the residuals from a regression model as a variable
  - Choose **Analyze** → **Regression** → **Linear**.
  - Define your regression model, as described in section 3.3.
  - Click the **Save** button.
  - Check the box next to **Unstandardized** in the section labeled **Residuals**.
  - Click the **Continue** button.
  - Click the **OK** button.

  SPSS will then perform the regression analysis you defined and add a new variable with a name beginning with **res** to the end of your data set containing the residuals.

  In addition to saving the raw residuals (which SPSS calls *unstandardized residuals*), SPSS provides you with the option of saving four other types of residuals from your regression model. For more information on these residuals see section 10.1. You can typically work with any of the residuals described in that section when checking the assumptions of regression - they will all give similar results. It can sometimes be easier to work with the various transformed residuals when looking for outliers (discussed in Chapter 10).

  In addition to saving the residuals, you can also save many other types of variables including the predicted values and diagnostic statistics. You simply check the box next to whichever variables you want SPSS to add to your data set. There is no limit to the number of variables you can save from an analysis.
- The first thing you can do to determine if your residuals have a normal distribution is to examine a graph of their distribution. You would then just visually examine the distribution to see if it has the characteristic bell-shaped curve. SPSS will produce three different types of graphs that can be used for this purpose.

A *frequency histogram* first divides the total range of the variable into a predefined number of bins, and then counts the total number of cases that have values in each of the bins. When arranged in ascending order, a graph of the bin counts can be used to get an idea of the approximate distribution of the variable. It is important to choose an appropriate number of bins when building a histogram. It is difficult to see the general form of the distribution in a histogram with a large number of bins, while a histogram with a small number of bins can obscure finer patterns. In can sometimes be useful to look at multiple histograms, each containing a different number of bins.

To create a histogram in SPSS

- Choose **Graphs** → **Histogram**.
- Move the variable whose distribution you want to see to the **Variable** box.
- Check the box next to **Display Normal Curve** if you want to see an image of a corresponding normal distribution overlayed with the histogram.
- Click the **OK** button.

After creating a histogram, you can change the number of bins it uses by taking the following steps.

- Double-click the histogram in the output window. This will open up the SPSS Chart Editor.
- Double-click the X-axis inside the Chart Editor.
- Click the button next to **Custom** in the **Intervals** box.
- Click the **Define** button.
- Enter the number of intervals you want to use in the box next to **# of intervals**.
- Click the **Continue** button.
- Click the **OK** button.
- Close the Chart Editor.

- You can also display a *stem-and-leaf plot*, which is very similar to a histogram. The main difference is that the bars in the stem-and-leaf plot are composed of stacks of the last digit of the observed values. Just as in a histogram, the height of a stack tells you how many observations fall inside a particular bin. The main advantage of a stem-and-leaf plot is that it provides information about the exact values that were used to create the plot. However, you do not have the option of overlaying a normal curve, so it can be slightly more difficult to determine whether the observed distribution is normal.

  To create a stem-and-leaf plot in SPSS

  - Choose **Analyze** → **Descriptive Statistics** → **Explore**.
  - Move the variable you want to examine to the box labeled **Dependent List**.
  - Click the **Plots** button.
  - Check the button next to **Stem-and-leaf** inside the **Descriptive** box.
  - Click the **Continue** button.
  - Click the **OK** button.

- The final distribution graph that you might want to examine would be a *boxplot*. A boxplot is essentially a graphical display of the 5-number summary. Your textbook provides detailed information about the numbers that go into a boxplot on page 108. In the boxplot for a variable that has a normal distribution, the median line appears in the center of the interquartile range (IQR) box, and the top and bottom whiskers are of approximately the same length.

  To create a boxplot in SPSS

  - Choose **Graphs** → **Boxplot**.
  - Click **Simple**.
  - Click the button next to **Summaries of separate variables** in the **Data in Chart Are** box.
  - Click the **Define** button.
  - Move the variable of interest to the **Boxes Represent** box.

- Click the **OK** button.

- A better alternative to just eyeballing a graph of a variable's distribution is to create a *normal probability plot*. A normal probability plot compares the percentiles of the values from your distribution to the percentiles of the standard normal distribution. For example, in a normal distribution, .3% of the values are more than 3 standard deviations above the mean, 5% of the values are more than 2 standard deviations above the mean, and 33% of the values are more than 1 standard deviation above the mean. If your variable has a normal distribution, then it should *also* have .3% of its values more than 3 sd above the mean, 5% of its values more than 2 sd above the mean, and 33% of its values more than 1 sd above the mean. The plot itself has the percentiles of the standard normal distribution on the Y axis and the percentiles of your variable on the X axis. If your variable has a normal distribution, then the normal probability plot will look like a straight line. If the plot has areas that look more like curves, it indicates that there are parts of your distribution that deviate from normality.

  To create a normal probability plot in SPSS

  - Choose **Graphs** → **P-P**.
  - Move the variable you want to examine to the **Variables** box.
  - Make sure that **Normal** is selected under **Test Distribution**.
  - Click the **OK** button.

  Examining a *Q-Q plot* will accomplish basically the same thing as examining a normal probability plot. It uses a different metric on the axes (the observed values of the variable versus the values from a variable with a perfectly normal distribution with the same mean and standard deviation) and so will not look exactly like a normal probability plot. However, normal distributions will again produce an approximately straight line while non-normal distributions will have systematic curves.

  To create a Q-Q plot in SPSS

  - Choose **Graphs** → **Q-Q**.
  - Move the variable you want to examine to the **Variables** box.
  - Make sure that **Normal** is selected under **Test Distribution**.
  - Click the **OK** button.

  Regression is relatively robust to violations of normality, so you typically do not have to worry about small or moderate deviations as long as you have a reasonable sample size (over 30 cases).

- The simplest way to check for homogeneity of variance across different levels of the IV (homoscedasticity) is to examine a scatterplot of the residuals against the predicted values. If the spread of the residuals appears to be constant across the different levels of the IV, then you likely do not have a problem. If the spread varies greatly or appears to have a pattern (such as increasing variability with increasing values of the IV) there could potentially be a problem.

  You can request both the residuals and the predicted values from a regression analysis in SPSS, as discussed at the top of this section. You can then simply create a scatterplot of these two variables to see how the residuals vary by the predicted value.

  Regression is robust to violations of homoscedasticity, so you usually do not have to worry about this assumption unless there is good evidence that the variance of the DV at one level of your IV is ten times the variance of the DV at another level of your IV.

- There is no general procedure to detect dependence among the error terms. Instead you must start with some theory as to how they may be dependent. If your different cases were measured at different times, you might check to see if observations made close together are more similar than those separated by a greater amount of time. This can be tested using an *autocorrelation* analysis, discussed on page 136 of the text. Similarly, if you ran your study in groups, you might check to see if people in the same group had more similar scores on the DV than those in different groups. The easiest way to test for this is to perform an ANOVA predicting the value of the DV from the group.

  Regression is *not* robust to violations of independence (Kenny & Judd, 1986), so it is important that your statistical model takes any and all dependencies among your cases into account.

- To determine whether a particular variable should be included in your statistical model because it is an important IV you must have made measurements of that variable at the same time that you collected the data from your original model. You can then determine whether adding that variable would be able to make a significant independent contribution to the prediction of your DV. You can do this graphically with a *partial regression leverage plot*, as discussed on page 127 of the textbook. However, it is easier to simply perform an additional regression analysis where you include the new variable in your statistical model. A significant regression coefficient on the new variable would indicate that it is a significant source of dependence, so you should include it in your model.

- You can determine whether your model accurately represents the relations between the IVs and the DV by examining scatterplots of the residuals against each IV. Any patterns that you might see in these graphs would indicate that the relation between the corresponding IV and the DV is more complex than is indicated in your statistical model. You can often determine the nature of the more complex relation by examining the pattern in the scatterplot.

- The amount of measurement error in the IVs can be determined by a reliability analysis. As mentioned previously, reliability is the proportion of variability in your observed values that can actually be attributed to variability in the underlying theoretical construct of interest. A measure is typically considered to be acceptable if it has a reliability of .7 or above.

  People typically think about computing the reliability of measured constructs, but it is equally possible to determine the reliability of a manipulation. The main requirement is a good measure of the theoretical variable being manipulated. You can then determine the extent to which people in the same level of the manipulation actually have the same amount of the manipulated variable.

4.4 Checking diagnostics using ARC

- In the last section we discussed how you can check the assumptions of your regression model in SPSS. However, we were forced to skip a discussion of several graphs presented in your textbook because SPSS does not have the ability to present these graphs. Luckily, there is a free program called ARC that you can use to create these graphs. In this section we will discuss how to obtain ARC and use it to generate these other graphs, many of which will make checking the assumptions of your model easier.

- The ARC software is available at the website

  `http://www.stat.umn.edu/arc/software.html`

  After you install ARC itself you will also want to install an ARC add-in to Excel to make it easier to load your Excel and SPSS data sets into ARC. A link to the add-in as well as the instructions on how to install and use it are available at the website

  `http://www.stat.umn.edu/arc/exceltoarc.html`

  Once you install this add-in you will be able to read an Excel spreadsheet into ARC by taking the following steps.

  - Start the ARC program.
  - Load the data set you want to examine in Excel. The names of the variables should be in the first row. If you want to examine an SPSS data set you must first save it in Excel format within SPSS before you load it into Excel.
  - Within Excel, choose **Tools** → **ARC** → **Export data**.
  - Click **Export**

  The data set will now be loaded into ARC. You should see a new entry in ARC's pull-down menu with the name of the new data set.

  Note: If you do not already have a directory called `temp` off of the root directory of your hard disk, you will need to create it before you can transfer data to ARC.

- ARC will not accept data sets whose names have spaces in them. You may therefore need to rename your Excel files before exporting them to ARC.

- If you are transferring a data set from SPSS to Excel with plans to then load it into ARC, you will first need to reformat it if you have any missing values. When you save an SPSS data set as an Excel spreadsheet it automatically replaces any missing values with the word **#NULL!**. ARC can not read variables that have letters in them, so before you can export such a file to ARC you will need to replace the **#NULL!** values with empty cells by taking the following steps within Excel.
  - Chose **Edit** → **Replace**.
  - Enter **#NULL!** in the **Find what** box.
  - Leave the **Replace with** box empty. Do not even put in a space or period.
  - Click the **Replace All** button.
- If you want to use ARC to examine the residuals from your regression model you can either run a regression model in SPSS and save the residuals (as discussed in the last section) or else you can actually perform the analyses within ARC. To perform a multiple regression in ARC
  - Choose **Graph&Fit** → **Fit linear LS**.
  - Move the IVs to the box labeled **Terms/Predictors**.
  - Move the DV to the box labeled **Response**.
  - Enter the name you want to give this analysis in the box labeled **Name for Normal Regression**. This will be used later on when referring to residuals and predicted values from this model.
  - Click the **OK** button.

  Whenever ARC performs an analysis it automatically saves the predicted values, the residuals, and several other diagnostic statistics for each case. They are stored in the data set connected to whatever name you gave to the analysis.
- To obtain a histogram in ARC
  - Choose **Graph&Fit** → **Plot of**.
  - Move the variable you want to examine to the box labeled **H**.
  - Make sure the **Plot Controls** box is checked.
  - Click the **OK** button.

  ARC has the ability to change the number of bins for a histogram "on the fly" using a slider. This allows you to easily see what your distribution looks like using a variety of different resolutions. To change the number of bins used by your histogram just move the **NumBins** slider to the value you want.

  When trying to determine if your residuals are normally distributed it can be helpful to add a *kernel density function* to your graph. This basically fits a smooth line to the data represented within your histogram. To add a kernel density function to your graph move the **GaussKerDen** slider to some value other than **NIL**. The value on this slider indicates how much "smoothing" is done to the data. Lower values will follow the specific values more closely while larger values will reflect more general trends.
- To obtain a scatterplot in ARC
  - Choose **Graph&Fit** → **Plot of**.
  - Move the variable you want on the X-axis to the box labeled **H**.
  - Move the variable you want on the Y-axis to the box labeled **V**.
  - Make sure the box labeled **Plot controls** is checked.
  - Click the **OK** button.

  ARC has the ability to add a *lowess line* to your scatterplot. A lowess line is a curve that is drawn to capture the relation between the two variables in a scatterplot. It is often easier to see if there is a relation by examining the lowess line than it would be by simply looking at the entire scatterplot. All you need to do to add a lowess line to your scatterplot is move the **lowess** slider to some value other than **NIL**. The actual value on this slider determines how smooth the lowess line will be. Lower values will follow the specific values more closely while larger values will reflect more general trends.

- As mentioned above, it can be useful to examine scatterplots graphing your residuals against your IVs when trying to determine if your residuals meet the assumption of homoscedasticity. To make this easier, you can have ARC display a lowess line with two additional lines representing a 67% confidence interval (1 SD in either direction) around the predicted value on the lowess line. Strong variability in the width of this interval would indicate a violation of homoscedasticity. To add the confidence interval lines to your scatterplot you should click the triangle next to the **lowess** slider and then select **lowess +- SD** from the menu.

- To obtain a normal probability plot in ARC

  - Choose **Graph&Fit → Probability plot of**.
  - Move the variables you want to plot to the box labeled **Selection**.
  - Click the **OK** button.

  This will produce a standard normal probability plot. Sometimes the patterns that appear in a normal probability plot can be difficult to see in its raw form. It is therefore often useful to look at the deviation scores from the diagonal to see if there are any patterns. If you check the box next to **Rem lin trend** on the normal probability plot it will redraw the graph removing the expected linear trend. Non-normal patterns will typically be easier to see in the de-trended graph. You can also have ARC insert an estimated curve to the data using the **lowess** slider, highlighting any patterns. You should, however, be careful when interpreting these detrended deviation plots. ARC will automatically rescale your plot to best display the deviations. This will make even the smallest deviations appear large on the graph, so be aware of their actual magnitudes.

4.5 Remedial measures for violations

- One option you always have is to not perform a regression analysis. Regression is a powerful tool to detect relations, and it gains some of that power by putting restrictions on the types of data and statistical models that it can handle. The remainder of this section will discuss ways to change your data or model so that it better fits the assumption of regression. However, it's good to keep in mind that you always have the option of performing a different analysis that better handles the characteristics of your data.

- Since regression is robust to many of its assumptions, we often focus on trying to make our analysis as valid as possible rather than simply deciding whether or not your data meet the assumptions.

- Non-normality of the residuals indicates that the DV does not have a linear relation to the IVs. One possible solution is to transform your variables to make the relation linear. For example, instead of trying to predict the value of Y from X you could try to predict the value of Y from $\log(X)$.

  We will hold off a detailed discussion of the possible transformations you might perform until Chapter 6. However, at this point it is important to mention that you have the option of either transforming your IVs or transforming your DV. Transforming your IVs works best when the only problem you have is non-normality of the residuals, while transforming the DV works best when you both have non-normal residuals and heteroscedasticity.

- There are two basic ways to deal with heteroscedasticity. The first is to perform a transformation on your DV. If your variance increases as the value of the DV increases, then a square root or logarithmic transformation will even out the variability. If the variance decreases as the value of the DV increases you can try a square or an inverse transformation. If you decide to transform your DV you may then need to transform some or all of your IVs to maintain the appropriate relations between them.

  The second option is to perform a *weighted regression analysis*. Weighted regression is a variant of standard regression analysis that allows you to indicate the importance of each observation in your data set. Instead of simply minimizing the total SSE (which would treat all of the data points equally), weighted regression puts a higher priority on minimizing the error for important observations. If you have nonconstant variance you can use weighted regression so that your regression equation best predicts those observations that have the lowest variance. This way you are putting the greatest weight on the observations that are the most accurate. In this case, you would want to assign the weights to be

$$w_i = \frac{1}{e_i^2}, \tag{4.2}$$

where e_i is the residual found under standard multiple regression.

To perform a weighted regression in SPSS

- You will first need to obtain the residuals from the standard least-squares regression equations so you can use them as weights. The beginning of this section discusses how you can do this.
- Once you have the residuals from the original analysis stored as a variable in your data set you need to create the weight variable. To create this variable you choose **Transform** → **Compute** and create your weight variable using equation 4.2.
- Choose **Analyze** → **Regression** → **Linear**.
- Move your DV to the **Dependent** box.
- Move your IVs to the **Independent(s)** box.
- Click the button in the lower-left corner labeled **WLS**.
- Move your weight variable to the **WLS Weight** box.
- Click the **OK** button.

The output from a weighted regression analysis can be interpreted in the same way as the output from standard multiple regression.

- If your residuals are not independent then you should change your statistical model so that it accounts for the dependency. Under certain conditions this may mean that you can no longer analyze your data with a standard regression analysis. Complicated groupings of your observations may require you to use *mixed models* or *multilevel modeling*, while complicated patterns of serial dependency may require you to use *time-series analysis*.

- If you discover an important IV that is absent from your statistical model you should revise your model to include the IV.

- If you determine that your statistical model misidentifies the relation between an IV and the DV then you should alter your model to reflect the correct relation. This may involve either transforming the IV or it may involve adding additional terms (like polynomial or interaction terms) to the statistical model. Polynomial regression will be discussed in Chapter 5, while interactions will be discussed in Chapter 7.

- If your IVs have a substantial amount of measurement error, the first thing you should do is try to increase the reliability of your measurements. It is much better to improve the consistency of your manipulations or measuring instruments than to try to correct for the error afterwards. However, it turns out that the presence of measurement error in your independent variables will not normally affect the distributions of your coefficients, and so this assumption can generally be ignored with relative safety (Aldrich, 2005). If you have any concerns, you can try analyzing your data using *structural equation modeling*, which is a statistical procedure that specifically accommodates measurement error in both the IVs and the DV.

Chapter 5

Sequential Regression, Stepwise Regression, and Analysis of IV Sets

5.1 Sequential (hierarchical) regression

- Recall from Chapter 3 that the tests of specific parameters in a regression model reflect the independent ability of an IV to account for variability in the DV. Any variance in the DV that is jointly predicted by two different IVs will actually not improve the statistical test of either parameter, a phenomenon known as multicollinearity. When there is substantial multicollinearity in a regression model, it is possible to have the full model account for a substantial amount of the variability in the DV without any tests of its individual parameters being significant. This is most likely to happen when you are examining the effect of a large number of measured IVs.

- Sometimes we may want variance that is explained by multiple IVs to be "credited" to a specific IV that has causal precedence, is theoretically more important than the others, or is a potential confounding variable. In this case we generate a specific ordering of our variables and enter them into our model one at a time, so that any variability in the DV that is explained by multiple IVs is credited to the IV that is earliest in the list. This typically results in more powerful tests of the model parameters compared to standard multiple regression.

- This procedure is referred to as either *sequential* or *hierarchical regression*. There is an entirely different statistical procedure called *hierarchical linear modeling* (discussed in Chapter 14 of the text), so we will refer to the procedure discussed in this chapter as "sequential regression" to prevent confusion.

- The difference between sequential regression and standard multiple regression parallels the distinction between *type I sums of squares* and *type III sums of squares* found in ANOVA. Type I sums of squares measure the addition to the total variance in the DV explained by a model when you add a new factor to the model (similar to sequential regression). On the other hand, the type III sums of squares measure the unique amount of variability that is explained by each factor (similar to standard multiple regression).

- The first thing you must do to perform a sequential regression is determine the order in which you want your IVs entered into your model. The test of any given IV will control for the influence of any other IVs that appear before it in the list. In other words, the test of a IV in sequential regression represents the relation between the IV and the DV independent of all of the variables earlier than it in the list. The test will *not* be affected by any relations with variables that appear after it in the list.

- Next you perform a number of different regression analyses equal to the number of IVs you want to examine. The first analysis regresses the DV on IV1 (the first IV in the list). The influence of IV1 on the DV is determined by the test of the regression coefficient for IV1 in this model. The second analysis regresses the DV on IV1 and IV2. The influence of IV2 on the DV is determined by the regression coefficient for IV2 in the model. The third analysis regresses the DV on IV1, IV2, and IV3. The influence of IV3 on the DV is determined by the test of the regression coefficient for IV3 in this model. This pattern then continues, so that each regression analysis adds one more IV compared to

the previous one, and the test for the effect of an IV on the DV is taken from the first model that included the IV.

- The coefficients for the individual regression equations used in sequential regression are determined and tested using the formulas provided for standard multiple regression in Chapter 3. The difference between sequential and standard multiple regression appears in the selection of the model you use as the basis for the test of each coefficient. The tests for standard multiple regression are all taken from a single model that includes all of the IVs, while sequential regression takes the test of each IV from a different model (as just described).

- An easy way to perform a sequential regression in SPSS is by using *blocks* in your analysis. When you tell SPSS to perform an analysis with more than one block, it will first run a regression predicting the DV from only the IVs in the first block. It will then run a regression model predicting the DV from the IVs in the first and the second block. Then it will run a regression model predicting the DV from the IVs in the first, second, and third blocks. This continues until it has gone through all of the blocks that you defined. You can use these blocks to perform a sequential regression by including every IV in a separate block, starting with the IVs that have the highest priority.

  To perform a sequential regression in SPSS

  - Choose **Analyze** → **Regression** → **Linear**.
  - Move the DV to the box labeled **Dependent**.
  - Make sure that the selector next to **Method** is set to **Enter**.
  - Move the first IV on the list to the box labeled **Independent(s)**.
  - Click the **Next** button.
  - Move the second IV on the list to the box labeled **Independent(s)**.
  - Click the **Next** button.
  - Continue adding variables one at a time to the **Independent(s)** box, clicking the **Next** button after each, until you go through your entire list of IVs.
  - Click the **OK** button.

- SPSS has a limit of nine blocks when performing a regression analysis. However, it is quite possible that you might want to perform a sequential regression with more than nine IVs. In this case what you can do is break your sequential regression into multiple analyses. You would first perform a standard sequential regression on the first nine IVs. You then conduct a second analysis where you put the nine IVs from the initial model in the first block, and then add in your new IVs one at a time to the remaining blocks, in the order they appear on your list. So the IV that was 10th on your list would appear in the second block, the IV that was 11th on your list would appear in the third block, and so forth. If you had more than 18 variables you would need to run a third analysis where the 18 variables from the first two models were included in the first block, and the variables 19-27 were added one at a time to the remaining blocks. Just like a standard sequential regression, you can draw the test of each IV from the first model in which it first appeared, even if this model was not in the first regression analysis.

- The SPSS output from a sequential regression analysis includes the following sections.

  - **Variables Entered/Removed**. This simply reports the steps in which your variables were added to your statistical model. You should see a single variable added to your model at each step, with no variables removed. You should check the order presented in this table to make sure that it matches the order for your sequential regression.
  - **Model Summary**. This section will contain a table where each row corresponds to one of the regression analyses you asked SPSS to perform. For a given model, the value in the **R** column is the multiple correlation between the IVs contained in the model and your DV. The value in the **R Square** column is the proportion of variance in your DV that can be accounted for by the IVs in the model. The value in the **Adjusted R Square** column is a measure of model fit, adjusting for the number of IVs in the model. The value in the **Std. Error of the Estimate** column is the standard deviation of the residuals for the model.

- **ANOVA**. This section provides an F test for each statistical model. If this F is significant, it indicates that the model as a whole (that is, all IVs combined) is able to account for a significant amount of variability in the DV.
- **Coefficients**. This section provides detailed information about the regression coefficients for each statistical model generated during the model-building process. See Chapter 3 for more information about interpreting the values in this section.

  For sequential regression, you should take the test for each IV from the model in which it was first introduced. You can ignore the tests of that variable in all of the other models.
- **Excluded Variables**. This section provides information about the IVs that are *not* included in each of your statistical models. The information contained in this table is not used in sequential regression.

5.2 Stepwise regression

- *Stepwise regression* is a model-building procedure that attempts to maximize the amount of variance you can explain in your DV while simultaneously minimizing the number of IVs in your statistical model.
- Stepwise regression is typically used when you have a large number of IVs and you want to determine the best combination to predict the value of the DV. Stepwise regression is designed to give you a model that predicts the as much variability as possible with the smallest number of parameters.
- Stepwise regression should be interpreted cautiously or avoided entirely when you are trying to understand theoretical relations. It makes its selection based purely on the amount of variance that variables can explain without any consideration of causal or logical priority. The IVs chosen through a stepwise regression are not guaranteed to be the most important factors affecting your DV. A theoretically meaningful variable that explains a large amount of variability in your DV could be excluded from the model if it also happens to cause changes in other IVs, because it would be collinear with those variables.

  Additionally, stepwise regression attempts to maximize the predictive ability for your IVs in the one specific sample that you collected. Its selections will therefore be affected by any relations that happen to appear due to chance alone. If you cannot come up with a theoretical explanation for an observed relation between an IV and the DV then it may just be an artifact only found in the particular sample you collected.
- Your textbook suggests two circumstances under which stepwise regression should be used. The first would be when you only want to determine the best predictive model and are specifically not interested in drawing inferences about the relations in your data. The second would be when you verify the results of a stepwise regression with data from an independent group of subjects. One option would be to use half of your subjects to build a model using stepwise regression and then use the other half to verify the results.
- There are two basic ways to perform a stepwise regression.
  - In *forward stepwise regression* we start with a simple model and gradually add IVs to it until we are unable to make any significant improvement. Some versions of forward stepwise regression also check the old IVs each time a new one is added to the model to make sure that they are still significant. If it turns out that an IV that was included in an earlier step is no longer making a significant contribution to the prediction of the DV, that IV is then dropped from the model.
  - In *backward stepwise regression* we start with a model containing all of the parameters and gradually drop IVs that make the smallest contribution until dropping the least important variable would significantly reduce the predictive ability of the model.
- To perform a stepwise regression in SPSS
  - Move the the DV to the box labeled **Dependent**.
  - Move all of the IVs to the box labeled **Independent(s)**.

- Set the selector next to **Method** to indicate the type of stepwise procedure that you'd like.
  - **Forward** chooses forward stepwise regression without checking the significance of prior variables included in the model.
  - **Stepwise** chooses forward stepwise regression *including* checks of the significance of prior variables included in the model.
  - **Backward** chooses backward stepwise regression.
- Click the **Options** button to set the significance levels the IVs must have to be added (for forward or stepwise) or removed (for stepwise or backward) from the statistical model.
- Click the **Continue** button.
- Click the **OK** button.

- The SPSS output from a stepwise regression analysis contains the following sections.
  - **Variables Entered/Removed**. This section tells you which IVs were added to or removed from your model at each step.
  - **Model Summary**. This provides you with the **R**, **R square**, **Adjusted R Square**, and the standard error of the residuals (listed below **Std. Error of the Estimate**) for the statistical model at each step of the model-building process. For more information on these statistics see Chapter 3.
  - **ANOVA**. This section provides the F tests for each statistical model created during the stepwise process.
  - **Coefficients**. This section provides detailed information about the regression coefficients for each statistical model generated during the model-building process. See Chapter 3 for more information about interpreting the values in this section. You should take the coefficients found in the very last model to create your prediction equation.
  - **Excluded Variables**. This section contains information about how each of the variables excluded from the model would behave if they were added to the current model. The value in the **Beta in** column is the standardized regression coefficient, the value in the **t** column is the t-test for this coefficient and the value in the **Sig** column is the p-value for the test. The column labeled **Partial Correlation** is the partial correlation for each excluded variable, controlling for all of the included variables. The p-value for testing the regression coefficient will be the same p-value obtained for a test of the partial correlation. **Tolerance** measures the multicollinearity between the excluded IV and the IVs already in the model. See Chapter 10 for more information on tolerance.

5.3 Testing the effects of IV sets

- Sometimes we may want to determine the unique contribution of a subset of our IVs to the prediction of the DV. There are several reasons why you might want to do this.
  - The IVs in the set are all measures of the same construct and you want to determine the unique influence of that construct on the DV.
  - You have several different types of IVs (such as IVs measuring situational, personal, or target characteristics), and you want to determine the influence of each IV type.
  - You are performing a polynomial regression (see Chapter 6) and want to determine the ability of the full polynomial function to predict the DV.
  - You have a categorical IV that is implemented in your statistical model as a collection of indicator variables (see Chapter 8). The effect of the categorization as a whole is the sum of the effects of the individual IVs.

- If you want to know the ability of a set of IVs to explain variability in a DV without controlling for any other variables, you can simply perform a standard multiple regression including only those variables in the set. The F test for the full model will test whether your set of IVs can account for a significant amount of the variability in the DV. The exact proportion of variance that your set can account for will be equal to the model R^2.

- Often times we may want to determine the ability of a set of IVs to account for variability in the DV after controlling for a number of other factors. That way we can determine the ability of your set of IVs to account for variability in the DV independent of the other IVs in your model. When examining the influence of a set of IVs, it can be useful to think of dividing the IVs into those in your set (which we will call set B) and those that are not in your set (which we will call set A).

- Just as two variables may both predict the same variability in the DV, so can two sets of IVs jointly predict the same variability. When testing the influence of an IV set, we therefore typically consider the unique ability of a set to explain the DV, above and beyond the variability that can be predicted by the other IVs in the model. However, it is also possible to establish a prioritized list of variable sets for sequential regression (see section 5.4).

- You can calculate the partial R^2 for a set of variables using the formula

$$pR^2_B = \frac{R^2_{y.AB} - R^2_{y.A}}{1 - R^2_{y.A}}, \tag{5.1}$$

  where $R^2_{y.AB}$ is the proportion of variance in the DV that can be explained by the combination of sets A and B (the R^2 from the full regression model), and $R^2_{y.A}$ is the proportion of variance in the DV that is explained by A (equal to the square of the multiple correlation between the DV and the variables in set A). Looking at this equation more closely, we can see that the partial R^2 for set B is equal to the variability in the DV not associated with the variables in set A that can be explained by the IVs in set B divided by the total amount of variability in the DV not associated with the variables in set A. It represents the increase in the overall model R^2 when we add the variables in set B to a model already containing the variables in set A. It has the same interpretation as the partial R^2 values that we can calculate for a single IV - it measures the relation between the part of B that is independent of A with the part of the DV that is independent of A.

- You can also perform a hypothesis test to determine whether the IVs in set B can uniquely account for a significant amount of variability in the DV. To do this you calculate the test statistic

$$F = \frac{\left(\frac{R^2_{y.AB} - R^2_{yA}}{k_B}\right)}{\left(\frac{1 - R^2_{y.AB}}{n - k_A - k_B - 1}\right)} = \left(\frac{R^2_{y.AB} - R^2_{yA}}{1 - R^2_{y.AB}}\right)\left(\frac{n - k_A - k_B - 1}{k_B}\right), \tag{5.2}$$

  where n is the number of cases, k_A is the number of IVs in set A, and k_B is the number of IVs in set B. This statistic follows an F distribution with k_B numerator and $(n - k_A - k_B - 1)$ denominator degrees of freedom. Conceptually, the F statistic in equation 5.2 is equal to the variability that can be uniquely attributed to the IVs in set B divided by the variability that can not be explained by the IVs in either set.

- Considering this same test from the framework of ANOVA, the test statistic in equation 5.2 can be expressed as

$$F = \frac{\left(\frac{\text{SSR}_{AB} - \text{SSR}_A}{k_B}\right)}{\text{MSE}_{AB}}, \tag{5.3}$$

  where SSR_{AB} is the SSR from a model containing the IVs from both sets A and B, SSR_A is the SSR from a model containing only the IVs from set A, and MSE_{AB} is the MSE from the model containing the IVs from both sets A and B. The statistic would again follow an F distribution with k_B numerator and $(n - k_A - k_B - 1)$ denominator degrees of freedom.

- The following steps test the unique contribution of a set of variables B in a regression model containing an additional set of variables A using SPSS.

  - Choose **Analyze** → **Regression** → **Linear**.
  - Move the DV to the box labeled **Dependent**.

- Move the IVs from set A (those that are *not* in the set you want to test) to the box labeled **Independent(s)**.
- Click the **Next** button.
- Move the IVs from set B (those that *are* in the set you want to test) to the box labeled **Independent(s)**.
- Click the **Statistics** button.
- Check the box next to **R squared change**.
- Click the **Continue** button.
- Click the **OK** button.

- The output from this analysis will be exactly the same as for a standard sequential regression with the exception of the **Model Summary** section. To the right of the standard statistics describing the fit of each model is a new section titled **Change Statistics**. The values in this section each describe the improvement of a later model over an earlier model (the first model is compared to a null model containing only an intercept). The column **R Square Change** contains the improvement in R^2 over the prior model. The value in this column for Model 2 will be the partial R^2 for IV set B. The column **F change** presents the results of a test statistic testing whether the new variables can explain significantly more variability than the prior model. The value in this column for Model 2 will be the F test determining whether the IVs in your set make a significant contribution to the prediction of the DV. Following the test statistic there are three columns containing the numerator degrees of freedom (labeled **df1**), the denominator degrees of freedom (labeled **df2**), and the p-value (labeled **Sig. F Change**) for this statistic.

- If you want to test the unique contribution of several different sets in the same analysis you just repeat the above procedure once for each set, redefining sets A and B each time so that B represents the set being tested and A represents all of the other variables in the model.

5.4 Sequential regression with IV sets

- The procedures described in section 5.3 allow you to determine the unique ability of an IV set to account for variability in the DV. However, just like with normal variables, the standard tests of the predictive ability of the set will not take into account any variability that can be accounted for by variables in the set that can also be explained by variables outside of the set.

- Just as we can establish a prioritized list of IVs for use in standard sequential regression, we can also establish a prioritized list of IV sets. In this case, any variability that is jointly explained by two or more sets will be "credited" to the set that appears earliest on the prioritized list. This will usually improve the statistical tests of those sets that appear early in the list.

- The procedure for performing a sequential regression using a set of IVs is basically the same as a standard sequential regression. You should start by developing a prioritized list of IV sets. In this list a "set" can be a single IV or a collection of IVs. The tests of sets that appear later in the list will be controlled for the effects of IVs contained in sets appearing earlier in the list.

- After establishing the list you then perform a series of regression analyses, each time adding in the variables from the next set in your list. For each set you can then compute the partial R^2 and test whether the set provides an increase in the overall model's ability to predict the DV using the procedures described in section 5.3.

- You can also examine the specific regression coefficients of individual variables within the IV sets. These coefficients should be taken from the first model in which they appear. The coefficients will control for the influence of all the other IVs included in the same set as well as all the IVs from sets appearing earlier in the list.

- To perform a sequential regression using IV sets in SPSS
  - Choose **Analyze** → **Regression** → **Linear**.

- Move the DV to the box labeled **Dependent**.
- Make sure that the selector next to **Method** is set to **Enter**.
- Move the IVs from the first set on the list to the box labeled **Independent(s)**.
- Click the **Next** button.
- Move the IVs from the second set on the list to the box labeled **Independent(s)**.
- Click the **Next** button.
- Continue adding the sets of variables to the **Independent(s)** box, clicking the **Next** button after each set, until you go through your entire list of IV sets.
- Click the **Statistics** button.
- Check the box next to **R squared change**.
- Click the **Continue** button.
- Click the **OK** button.

- The output from a sequential regression using IV sets is basically the same as that for a standard sequential regression. The significance of each IV set can be found in the **Model Summary** section in the columns labeled **Change Statistics**. Here you will see an F test examining whether each of the different sets you defined is able to explain a significant amount of the variability in the DV above and beyond that explained by the IVs included in earlier sets. You can find the tests of individual IVs in the **Coefficients** section. The test for a given IV should be taken from the model in which it first appeared.

5.5 General strategies for multiple regression analysis

- **Separate exploratory and inferential analyses**. This strategy suggests that you should not perform inferential tests on a data set on which you have already performed exploratory analyses. Specifically, if you use the results of a statistical analysis to determine what variables you should include in your statistical model you should always test your model on a different data set.

  One benefit having a separate exploratory stage is that you do not have to worry about "following the rules" established by statisticians when you are just trying to understand the relations in your data set. You can run a large number of analyses and not worry about their effects on your overall experimental error. You can choose to discard variables that don't behave in the way that you expect them to. The statistics that you calculate during this stage are only for your own use, so you don't have to worry about violating any of their analytic assumptions.

  The second benefit of separating your exploratory and inferential stages is that any changes you might choose to make to your variables or the statistical model will not alter the actual p-values of your statistics. If you use the same data set to both determine what variables you should include in a statistical model and to actually test that model, the p-values that you calculate for your statistics will be too low. In this case, you have specifically selected the IVs that best match the relations appearing in the data set. Some of the relations you observe, however, may just be caused by random chance. By selecting the variables that had the strongest observed relations, you are also selecting the variables that most likely had random chance working in their favor. If you were to test the strength of these relations in a second data set you would find that the average strength would decrease (a phenomenon called *regression toward the mean*). In fact, the statistics calculated in a second data set will be more accurate predictors of the true relations found in the population.

  One strategy that people sometimes employ is to use a subset of their collected data for exploratory purposes and to use the remainder to test the relations that the exploration happens to turn up. In this case you may want to use a smaller percentage of your data for exploratory purposes since you will not be reporting any statistics from these analyses. In this case, you might also want to use a more liberal significance level during exploration to counter the reduction in the sample size.

- **Less is more**. This strategy, articulated by Cohen (1990), suggests that you should always try to use the simplest type of analysis to test the hypothesis of interest. Unless you are specifically dealing with an exploratory analysis, you should try to limit the number of IVs that you use in a given regression

analysis. This advice, however, seems to conflict with one of the assumptions of regression itself, namely that your model includes all significant predictors of the DV. The way to balance these two requirements is to always include any well-known predictors in your statistical model, but to avoid including exploratory IVs when testing the major hypotheses of interest. If you perform a separate exploratory analysis, you can then limit the variables you include in your inferential analyses to those with strong relations.

This principle can also be generalized to "simple is better." In this form it applies well beyond regression analysis. When writing a paper, you should present it in such a way that it could be easily understood by the least-educated member of your intended audience. When designing an experiment, you should try not to manipulate too many things in a single study since the added variance can make it difficult to detect effects. When graphing your data, choose the form that demonstrates the relation that you want to present as simply as possible. Many graphing programs now provide 3D variations on traditional 2D graphs. Often times, however, the 3D versions are actually less clear than the 2D equivalent.

- **Least is last**. If you do decide to combine some data exploration with analyses designed to test your central research hypotheses, you should make use of sequential regression. If you order your sets so that the IVs associated with your central research question are tested before the IVs that are purely exploratory in nature, then the statistics from which you draw your inferences will not be complicated by the inclusion of the extra variables. It can also be very useful to use sequential regression to test more important hypotheses prior to examining less important hypotheses so that any collinearity between the central and the exploratory variables does not reduce the power of the more important tests.

- **Use full model tests to protect against error inflation**. Whenever you perform a number of tests examining the same hypothesis, the actual probability of any one of those tests being significant due to random chance alone will actually be greater than the established confidence level. Statisticians have therefore developed a number of methods that allow you to modify the results of your individual tests so that the overall experimentwide error rate stays below the confidence level. The simplest of these methods is to apply a *Bonferroni correction* by dividing the confidence levels of the individual tests by the total number of tests that you perform investigating the same hypothesis. In multiple regression we have reason to consider the tests of the different IVs as all examining the same hypothesis because they are all being used to predict the same DV. Therefore, some researchers have suggested that you should apply a Bonferroni correction to the tests of your individual coefficients by dividing the confidence level by the total number of IVs.

  There is, however, a different procedure that allows more powerful tests of your coefficients and provides a closer match between the established confidence level and the true probability of obtaining a significant result due to chance alone (Carmer & Swanson, 1973). This method, called a *protected t-test*, calls for us to first examine the overall F test whenever we fit a regression model. If the overall test is significant, we then go on to examine the significance of individual coefficients in that model using the standard confidence level. If the overall F is not significant, we stop without considering the tests of the individual coefficients.

  This idea of using a general test to provide protection for more specific analyses can be applied in a number of different areas. At a broader level, a significant multivariate analysis can be used as justification for examining the effect of an IV on a set of individual DVs (Rencher & Scott, 1990). At a more specific level, a significant F test in an ANOVA can be used as a justification for examining the differences between pairs of means (Fisher's LSD test).

Chapter 6

Dealing with Nonlinear Relationships

6.1 Using linear regression to model nonlinear relations

- Linear regression determines the best line that can express the relation between the IV and the DV. Even when you are working with multiple regression, tests of your coefficients tell you about the linear relation between the independent part of the IV with the independent part of the DV. However, there are certain types of nonlinear relations that can actually be examined using linear regression. The key is to find some function of the DV that has a linear relation to some function of the IV. If we can show that there is a linear relation between a function of the IV and a function of the DV, then we have simultaneously demonstrated that the IV and the DV are actually related to each other in a way that is determined by the functions that we used.
- The reason why this works is that linear regression only requires that the *parameters* be linear, and not the variables themselves. It doesn't matter if the values of the Xs and the Y in your regression are a function of other variables. As long as there is a linear relation between them then it can be estimated. The relation between your IV and your predictor variables and the relation between your DV and the response variable *does* influence the way that you interpret the coefficients, but it does *not* affect your ability to estimate those coefficients using least-squares linear regression.
- In this chapter we discuss a number of ways that a complex relation between your DV and an IV can actually be reinterpreted as one or more linear relations that can then be examined using the linear regression procedures we have discussed so far.

6.2 Nonlinear transformations

- Sometimes your IV and DV do not have a linear relation in their raw forms, but there is a linear relation after you transform one or both of them. In such a case, you can use linear regression to estimate the strength of the linear relation between the transformed variables. A test of the linear regression coefficients from this model will then act as a test of the nonlinear relation between the two original variables.
- Below is a list of some of the more common transformations that can be applied to your variables and suggests as to when they would most likely be relevant. The formulas for the transformations are all written in terms of X, although they can be applied to either the IV or the DV.
    - *Logarithm* ($\log[X]$). Logarithmic transformations are most often applied when the increases in one of your variables are related to the proportion by which another variable is increased. For example, if the value of your DV is increased by 1 every time you double the value of an IV, then you should apply a logarithmic transformation of the IV to make the relation linear. When examining the effect of practice it is typically necessary to perform a logarithmic transformation on both performance and time variables to obtain a linear relation between them (Newell & Rosenbloom, 1981).

It is also possible that proportionate increases in your IV be related to proportionate increases in your DV. For example, your DV may double every time your IV doubles. In this case you could linearize the relation by applying logarithmic transformations to both the IV and the DV.

There actually is not a single logarithmic transformation. Every logarithm is taken with reference to a "base" number. Mathematically, the logarithm tells you what power you have to raise the base to in order to get the observed number. The most common bases are 10 or the mathematical constant e (equal to approximately 2.7183). Logarithms with base e are called *natural logarithms*. However, none of this matters when transforming data since every logarithm will have the same linearizing effect on your variables (although the actual numbers they produce will vary).

- *Reciprocal* ($\frac{1}{X}$). Reciprocal transformations are used when the variable being transformed is expressed in terms of rates (how often something occurs in a given time period) while the other variable is related to the typical amount of time between occurrences.
- *Square root* ($\sqrt{X}$). A square root transformation is useful when the variable of interest represents the number of times a specific event occurs within a fixed time period. If the frequency of occurrences is low, Freeman and Tukey (1950) suggest using a transformation of $\sqrt{X} + \sqrt{X+1}$ instead of the standard square root because it is more likely to lead to homoscedasticity.
- *Arcsine* ($2 \sin^{-1}\sqrt{X}$). The arcsine transformation is typically applied when your variable represents a proportion. For this transformation, it is important that the arcsine is taken in radians rather than degrees (both are ways of measuring the size of an angle). You might also try using either *probit* or *logit* transformations on proportion data, as described on pages 241-244 of the textbook. However, the arcsine transformation is often preferred because it tends to stabilize the variance across the different levels of X.

- By changing the nature of the relation between your IV and your DV you will also change the distribution of your residuals. A lack of normality in your residuals can often be corrected by applying the appropriate transformation to your IV or DV. In fact, the patterns found in a plot of your residuals against an IV can not only indicate the need for a transformation, but can provide you with information about what transformation might work best.

- It is always best to have a historical or theoretical justification for any transformations that you might apply to your variables. However, you can also use exploratory techniques to determine what type of transformations you should perform on your data. Sometimes you can just look at a graph of your IVs or DV to determine what type of transformation would likely make the residuals more normal. Similarly, you might look at a graph of the residuals against your predicted values or your IVs to see if a transformation might make the distribution of your residuals more normal or their variance more equal across different levels of X. Whenever you obtain a scatterplot in ARC you can request "transformation sliders" by clicking the triangle next to **Options** and checking the box next to **Transform slidebars**. These allow you to immediately see what the scatterplot would look like if you performed a *power transformation* to either the variable on the horizontal or vertical axis. When you move the slider, the chart changes to represent what the graph would look like if you raised the corresponding variable to the value indicated by the slider. This allows you to accomplish a whole range of transformations, as illustrated in Figure 6.1.

Figure 6.1: Power transformations.

Power	Transformation
$p > 1$	X^p
$p = 1$	no transformation $\left(\text{since } X^1 = X\right)$
$0 > p > 1$	root transformation $\left(\text{since } X^{\frac{1}{p}} = \sqrt[p]{X}\right)$
$p = 0$	logarithmic transformation (conventionally assigned to $p = 0$)
$p < 0$	inverse transformation $\left(\text{since } X^{-p} = \frac{1}{X^p}\right)$

You can shift both sliders around on a scatterplot to determine which power transform would give the data the appearance you want. You can have ARC automatically find the transformation that would

make the dispersion of the graph around one axis best approach a normal distribution by clicking the triangle next to the appropriate axis and selecting **Find normalizing transformation**.

- The *Box-Cox transformation* provides a method for choosing the best transformation for your DV based directly on your data. Specifically, this transformation will compare the ability of various power transformations to linearize the relation between your IVs and the DV.

  You cannot perform a Box-Cox transformation in SPSS, but you can in ARC. To perform a Box-Cox transformation in ARC

  - Run a regression model using the **Graph&Fit** → **Fit linear LS** option.
  - Choose <**regression model**> → **Choose response transform** where <**regression model**> is the name you gave to the regression analysis you just ran.
  - Make sure the **Box-Cox Power** radio button is selected.
  - Click the **OK** button.

  ARC will then present a graph that presents the optimal power transformation for your DV next to the label **Lambda-hat**. More specific detail on how the Box-Cox transformation works can be found on pages 237-238 of your textbook.

- If you use one of the above methods to choose a transformation based on your data, you will usually get a value that does not correspond to a standard transformation. For example, the Box-Cox procedure might indicate that the optimal transformation for your DV would be to raise it to a power of .57. In such a case you might choose the standard transformation that most closely matches the exact power produced by the procedure. In this case we might just choose to do a square root transform (which would correspond to a power of .5) instead of actually raising the DV to the power of .57.

- To transform a variable in SPSS.

  - Choose **Transform** → **Compute**.
  - Type in the name of the variable you want to create in the box labeled **Target Variable**. The name should let you know both the name of the original variable as well as the function that you used.
  - Type the formula relating the new variable to the old one in the box labeled **Numeric Expression**. You can use the provided keypad if you like, but you can also just type the formula in directly.
  - Click the **OK** button.

  The new variable should then be added to your data set based on the function that you defined. The transformations discussed above can be implemented in SPSS using the formulas presented in Figure 6.2.

Figure 6.2: Transformation functions in SPSS.

Transformation	SPSS Formula
Raising a variable to the power p	$X**p$
Square root	sqrt(X)
Taking the rth root of a variable	$X**(1/r)$
Logarithm	ln(X) or lg10(X)
Reciprocal	$1/X$
Arcsine	2*arsin(sqrt(X))

- You should be cautious when applying transformations to your DV. Transforming your DV will not only affect the normality of your residuals, but will also differentially change the variance of your residuals across the different levels of your IVs. Used wisely, transformations of your DV can be

used to eliminate problems of heteroscedasticity. Used unwisely, you may inadvertently introduce heteroscedasticity that was not there originally. When you have a choice, you should therefore only transform your DV if you specifically want to change the variance of the residuals across the different levels of your IVs. Otherwise you should try to apply the opposite transformation to your IVs.

6.3 Polynomial regression

- Polynomial regression is actually just a type of general linear model where you take your terms to higher powers, such as squares (raising your IV to the power of 2) and cubes (raising your IV to the power of 3). It is different than simply performing a transformation on your variables in that you use a full polynomial equation to predict your DV.

- In polynomial regression you will often create multiple predictors from the same IV. For example, you might want to predict the value of your DV from both the raw IV as well as the square of the IV. In this case we refer to each of the predictor variables in your statistical model as a separate *term*, even if they are actually all just functions of the same IV.

- The power that you raise your IV to for a particular term is called the *order* of that term. "Lower-order" terms are those that are raised to low powers, while "higher-order" terms are those that are raised to high powers.

- You need to center your IVs (subtract off the mean from the value for each case) before performing a polynomial regression. Specifically, you should first center each IV, and then afterwards create any higher-order terms that you want to include in your model. This reduces the collinearity between terms that are derived from the same IV and makes the coefficients obtained for lower-order terms easier to interpret. Standardizing your IVs before creating your higher-order terms works just as well, since standardized variables are always centered.

- A simple type of polynomial regression is the *quadratic model* with one predictor:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i, \tag{6.1}$$

  where $i = 1 \ldots n$. This type of model can not only account for any linear trends in your data, but can also explain a U-shaped or inverted U-shaped curve.

- The best way to choose the highest order polynomial is through a historical or theoretical analysis. There are certain types of relations that are well-known to be fitted by quadratic or cubic models. You might also determine that a specific type of relation should exist because of the mechanisms responsible for the relation between the IV and the DV.

  If you are building your model in an exploratory fashion, however, you can estimate how high of an order function you should use by the shape of the relation between the DV and that IV. If your data appears to reverse p times (has p curves in the graph), you should use a function whose highest order parameter is raised to the power of $p+1$. In multiple regression you can see whether you should add an additional term for an IV by examining a graph of the residuals against the IV. Again, if the relation between the residuals and the IV appears to reverse p times, you should add terms whose highest order parameter is raised to the power of $p + 1$.

  You can also determine the highest-order term for your polynomial from a graph of the residuals from a regression model plotted against your IV. If your original regression model only contained a linear parameter but your relation was of a higher order, then you should see a curvilinear pattern in your residuals. If this curvilinear pattern appears to reverse p times, then the actual relation between your IV and the DV follows a polynomial function whose highest order parameter is raised to the power of $p+1$. It actually doesn't matter what the exact relation is = a polynomial whose highest term is raised to the power of $p + 1$ can actually fit *any* pattern that has p turns in it. If you have a number of other IVs in your model it can often be easier to see the presence of a higher-order relation in the residuals than in the raw values of the DV, since the residuals do not include the influences of the other IVs.

- Note that you always include the lower-order terms in your model. So if you want to test a polynomial whose highest order term is X^k, you should also include factors X^1 to X^{k-1} in your model. If you do not include all of the lower-order terms, any tests of higher-order terms will not reflect what you expect them to because they will also include variability that should be accounted for by the missing terms. In sequential regression, the lower-order terms should always be entered in your model before or at the same time as a higher-order term from the same IV.

- You can also have polynomial models with more than one IV, such as

$$Y_i = \beta_0 + \beta_{11}X_{i1} + \beta_{12}X_{i1}^2 + \beta_{21}X_{i2} + \beta_{22}X_{i2}^2 + \epsilon_i, \tag{6.2}$$

  where $i = 1 \ldots n$.

- You can only make very limited inferences based on the coefficients for the individual terms in a polynomial model. Taken by themselves, the estimates do not have great meaning. For example, consider the following estimated cubic model.

$$Y_i = b_0 + b_1X_i + b_2X_i^2 + b_3X_i^3. \tag{6.3}$$

- The coefficients in equation 6.3 can be interpreted in the following way.
  - b_0 is the expected value of Y when X = 0.
  - b_1 is the slope of the estimated function when X = 0.
  - b_2 is the acceleration (change in slope) of the estimated function when X = 0.
  - b_3 represents the strength of the cubic relation between X and Y.

  When your IVs are uncentered, the first three coefficients represent things that are not usually of great interest to social scientists. However, when you center your IVs, the value 0 refers to the mean of the IV. In this case, your lower-order coefficients can be interpreted in two meaningful ways.

  1. b_1 is the estimated slope at the mean value of the IV, which is also the average slope across the entire estimated function.
  2. b_2 estimates the strength of a quadratic relation at the mean value of the IV, which is also the strength of a quadratic relation across the entire estimated function.

  The test of b_3 is a direct test of the cubic relation between X and Y, independent of the other terms in the model. The test of this parameter specifically tells you whether the cubic term adds a significant amount to prediction above and beyond that of the quadratic equation (consisting of the intercept, linear, and quadratic terms). This can be useful because it specifically informs you as to which of two models (one including the cubed term and one that does not) best fits the data. If your test for b_3 was not significant you would likely remove the cubed term from the model and retest it again using only the linear and quadratic terms.

- When performing polynomial regression, you may want to consider the ability of the full polynomial function to account for variability in your DV. In this case you can consider the terms derived from the same IV as an IV set and test their collective ability to predict the DV using the procedures described in section 5.3. The test of the IV set will tell you whether or not your IV has an overall influence on the DV.

- You should be very careful about trying to use polynomial models to predict values of the DV beyond the range of values from which the model was estimated. High-order polynomials can have sudden changes of slope and can even reverse their direction. In such a case the predicted values will likely have little to do with the actual values of the DV.

- To perform a polynomial regression in SPSS
  - Determine the highest order term that you will use for each IV.
  - Center any IVs for which you will examine higher-order terms.

- For each IV, create new variables that are equal to your IV raised to the powers of 2 through the power of your highest order term. Details on constructing new variables using formulas are presented in section 6.2. Be sure to use the centered version of your IV.
- Conduct a standard multiple regression including all of the terms for each IV.
- If you want, you can also use the procedures for testing IV sets discussed in section 5.3 if you want to determine the total influence of an IV on the DV.

6.4 Nonlinear and nonparametric regression

- We have just spent this chapter discussing how you can use transformations and polynomial regression to analyze data with nonlinear relations using linear regression. However, statisticians have also developed a number of analytic techniques that can be used to test specific types of nonlinear relations. For example, *logistic regression* is a procedure that has been developed to analyze models with categorical DVs, and *Poisson regression* is a procedure that has been developed to analyze models with count DVs. When such procedures exist they will typically provide more accurate and more powerful tests of the relations between your variables compared to performing an analysis in linear regression on a transformed variables.
- There is also a new class of procedures referred to as *nonparametric regression* that specifically do not make *any* assumptions about the nature of the relations between your IVs and your DV. These analyses are often based on the fit of a lowess line instead of trying to fit the data to a specific line or curve. However, the work on nonparametric regression is still in its initial stages, so it will likely be some time before they can be easily used by basic practitioners to test hypotheses about their data.

Chapter 7

Interactions Among Continuous IVs

7.1 Main effects and interactions

- In Chapter 2 we discussed simple linear regression, where we predicted a single DV from a single IV. We followed this up in Chapter 3 with a discussion of multiple regression, where we predicted a single DV from the combination of two or more IVs. The coefficients on the IVs in both of these cases represent the *main effects* of the IVs in question. A main effect refers to the influence of an IV on the DV averaging over the values of other IVs in the statistical model.

  In this chapter we will discuss the possibility of testing for *interactions* among your IVs. An interaction measures the extent to which the relation between one IV and a DV depends on the level of other IVs in the model. If you have an interaction between two IVs (called a *two-way interaction*) then you expect that the relation between the first IV and the DV will be different across different levels of the second IV.

- Interactions are symmetric. If you have an interaction such that the effect of IV1 on the DV depends on the level of IV2, then it is also true that the effect of IV2 on the DV depends on the level of IV1. It therefore does not matter whether you say that you have an interaction between IV1 and IV2 or an interaction between IV2 and IV1.

- You can also have interactions between more than two IVs. For example, you can have a *three-way interaction* among IV1, IV2, and IV3. This would mean that the two-way interaction between IV1 and IV2 depends on the level of IV3. Just like two-way interactions, three-way interactions are also independent of the order of the variables. So the above three-way interaction would also mean that the two-way interaction between IV1 and IV3 is dependent on the level of IV2, and that the two-way interaction between IV2 and IV3 depends on the level of IV1.

- It is possible to have both main effects and interactions at the same time. For example, you can have a general trend that the value of the DV increases when the value of a particular IV increases, but that the relation is stronger when the value of a second IV is high than when the value of that second IV is low. You can also have lower order interactions in the presence of a higher order interaction. Again, the lower-order interaction would represent a general trend that is modified by the higher-order interaction.

- An interaction effect is *not* the same thing as the effect of multicollinearity on your coefficients. In multiple regression your coefficients represent the relation between the part of each IV that is independent of the other IVs in the model with the part of the DV that is independent of the other IVs in the model. While this does mean that the coefficient that you get for a given IV will be dependent on the other IVs in the model, in a standard multiple regression you would still expect that coefficient to express the relation between the IV and the DV no matter what other values your other IVs may have. The important distinction is that while the relation between a given IV and the DV is always dependent on the other *variables* in the model, when you have an interaction the relation between the IV and the DV is also dependent on the particular *values* that the other variables have.

7.2 Regression models with two main effects and one interaction

- You can use linear regression to determine if there is an interaction between a pair of IVs by adding an *interaction term* to your statistical model. To detect the interaction effect of two IVs (X_1 and X_2) on a DV (Y) you would use linear regression to estimate the equation

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i1} X_{i2}, \tag{7.1}$$

  where $i = 1 \ldots n$. You construct the variable for the interaction term $X_{i1}X_{i2}$ by literally multiplying the value of X_1 by the value of X_2 for each case in your data set.

- If the b_3 is significantly different from zero in the model presented in equation 7.1 then you have a significant interaction effect. However, for this to be true you must include both the main effects for X_1 and X_2 in the model. If you were to estimate a model such as

$$Y_i = b_0 + b_3 X_{i1} X_{i2}, \tag{7.2}$$

  then the test of b_3 would actually confound the interaction effect with the main effects of X_1 and X_2, and so would not be readily interpretable.

- In the model represented in equation 7.1, the expected change in Y with a unit increase in X_1 when X_2 is held constant is $b_1 + b_3 X_2$. Similarly, the expected change in Y with a unit increase in X_2 when X_1 is held constant is $b_2 + b_3 X_1$. This parallels what we said before about interaction effects: The effect of each IV on the DV depends on the level of the second IV in the interaction.

- The remaining coefficients from equation 7.1 can be interpreted in the following way.

  - b_0 is the expected value of Y when $X_1 = 0$ and $X_2 = 0$.
  - b_1 is the slope of the relation between X_1 and Y when $X_2 = 0$.
  - b_2 is the slope of the relation between X_2 and Y when $X_1 = 0$.

  When your IVs are uncentered (meaning when their means are not equal to zero), these coefficients are rarely of much interest. If X_1 and X_2 are centered (performed by subtracting the mean from each observation), however, then the test of b_1 tells you if there is a main effect of X_1 and the test of b_2 tells you if there is a main effect of X_2. For this reason it is suggested that you always center your IVs when testing for interactions in regression.

- In addition to providing you with tests of the main effects of your IVs, centering your IVs will also reduce the collinearity between the main effects and the interaction term. Just like having an uncentered predictor can inflate the collinearity between the terms in a polynomial regression, having an uncentered predictor will also inflate the collinearity between your main effects and your interaction term.

  It is not necessary to center your DV when testing for an interaction. Having an uncentered DV will not affect either your ability to test for main effects of your IVs or the amount of collinearity between the main effect and interaction terms.

- When you are building your interaction term you should first center each of your IVs and then multiply the centered variables together to get the variable for the interaction term. This does *not* give you the same results as multiplying your variables together first and then centering the resulting product.

7.3 Exploring a significant interaction

- Once you determine that there is an interaction between a pair of IVs, you will typically want to graphically display the relations among these three variables so that you can explain the effect. One way to do this is by examining a table of means for *dichotomized* versions of your variables. When you dichotomize a variable you basically create two groups, one of which contains cases that have high values and one of which contains cases that have low values. You can choose to make this split on either the mean or the median, although the median is commonly preferred because it is more likely to give you groups of equal size. To dichotomize a variable in SPSS you would take the following steps.

- Choose **Analyze → Descriptives → Explore**.
- Move the variable you want to dichotomize to the box labeled **Dependent List**.
- Click the **OK** button. The output from this procedure will contain both the mean and the median of the variable.
- Choose **Transform → Recode → Into different variables**.
- Move the variable to the **Input Variable -> Output Variable** box.
- Type the name of the variable you want to hold the dichotomized values in the **Name** box beneath **Output Variable**.
- Click the **Change** button.
- Click the **Old and New Values** button.
- Click the radio button on the left-hand side next to **Range: Lowest through**.
- Enter the mean or median into the box next to **Lowest through**.
- Type the number **-1** in the box next to **Value** on the right-hand side.
- Click the **Add** button.
- Click the radio button on the right-hand side next to **All other values**.
- Type the number **1** in the box next to **Value** on the left-hand side.
- Click the **Add** button.
- Click the **Continue** button.
- Click the **OK** button.

Once you have dichotomized your two IVs you can then obtain a table of the means for each combination of your IVs by taking the following steps.

- Choose **General Linear Model → Univariate**.
- Move your DV to the box labeled **Dependent Variable**.
- Move your two dichotomized variables to the box labeled **Fixed Factor(s)**.
- Click the **Options** button.
- Move the term related to the interaction of your two dichotomized variables from the **Factor(s) and Factor Interactions** box to the **Display Means for** box.
- Click the **Continue** button.
- Click the **OK** button.

This procedure will provide you with the mean value of your DV for people within each possible combination of high vs. low on your two IVs in the section labeled **Estimated Marginal Means**. It will also produce an analysis actually predicting the value of your DV from the values on your two dichotomized variables. However, you should not use or report the statistics presented here. When you have a truly continuous measure you should *never* use dichotomies based on those variables in statistical analysis. Instead you should always report the statistics from the regression analysis using the centered continuous variable. Research has consistently demonstrated that analyses using dichotomized variables will have significantly less power and can actually lead to falsely significant tests under certain circumstances (MacCallum, Zhang, Preacher, & Rucker, 2002).

- Another way to explore an interaction is to plot a graph of the regression equations predicting the DV from IV1 at different levels of IV2. This method is actually preferred because it provides more information than just the cell means. These equations, called the *simple slopes*, can be determined just by substituting in specific values of X_2 into equation 7.1. For example, let us say that the least squares regression equation relating a DV to two IVs is

$$Y_i = b_0 + b_1 X_{i1}^c - b_2 X_{i2}^c + b_3 X_{i1}^c X_{i2}^c, \tag{7.3}$$

where X_{i1}^c is the centered version of X_{i1} and X_{i2}^c is the centered version of X_{i2}. We can determine the simple slope relating X_1^c to Y at a given value of X_2^c by simply substituting the corresponding value of X_2^c in the above equation.

- Commonly people will choose to display the slopes where X_2^c is equal to its mean, the mean plus 1 sd, and the mean minus 1 sd. Looking at these three slopes you can usually get a decent picture of how changes in the value of X_2^c affects the relation between Y and X_1^c. Since centered variables have a mean of zero, you can obtain these equations by substituting 0, -sd, and +sd for X_2^c in formula 7.3.

  Although we frame this example in terms of seeing how X_2^c affects the simple slopes for X_1^c, we could just as easily have chosen to plot how X_1^c affects the simple slopes for X_2^c because of the symmetric nature of interactions.

- Since X_1^c and X_2^c are simply the centered versions of X_1 and X_2, the plot of the simple slopes for equation 7.3 will also conceptually represent the interaction between the original, uncentered variables. The only difference would be that the axes have different scales, which should not influence your interpretation of the interaction.

- Aiken and West (1991) show that for a 2-way interaction, the standard error of the simple slope relating X_1 to Y when X_2 is held constant at the value Z can be calculated using the equation

$$s_b = \sqrt{s_{11} + 2Zs_{13} + 2Z^2 s_{33}}, \tag{7.4}$$

  where s_{11} is the variance of the coefficient for X_1, s_{13} is the covariance between the regression coefficient for X_1 and the regression coefficient for the X_1X_2 interaction, and s_{33} is the variance of the regression coefficient for the X_1X_2 interaction. These variances and covariances do not appear as part of the standard SPSS output, but can be obtained clicking the in the **Statistics** button and then checking the box next to **Covariance matrix** in the linear regression menu.

  After obtaining the standard error for the simple slope, you can use it to perform a hypothesis test of a point estimate testing whether the simple slope is significantly different from zero, or you can use it to calculate a confidence interval around the simple slope estimate.

- The staff of Stat-Help.com have created a Microsoft Excel spreadsheet that will graph the simple slopes for a two-way interactions, and test whether each is significantly different from zero. This spreadsheet is available online at

  `http://www.stat-help.com/spreadsheets.html`

- You can have SPSS generate a graph illustrating the simple slopes of X_1^c at different levels of X_2^c. However, the process is somewhat complicated, so if you have access to Microsoft Excel you will likely find it easier to graph your interactions using the spreadsheet described above. To graph a two-way interaction in SPSS you would take the following steps.

  - Determine the mean and the standard deviation of X_2. You can obtain these using the **Analyze** → **Descriptives** procedure in SPSS.
  - Perform a standard regression analysis testing the interaction you want to examine. You should be sure to use the centered versions of all of your IVs in the equation.
  - Determine the equations for the simple slopes between X_1^c and Y, where X_2^c is equal to 0, +sd, and -sd as described above.
  - For each equation, create a new variable to represent its predicted values by taking the following steps in SPSS.
    - Choose **Transform** → **Compute**.
    - Type the variable to hold the predicted value in the box labeled **Target Variable**.
    - Type the equation into the box labeled **Numeric Expression**.
    - Click the **OK** button.
  - Display an overlayed scatterplot graphing the value of the predicted value by the value of X_1^c for each of the three equations by taking the following steps.
    - Choose **Graphs** → **Scatter** → **Overlay**.
    - Click the **Define** button.
    - Click the variable corresponding to X_1^c followed by the variable corresponding to the predicted value from the first equation.

    - Move the pair of variables to the **Y-X Pairs** box.
    - Click the variable corresponding to X_1^c followed by the variable corresponding to the predicted value from the second equation.
    - Move the pair of variables to the **Y-X Pairs** box.
    - Click the variable corresponding to X_1^c followed by the variable corresponding to the predicted value from the third equation.
    - Move the pair of variables to the **Y-X Pairs** box.
    - Make sure that the pairs are all organized such that the variable on the left is the predicted value and the variable on the right is X_1^c. SPSS always puts the variable that is first alphabetically on the left, which may not be what you want. In this case you should select any misaligned pairs and click the **Swap Pair** button.
    - Click the **OK** button.

  The final graph will display the value of X_1^c on the X-axis and the predicted value of Y on the Y-axis for each of the three different levels of X_2^c. However, the graph will display a sequence of dots instead of straight lines. To modify the graph so that it displays lines you can take the following steps.

    - Double-click the graph in the output window to open the Chart Editor.
    - Select one of the markers in the legend. These are the colored boxes next to the line descriptions.
    - Choose **Format** → **Interpolation**.
    - Click the box next to **Straight**.
    - Uncheck the box next to **Display markers**.
    - Click **Apply All**.
    - Click **Close**.
    - Close the Chart Editor.

- There are two common patterns of interactions that have important conceptual meanings.

  1. If the signs of b_1, b_2, and b_3 are all the same then you have a *synergistic interaction*. In this case the benefit or detriment of having strong values on both of your IVs exceeds individual effects of the IVs. This conceptually maps onto phenomena where you need to have a certain level of one IV to take advantage of the benefits of a high value on the second IV.
  2. If the signs of b_1 and b_2 are the same but b_3 has the opposite sign the then you have an *antagonistic interaction*. In this case, the effect of having a high value on one of the IVs is reduced when the value on the other IV is also high. This conceptually maps onto phenomena where high values on one IV are providing a benefit on the DV that is at least partially redundant with the effect of having a high value on the second IV.

- It is important to note that you typically need a larger sample size to detect an interaction effect than you would to detect an equally strong main effect. The reason is that the reliability of the interaction term is equal to the product of the reliabilities of the two IVs. Since reliabilities are always between 0 and 1, this means that the reliability of the interaction term will often be substantially lower than the reliabilities of the original IVs.

7.4 Interactions in more complicated models

- As mentioned in section 7.1, you can also have interactions between three different IVs. To detect such an interaction you would use linear regression to estimate the equation

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + b_4 X_{i1} X_{i2} + b_5 X_{i1} X_{i3} + b_6 X_{i2} X_{i3} + b_7 X_{i1} X_{i2} X_{i3}, \tag{7.5}$$

  where $i = 1 \ldots n$. The variables for the two-way interactions ($X_{i1}X_{i2}$, $X_{i1}X_{i3}$, and $X_{i2}X_{i3}$) and the three-way interaction ($X_{i1}X_{i2}X_{i3}$) can all be determined by simply multiplying together the appropriate IVs.

- According to the model in equation 7.5, we would expect that a one-unit increase in X_1 will change the value of Y by $b_1 + b_4X_2 + b_5X_3 + b_7X_2X_3$, assuming that X_2 and X_3 are held constant. We would expect that a one-unit increase in X_2 will change the value of Y by $b_2 + b_4X_1 + b_6X_3 + b_7X_1X_3$, assuming that X_1 and X_3 are held constant. Finally, we would expect that a one-unit increase in X_3 will change the value of Y by $b_3 + b_5X_1 + b_6X_2 + b_7X_1X_2$, assuming that X_1 and X_2 are held constant. From these we can see that the expected change in Y due to a change in one IV not only depends on the values of the other IVs independently, but also on their particular combination (represented by the coefficient b_7).

- The coefficient b_7 in equation 7.5 provides a test of the three-way interaction between variables X_1, X_2, and X_3, but only when all of the *lower-order effects* are also included in the model. The lower-order effects for a three-way interaction include the main effects of the three IVs involved in the interaction as well as all of the possible two-way interactions between those IVs.

- The recommendation for using centered IVs when testing for interactions holds no matter how many IVs are involved in the interaction you are trying to detect. If you use centered main effects and construct the interaction terms by multiplying together the centered values, you will greatly reduce the collinearity between your main effects and your interactions and will also make the coefficients on your lower-order effects interpretable. When you center your variables, the other coefficients in equation 7.5 can be interpreted as follows.

  - b_0 is the expected value of Y at the mean of all of the IVs.
  - b_1 provides a test of the main effect of X_1.
  - b_2 provides a test of the main effect of X_2.
  - b_3 provides a test of the main effect of X_3.
  - b_4 provides a test of the interaction between X_1 and X_2, averaging over the different levels of X_3.
  - b_5 provides a test of the interaction between X_1 and X_3, averaging over the different levels of X_2.
  - b_6 provides a test of the interaction between X_2 and X_3, averaging over the different levels of X_1.

- Once you determine that there is a three-way interaction between your IVs you will want to perform follow-up analyses to determine the nature of the interaction. First, you can dichotomize your IVs and then look at a table of means by each of your IVs. You would follow the same basic procedure described in section 7.2 except that you would need to dichotomize all three of your IVs and you would need to obtain the marginal means for the three-way interaction of the ANOVA. Second, you could examine graphs of the simple slopes relating your IVs to the DV. The simple slopes for a three-way interaction show the relation between one of the IVs and the DV and specific values of the other two IVs involved in the interaction. You can obtain these simple slopes by substituting the values of the fixed IVs into the original regression equation, just like you would for a two-way interaction.

- Aiken and West (1991) show that for a three-way interaction, the standard error of the simple slope relating X_1 to Y when X_2 is held constant at the value Z and X_3 is held constant at the value of W can be calculated using the equation

$$s_b = \sqrt{\begin{array}{l} s_{11} + Z^2 s_{44} + W^2 s_{55} + Z^2 W^2 s_{77} + 2Z s_{14} + 2W s_{15} + \\ 2ZW s_{17} + 2ZW s_{45} + 2Z^2 W s_{47} + 2ZW^2 s_{57} \end{array}} \tag{7.6}$$

  where s_{11} is the variance of the coefficient for X_1, s_{44} is the variance of the regression coefficient for the X_1X_2 interaction, s_{55} is the variance of the regression coefficient for the X_1X_3 interaction, s_{77} is the variance of the regression coefficient for the $X_1X_2X_3$ interaction, s_{14} is the covariance between the regression coefficient for X_1 and the regression coefficient for the X_1X_2 interaction, s_{15} is the covariance between the regression coefficient for X_1 and the regression coefficient for the X_1X_3 interaction, s_{17} is the covariance between the regression coefficient for X_1 and the regression coefficient for the $X_1X_2X_3$ interaction, s_{45} is the covariance between the regression coefficient for the X_1X_2 interaction and the regression coefficient for the X_1X_3 interaction, s_{47} is the covariance between the regression coefficient for the X_1X_2 interaction and the regression coefficient for the $X_1X_2X_3$ interaction, and s_{57} is the covariance between the regression coefficient for the X_1X_3 interaction and the regression coefficient for the $X_1X_2X_3$ interaction. These variances and covariances do not appear as part of the standard SPSS

output, but can be obtained clicking the in the **Statistics** button and then checking the box next to **Covariance matrix** in the linear regression menu.

After obtaining the standard error for the simple slope, you can use it to perform a hypothesis test of a point estimate testing whether the simple slope is significantly different from zero, or you can use it to calculate a confidence interval around the simple slope estimate.

- When people want to graph the simple slopes for a three-way interaction, they usually create two graphs. The first contains the simple slopes for the two-way interaction between X_1 and X_2 for cases that are below the average value of X_3, while the second contains the simple slopes for the two-way interaction between X_1 and X_2 for cases that are above the average value of X_3.

- The staff of Stat-Help.com have created a Microsoft Excel spreadsheet that will graph the simple slopes for a two-way interactions, and test whether each is significantly different from zero. This spreadsheet is available online at

  `http://www.stat-help.com/spreadsheets.html`

- You can also plot the simple slopes from a three-way interaction using SPSS, but the procedure is much more complicated. If you have access to Microsoft Excel you will find that using the above spreadsheet to be the easiest way to explore your interaction. If you want to use SPSS to plot the simple slopes from a three-way interaction, however, you would take the following steps.

  - Create a dichotomized version of X_3 as described in section 7.3.
  - Choose **Data** → **Split file**.
  - Click the radio button next to **Organize output by groups**.
  - Move the dichotomized version of X_3 to the box labeled **Groups based on**.
  - Click the **OK** button.
  - Perform a regression analysis predicting the value of the DV from X_1, X_2, and the X_1X_2 interaction term. You should not include X_3 or any interactions with X_3 in the model.
  - SPSS will produce the output from two different regression models: one only including people with high values of X_3 and one only including people with low values of X_3.
  - Choose **Data** → **Split file**.
  - Click the radio button next to **Analyze all cases, do not create groups**.
  - Click the **OK** button.
  - Use the output of the regression analyses to create six variables, representing the simple slopes both under high values of X_3 and under low values of X_3.
  - Tell SPSS to create a simple slopes plot for the two-way interaction between X_1 and X_2 under high values of X_3, as described in section 7.2. Do this a second time to obtain a simple slopes plot for the two-way interaction between X_1 and X_2 under low values of X_3. You can use the full data set when generating these graphs (you don't have to separate it by the values of X_3).

- The logic presented here to test a three-way interaction can be extended to test interactions between four, five, or any number of variables. In each case you must make sure that your statistical model contains all of the lower-order effects that lead up to the interaction. So, to test for a four-way interaction you need to make sure that your model contains the main effects for all of the IVs involved in the interaction as well as all possible two-way and three-way interactions between those variables.

  Even though you can look for interactions between any number of variables, it is rare for researchers in the social sciences to look for interactions higher than a three-way interaction. First, the reliability of your interaction term is equal to the product of the reliabilities of all of the IVs composing it. This means that increasing the number of variables in your interaction decreases the reliability of the interaction term, which will in turn make it more difficult to find a significant relation between that term and the DV. Additionally, the more variables you have in your interaction the more complicated will be the relations between your variables. It is very difficult to explain patterns represented by a four-way or higher order interaction.

- It is possible to interact one IV with a polynomial function of another IV. For example, you might suspect that the degree of curvature in a quadratic model may depend on the level of another IV. However, it is inappropriate to include an interaction of one IV with a higher-order polynomial term in a statistical model without also including the interaction of that IV with all of the lower-order polynomial terms. For example, to test whether there is an interaction between the variable X_1 and X_2^2 we would need to use a model such as

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i2}^2 + b_4 X_{i1} X_{i2} + b_5 X_{i1} X_{i2}^2, \tag{7.7}$$

  where $i = 1 \ldots n$. The coefficient b_5 will only represent the interaction between the linear effect of X_1 and the quadratic effect of X_2 when you include all of the main effects and the lower-order interactions in the model. Otherwise b_5 will represent a combination of the desired interaction and these missing terms.

- You can include tests of interaction effects in models that also contain variables unrelated to the interaction. The inclusion of these additional terms does not affect the estimation of your main effects or interactions, except possibly through multicollinearity. You can even test for the effect of two unrelated interaction effects in the same model, as in the estimated regression equation

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i1} X_{i2} + b_4 X_{i3} + b_5 X_{i4} + b_6 X_{i3} X_{i4}, \tag{7.8}$$

  where $i = 1 \ldots n$.

- You can use a simple slopes plot to explore the nature of an interaction, even when there are other IVs in the model. In this case all you need to do is assume that all of the IVs that are not involved in the interaction are held constant at their mean values when determining the lines for your graph. This will allow you to see the effect of the interaction averaging over all of the other variables in the model.

- When analyzing models that contain a large number of interaction terms it can become tedious to manually construct all of the product terms. Luckily, if you choose to analyze your data using SPSS's general linear model procedure it can create these interaction terms for you. However, you still need to center all of your original IVs before using the procedure. To analyze a regression model this way in SPSS

  - Choose **Analyze → General Linear Model → Univariate**.
  - Move your DV to the box labeled **Dependent Variable**.
  - Move all of the main effect terms for your IVs to the box labeled **Covariate(s)**.
  - Click the **Options** button.
  - Check the box next to **Parameter estimates**. By default this procedure will only provide you with tests of your IVs and not the actual parameter estimates.
  - Click the **Continue** button.
  - By default SPSS will not include interactions between continuous variables in its statistical models. However, if you build a custom model you can include whatever terms you like. You should therefore next build a model that includes all of the main effects of your IVs as well as any desired interactions. To do this
    - Click the **Model** button.
    - Click the radio button next to **Custom**.
    - Select all of your IVs, set the drop-down menu to **Main effects**, and click the arrow button.
    - For each interaction term, select the variables involved in the interaction, set the drop-down menu to **Interaction**, and click the arrow button.
    - If you want all of the possible two-way interactions between a collection of IVs you can just select the IVs, set the drop-down menu to **All 2-way**, and click the arrow button. This procedure can also be used to get all possible three-way, four-way, or five-way interactions between a collection of IVs by setting the drop-down menu to the appropriate interaction type.

- Click the **Continue** button.
- Click the **OK** button.

- The SPSS output from running an analysis using the General Linear Model contains the following sections.
  - **Between-Subjects Factors**. This table just lists out the different levels of any categorical variables included in your model.
  - **Tests of Between-Subjects Effects**. This table provides an F test of each main effect or interaction that you included in your model. It indicates whether or not the effect can independently account for a significant amount of variability in your DV. This provides the same results as testing the change in model R^2 that you get from the test of the set of terms representing the effect.

7.5 Interactions with IV sets

- In section 5.3 we talked about how you can determine if a specific collection of IVs can jointly account for a significant amount of variance in your DV. It is also possible to test whether the effect of one set of IVs interacts with another set of IVs.

- The interaction between two IV sets is represented in your regression model as a third set of IVs. This interaction set consists of the interaction between each IV in the first set with each IV in the second set. This means that if your first set contains j variables and your second set contains k variables, the interaction set would contain jk variables.

- As an example, let us assume that we want to determine the interaction between IV set M, consisting of variables M_1, M_2, and M_3, and IV set N, consisting of variables N_1 and N_2. In this case, the interaction set would consist of the variables M_1N_1, M_1N_2, M_2N_1, M_2N_2, M_3N_1, and M_3N_2.

- You can test for the effect of an interaction set in the same way that you test for the effect of any other set of variables, as described in section 5.3. In the formulas presented there, the terms in the interaction set would be set B, while all the other variables in the model would be in set A. However, you must make sure that the statistical model in which you are testing the interaction also includes the two sets of variables that compose the interaction. For example, if you wanted to test the interaction between sets M and N defined above, you must test whether the six interaction variables can account for a significant amount of variability in the DV beyond that accounted for by the five main effect variables.

- When testing the interaction between two IV sets you should center the IVs in those sets and construct your interaction terms by multiplying together your centered variables. Centering will reduce the collinearity between the main effect and interaction sets, and will also provide meaning to the tests of the main effects.

- If you center the variables in the original sets, the test of each of the original sets when you include the interaction set in the model will represent the main effect of those sets on the DV. Specifically, it will represent the average effect of each set on the DV.

- The presence of an interaction effect between two IV sets indicates that the interactions between the variables in the two sets can account for a significant amount of variability in the DV.

- It is possible that you might have an interaction between one IV and the polynomial function of another IV. You can conceptualize this as an interaction between IV sets. In this case one of your sets is a single IV, and the other set contains the terms from the polynomial function of another IV. It is even possible to have interactions between the polynomial functions of two IVs. In this case each of the polynomial functions would be considered to be its own set.

Chapter 8

Regression with Categorical IVs

8.1 Representing categorical variables in regression

- In this chapter we will discuss how you can perform a regression analysis that includes categorical IVs. Even if you code the different groups using different numbers, it is typically not acceptable to directly include a categorical variable in a regression analysis. For example, let us say that you had a variable representing the race of a respondent where whites were coded as 1, blacks were coded as 2, and all other races were coded as 3. It is certainly possible to enter this variable into a regression analysis, but the results would not have any meaning. Let us then say that you performed a regression analysis predicting a DV from this IV and found that the regression coefficient was significantly greater than zero. The coefficient in linear regression measures the linear relation between your variables, so a positive coefficient would mean that as people had higher values on your IV they tended to have higher values on the DV. This could either mean that blacks had higher values than whites, other races had higher values than whites, or that other races had higher values than blacks. It is impossible to tell from this analysis which of these is true.

- However, consider the case where you have a categorical variable that only has two groups. Let us say that you had a variable that had the value 1 when your participant's gender was male and 2 when your participant's gender was female. If you performed a simple linear regression analysis and determined that this variable had a positive relation with a DV, you actually *would* be able to interpret this. Specifically, this would mean that women had higher values on the DV than males. So, while a categorical variable that has more than three groups cannot be directly tested using regression, you can use regression to test a categorical variable that compares only two groups.

- We can use the fact that tests of categorical variables with two levels can be included in regression models to provide a method to test categorical variables with more than two groups. Specifically, if we can conceptualize the distinctions made by a categorical variable as a whole as a set of two-group comparisons, we can test the effect of the categorical variable using regression. For example, while we cannot directly test the race variable described above, we could independently test whether whites are different from blacks using one variable and whether whites are different from other races using a second variable. The set of comparison variables you use to represent a categorical IV are called its *codes*.

- Once we have created variables representing $g-1$ comparisons from a categorical variable with g levels (assuming that none are completely collinear with the other codes) we have fully represented that variable in our regression model. To test for an influence of the categorical IV on the DV you would perform a test of the influence of the entire set of codes, as discussed in section 5.3. If you try to include g different codes from a categorical variable with g levels in the same regression you will get an error because the extra code will be completely collinear with the other variables in the model. For example, in the race example above, if we knew that whites were 4 points higher than blacks and that whites were 3 points less than other races, we would know that blacks would have to be 7 less than other races. You can see this by setting up the simultaneous equations presented in Figure 8.1. Since the comparison of blacks with other races can be determined if we know the other two comparisons, then a variable that represents this comparison would be completely collinear with the other two variables.

Figure 8.1: Equations used to calculate the difference between blacks and other races.

Equation 1	$\bar{X}_w - \bar{X}_b = 4$
Equation 2	$\bar{X}_w - \bar{X}_o = -3$
Taking Equation 2 - Equation 1	$(\bar{X}_w - \bar{X}_o) - (\bar{X}_w - \bar{X}_b) = -3 - 4$
Resolving terms	$\bar{X}_b - \bar{X}_o = -7$

In fact, any comparison among the different levels of a categorical IV with g levels can be computed as a linear combination of the coefficients on $g - 1$ code variables, assuming none of the codes are redundant.

- While you can represent a categorical IV with any set of $g-1$ independent codes, there are few preferred ways of defining your code variables. The next section in these notes describes three different ways of defining your codes: dummy coding, effect coding, and contrast coding. The different coding methods all work equally well when testing the overall effect of your categorical IV. They differ only in the interpretations that you can give to the estimated coefficients for the code variables.

  It should be reiterated that you can actually use any set of $g - 1$ code variables (as long as none are completely collinear) to represent your categorical variable in a regression analysis. The way you define your codes does not have any influence on the overall test of the categorical IV, which is accomplished by looking at the total effect of your set of code variables on the DV. However, your coding method *does* influence the interpretation of the estimated coefficients on your codes, so if you develop a nonsensical coding method then you will not be able to make any inferences based on the values of the coefficients. The methods we describe are preferred primarily because they result in coefficients that have specific, interpretable meanings.

8.2 Coding schemes

- For each of the coding schemes we will first explain how you would interpret the coefficients that you obtain using that system. We will then describe exactly how you create the code variables. When explaining the coding scheme we will show you how the codes would be defined for a categorical variable designed to represent which gulf state an individual is from, which can take on the values of Florida, Alabama, Mississippi, Louisiana, and Texas. We will assume that you wish to use regression analysis to determine the effect of state on the SAT college entrance exam.

- The first coding scheme is called *dummy coding*. Each dummy code corresponds to a specific group in your categorical IV. Note that since there are only $g - 1$ codes for a categorical variable with g groups, one of your groups will not be associated with a dummy code. This missing group is called the *reference group.* The coefficient on each dummy code obtained through least squares regression estimates the difference between the mean value of the DV for its corresponding group and the mean value of the reference group. A significant coefficient on a dummy code variable indicates that the corresponding group is significantly different from the reference group. Dummy coding is a good choice when you want to compare the mean values of the different levels of your categorical IV to a specific reference group. For example, if you have several different treatments plus a control group you will likely want to compare the value of the DV for each treatment to the control group.

  To create dummy codes you must first decide which of your variables will be the reference group. This will generally be the group that is of greatest interest since the coefficients will each estimate between the other groups and the reference group. Once you determine your reference group, you define each code as corresponding to one of the *other* groups of your categorical variable. The code will have a value of 1 if a given case is a member of the corresponding group and will have a value of 0 if it is not. Specifically note that you do not have a code for the reference group.

  Taking the example of the gulf state variable, let us assume that we want to use Alabama as the reference group. Figure 8.2 illustrates the values the code variables would take for cases from each state in the analysis. Variable $\mathrm{D_F}$ codes cases from Florida, $\mathrm{D_M}$ codes cases from Mississippi, $\mathrm{D_L}$ codes cases from Louisiana, and $\mathrm{D_T}$ codes cases from Texas. Cases from Alabama will have the value

Figure 8.2: Dummy codes for gulf state example.

	D_F	D_M	D_L	D_T
Florida	1	0	0	0
Alabama	0	0	0	0
Mississippi	0	1	0	0
Louisiana	0	0	1	0
Texas	0	0	0	1

of 0 on all four variables. The interpretations of the coefficients obtained using this coding scheme are presented in Figure 8.3

Figure 8.3: Interpreting the coefficients on dummy codes.

Variable	Interpretation of Coefficient
D_F	Mean of Florida cases - Mean of Alabama cases
D_M	Mean of Mississippi cases - Mean of Alabama cases
D_L	Mean of Louisiana cases - Mean of Alabama cases
D_T	Mean of Texas cases - Mean of Alabama cases

- The second coding scheme is called *effect coding*. The coefficients for the effect codes obtained through least-squares regression each represent the difference between one of your groups and the unweighted mean value of all of your groups. The unweighted mean is obtained by taking the average of the scores of all five of your groups *without* weighting them by the number of observations in each group. A significant coefficient on an effect code indicates that the mean of the corresponding group is significantly different from the unweighted grand mean. Since you only have $g - 1$ codes for a variable with g levels, you will not obtain an estimate of the difference for one of your groups just by looking at the coefficients. Effect coding is useful when you do not have a specific reference group you want to use for comparison, and your groups were either manipulated or else occur in relatively equal amounts in the real world.

  Each effect code corresponds to one group in your categorical IV. However, as mentioned above, one of your groups will not be associated with any of your effect codes. Unlike dummy coding, you typically want the missing group to be one in which you do *not* have a lot of interest, since it will not be directly reflected in any of the coefficients. The value for each effect code will be equal to 1 if the case is from the corresponding group for that code variable, -1 if the case is from the missing group (the one not coded by the any of the codes), and 0 otherwise.

  Taking the example of the gulf state variable, let us assume that we want to treat Florida as the missing group. Figure 8.4 illustrates the values the code variables would take for cases from each state in the analysis. Variable E_A corresponds to Alabama, E_M corresponds to Mississippi, E_L corresponds

Figure 8.4: Effect codes for gulf state example.

	E_A	E_M	E_L	E_T
Florida	-1	-1	-1	-1
Alabama	1	0	0	0
Mississippi	0	1	0	0
Louisiana	0	0	1	0
Texas	0	0	0	1

to Louisiana, and E_T corresponds to Texas. Cases from Florida will have the value of -1 on all four variables. The interpretations of the coefficients obtained using this coding scheme are presented in Figure 8.5

Figure 8.5: Interpreting the coefficients on effect codes.

Variable	Interpretation of Coefficient
E_A	Mean of Alabama cases - Unweighted grand mean
E_M	Mean of Mississippi cases - Unweighted grand mean
E_L	Mean of Louisiana cases - Unweighted grand mean
E_T	Mean of Texas cases - Unweighted grand mean

- The third coding scheme is called *weighted effect coding*. The coefficients for the weighted effect codes obtained through least-squares regression each represent the difference between one of your groups and the weighted mean value of all of your groups. The weighted mean is equivalent to simply taking the mean value of all of your observations on the DV. In this way, groups that have more cases will have a greater influence on the estimated mean. A significant coefficient on a weighted effect code variable indicates that the corresponding group is significantly different from the weighted grand mean. Since you only have $g-1$ codes for a variable with g levels, you will not obtain an estimate of the difference for one of your groups just by looking at the coefficients. Weighted effect coding is useful when you do not have a specific reference group you want to use for comparison, and your groups reflect some real-life categorization where you expect that some groups are more common than others.

  Each weighted effect code corresponds to one group in your categorical IV. As with regular effect coding, one of your groups will not be associated with any of your effect codes. You typically want the missing group to be one in which you do *not* have a lot of interest, since it will not be directly reflected in any of the coefficients. The value for each weighted effect code will be equal to 1 if the case is from the corresponding group for that code variable and 0 if the case is from a group corresponding to a different code variable. If the case is from the missing group, the value will be equal to

$$-\frac{n_{\text{code}}}{n_{\text{missing}}}, \tag{8.1}$$

  where n_{code} is the sample size of the group corresponding to the code variable and n_{missing} is the sample size of the missing group.

  Taking the example of the gulf state variable, let us again assume that Florida is our missing group. Additionally, let us say that we had 100 cases from Florida, 60 cases from Alabama, 50 cases from Mississippi, 80 cases from Louisiana, and 120 cases from Texas. Figure 8.6 illustrates the values the code variables would take for cases from each state in the analysis. Variable W_A corresponds to Alabama,

Figure 8.6: Weighted effect codes for gulf state example.

	W_A	W_M	W_L	W_T
Florida	$-\frac{60}{100}$	$-\frac{50}{100}$	$-\frac{80}{100}$	$-\frac{120}{100}$
Alabama	1	0	0	0
Mississippi	0	1	0	0
Louisiana	0	0	1	0
Texas	0	0	0	1

  W_M corresponds to Mississippi, W_L corresponds to Louisiana, and W_T corresponds to Texas. The interpretations of the coefficients obtained using this coding scheme are presented in Figure 8.7

Figure 8.7: Interpreting the coefficients on weighted effect codes.

Variable	Interpretation of Coefficient
W_A	Mean of Alabama cases - Weighted grand mean
W_M	Mean of Mississippi cases - Weighted grand mean
W_L	Mean of Louisiana cases - Weighted grand mean
W_T	Mean of Texas cases - Weighted grand mean

- The final coding scheme that we will discuss is called *contrast coding*. The coefficients for the contrast codes obtained through least-squares regression each represent a predefined *contrast* among your group means. A contrast is basically the mean difference between two sets of your groups. Using the gulf state example, you could define a contrast to be equal to the average of the SAT scores for Alabama and Mississippi minus the average SAT score for Florida. Each of the code variables in this scheme represents a different contrast among your groups. A significant coefficient on a contrast code variable would indicate that the corresponding contrast is significantly different from zero. However, you cannot just choose any set of contrasts: The contrasts being tested by your codes must all be independent. If you choose a set of contrasts that are not independent then the coefficients on each code variable will *not* be equal to the value of the contrast. Contrast coding is useful when you are primarily interested in a specific set of comparisons among your group means.

  The first thing you need to do to set up contrast codes is to determine what contrasts will be represented on each of your code variables. You should start by deciding exactly what contrast among the group means is of greatest interest for you. Most commonly this will be equal to the mean of one set of groups minus the mean of another set of groups. It is not necessary that the two sets have the same number of groups, nor is it necessary that a given contrast involve all of the groups in your IV.

  Once you have determined the specific contrast you want, you must then determine a coding scheme that will represent that contrast. If your contrast represents the comparison of two sets of groups then you should choose one of your sets to be the "positive set" and which to be the "negative set." This is an arbitrary decision and only affects the sign of the resulting contrast. If the estimated value of the contrast is positive, then the mean of the positive set is greater than the mean of the negative set. If the estimated value of the contrast is negative, then the mean of the negative set is greater than the mean of the positive set. Once you do this you can determine the values of your contrast code.

  - For groups in the positive set, the value of the contrast code will be equal to $\frac{g_n}{g_p+g_n}$, where g_p is the number of groups in the positive set and g_n is the number of groups in the negative set.
  - For groups in the negative set, the value of the contrast code will be equal to $-\frac{g_p}{g_p+g_n}$, where g_p is the number of groups in the positive set and g_n is the number of groups in the negative set.
  - For groups that are not in either set, the value of the contrast code will be equal to 0.

  As an example, let us say that you wanted to compare the SAT scores of Alabama and Mississippi to the values of Florida, Louisiana, and Texas in the gulf states example. We will define the Alabama and Mississippi as our positive set and Florida, Louisiana, and Texas as our negative set. The values of the corresponding contrast code variable are presented in Figure 8.8

Figure 8.8: Example contrast code.

Florida	$-\frac{2}{5}$
Alabama	$\frac{3}{5}$
Mississippi	$\frac{3}{5}$
Louisiana	$-\frac{2}{5}$
Texas	$-\frac{2}{5}$

Using this coding system does two things. First, it ensures that the sum of your values is equal to zero, which is required for a comparison among your groups to be a valid contrast. Second, when you scale the contrast code this way, the value of its coefficient obtained in a regression analysis will be equal to the desired comparison.

As with any other coding scheme, a categorical variable with g groups will be represented by $g-1$ code variables. Therefore, after determining your first contrast you must then come up with $g-2$ additional independent contrasts to fully represent the variability in your categorical IV. For two contrasts to be independent, knowing the value of one contrast should give you no information about the value of the other contrast. You can mathematically determine whether two specific contrasts are independent by multiplying together the values for each group and then summing the products up over all the groups. If your two contrasts are independent then this sum will be equal to zero. Using the gulf states example, consider the contrasts presented in Figure 8.9.

Figure 8.9: Testing independence example 1.

	C1	C2	C1 × C2
Florida	$-\frac{2}{5}$	$\frac{2}{3}$	$-\frac{4}{15}$
Alabama	$\frac{3}{5}$	0	0
Mississippi	$\frac{3}{5}$	0	0
Louisiana	$-\frac{2}{5}$	$-\frac{1}{3}$	$\frac{2}{15}$
Texas	$-\frac{2}{5}$	$-\frac{1}{3}$	$\frac{2}{15}$

Adding up the products of the coefficients we can see that these two contrasts are independent ($-\frac{4}{15} + \frac{2}{15} + \frac{2}{15} = 0$). This makes sense because knowing whether the average of Alabama and Mississippi is different from the average of Florida, Louisiana, and Texas does not tell us anything about whether the scores from Florida are different from the average of the scores from Louisiana and Texas. Now consider the contrasts presented in Figure 8.10 Adding up the products of the coefficients we can see

Figure 8.10: Testing independence example 2.

	C1	C2	C1 × C2
Florida	0	0	0
Alabama	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$
Mississippi	$\frac{1}{2}$	0	0
Louisiana	$-\frac{1}{2}$	$-\frac{1}{2}$	$\frac{1}{4}$
Texas	$-\frac{1}{2}$	0	0

that C1 and C2 are not independent ($\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$). This makes sense because knowing that Alabama is greater than Louisiana would lead us to predict that the average of Alabama and Mississippi would probably be greater than the average of Louisiana and Texas.

When using contrast codes, each contrast must be independent of all of the other contrasts for the coefficients to actually estimate the contrast values. It can be difficult to come up with independent contrasts just by testing them using the above procedure. Luckily there are some general guidelines you can use to make creating sets of independent contrasts easier. If your initial contrast compares two sets of groups, then any contrasts that only compare groups within the same set will be independent of the initial contrast. Additionally, anytime that the two contrasts do not have any groups in common (i.e., groups with nonzero values for one contrast always have zero values on the other contrast) the two contrasts will be independent. When defining these additional contrast codes you should follow the same procedure we discussed above for the initial contrast to determine the coefficients for these secondary contrasts. If you do not then your unstandardized regression coefficient is not guaranteed to be equal to the value of the contrast.

Taking the example of the gulf state variable, let us assume that the main contrast of interest was a comparison of SAT scores from Alabama to the average scores of the other states. A set of independent contrast codes (the first of which is the main contrast of interest) is presented in Figure 8.11.

Figure 8.11: Contrast codes for gulf state example.

	C1	C2	C3	C4
Florida	$-\frac{1}{5}$	$\frac{1}{2}$	$\frac{1}{2}$	0
Alabama	$\frac{4}{5}$	0	0	0
Mississippi	$-\frac{1}{5}$	$-\frac{1}{2}$	0	$\frac{1}{2}$
Louisiana	$-\frac{1}{5}$	$-\frac{1}{2}$	0	$-\frac{1}{2}$
Texas	$-\frac{1}{5}$	$\frac{1}{2}$	$-\frac{1}{2}$	0

The interpretations of the coefficients obtained using this coding scheme are presented in Figure 8.12. However, you should keep in mind that this is specific to the way we decided to define our contrast codes. If we had chosen an alternate set of codes then the interpretations would be entirely different.

Figure 8.12: Interpreting the coefficients on contrast effect codes.

Variable	Interpretation of Coefficient
C1	Mean of Alabama cases - Mean of Florida, Mississippi, Louisiana, and Texas cases
C2	Mean of Florida and Texas cases - Mean of Mississippi and Louisiana cases
C3	Mean of Florida cases - Mean of Texas cases
C4	Mean of Mississippi cases - Mean of Louisiana cases

8.3 Regression analysis using code variables

- The very first thing that you must do is to actually create the code variables defined in your coding scheme. To create a code variable in SPSS
  - Choose **Transform** → **Recode** → **Into different variables**.
  - Select your categorical IV and click the arrow button.
  - Enter the name of your code variable in the **Name** box on the right-hand side.
  - Click the **Change** button.
  - Click the **Old and New Values** button.
  - Specify the value of the code variable for each group in the categorical IV by taking the following steps.
    - Enter the value used to identify the group in the categorical IV in the box to the right of **Value** in the **Old Value** section.
    - Enter the value you want the code variable to take for this group in the box to the right of **Value** in the **New Value** section.
    - Click the **Add** button.
  - Once you have defined the value of the code variable for each group in the categorical IV, click the **Continue** button.
  - Click the **OK** button.

You will need to repeat this procedure for each of your code variables.

- Once you have generated a set of code variables to represent your categorical IV, you can test for the influence of the IV on the DV by performing a standard regression analysis in which you include the entire set of your code variables.
- If your model only contains the code variables from your categorical IV then the F test of your overall model will test the ability of your IV to predict the DV. The model R^2 will be the proportion of variability in the DV that can be explained by your categorical IV. If you used one of the coding schemes that we described above, you will also be able to interpret the coefficients on the individual code variables.
- You can also include the code variables for your categorical IV in a statistical model that contains other IVs. You could include your code variables in a model that contains one or more continuous IVs, or you could include them in a model that contains code variables for other categorical IVs. You could even create a statistical model that contains multiple continuous IVs along with multiple categorical IVs. No matter how you decide to define your model, you can still estimate your coefficients using the standard least-squares regression procedures.
- If you have other IVs in the model, then you can test whether there is a unique effect of your categorical IV on the DV by testing the influence of the full set of code variables for that IV. You would do this using the procedures for testing the effect of IV sets presented in section 5.3. In this case, the resulting F-test will tell you whether your categorical IV has an independent effect on the DV above and beyond the influence of other IVs in the model.
- The inclusion of other IVs in your model will change the way that you can interpret your coefficients if you used one of the coding schemes described above. If there are any relations between the categorical IV and the other IVs in the model, then the interpretation of the regression coefficients on the code variables will change. Instead of testing differences among the actual group means on the DV, the coefficients will now test for differences among the *adjusted means*. The adjusted means for your categorical IV are corrected for any differences between the groups on other IVs in the model. For example, let us say that you performed an analysis including both the gulf state variable as well as another variable indicating the SES of the participants. After generating a set of dummy codes for our gulf state variable (using Alabama for the reference group) we decide to include both the gulf state code variables and the SES variable in a regression analysis. Originally the test of the coefficient on the first code variable told us whether there was a difference between the average SAT scores for Florida and Alabama. If we included SES in our model, however, the test of the coefficient on the first code variable would instead tell us whether there is a difference between the average SAT scores for Florida and Alabama ignoring any differences between the two that could be explained by differences in the average SES of the two states.
- If all of the continuous IVs in your model are centered, then the coefficients on your code variables will actually estimate the corresponding comparisons among the adjusted means. The t-test for each coefficient will tell you whether the comparison is significantly different from zero. If the other variables are not centered then the tests of the coefficients will still tell us whether the corresponding comparisons among the adjusted means are significantly different from zero, but the coefficients themselves will not estimate the actual values of the comparisons.

8.4 Relating ANOVA and regression

- For quite some time ANOVA and regression were viewed as opposing ways to conduct studies and analyze data.
    - ANOVA was primarily used to test the influence of one or more categorical IVs on a continuous DV. Researchers concentrated on testing the IVs in their models and typically had little interest in prediction. Since they rarely generated predicted values, researchers using ANOVA typically never examined whether the distribution of the residuals met the model assumptions. The effect of the IVs was measured by the amount of variability that each variable could uniquely explain in the DV. Researchers using ANOVA would commonly examine tests of interactions among their variables. The dominant testing strategy involved comparing full and reduced models, where the predictive ability of a model containing all of the IVs is compared to the predictive ability of a

model containing all of the IVs except the one being tested. The effect was tested using an F statistic that represented the variability in the DV that could be uniquely explained by a given IV divided by the unexplained variability in the DV. Researchers who used ANOVA tended to work in experimental settings where they manipulated their IVs. ANOVA emphasized the use of balanced (uncorrelated) designs, so multicollinearity was not usually an issue.

- Regression was primarily used to test the influence of one or more continuous IVs on a continuous DV. Researchers using regression were typically interested in both testing the effects of their IVs on the DV as well as predicting the expected values of the DV for new observations. They would commonly test whether the distribution of their residuals fit the model assumptions. The effect of the IVs was measured by the size of the corresponding coefficient in the estimated regression equation. Researchers using regression would typically only examine the main effects of the variables of interest. The dominant testing strategy involves hypothesis tests of point estimates, comparing the observed coefficients to a null value of zero. This effect was tested using a t-test equal to the estimated coefficient for the IV being tested divided by the standard error of the estimate. Researchers who used regression tended to work in field settings where they observed the values of their IVs. Observed variables are often correlated, so multicollinearity was commonly an issue.

- Mathematically, however, both ANOVA and regression are specialized versions of the GLM, a general method of examining the statistical relations among variables. Therefore, the differences in the way that people examine the relation between variables in ANOVA and regression are really just a matter of convention. Modern researchers commonly blur the lines between ANOVA and regression.

  - Researchers using ANOVA can obtain an estimated equation allowing them to predict new values of the DV from the IVs in their model. They are also more likely to be concerned about the distribution of the residuals from their model. Instead of insisting on balanced designs, modern statistical software enables researchers to easily analyze ANOVA models that contain correlated IVs. This has led to ANOVA being used by more researchers in field settings. However, it has also introduced the potential concern of multicollinearity to these analyses.
  - Researchers using regression now commonly consider the proportion of variability that their IVs can explain in the DV through the examination of partial and semipartial correlations. It is now more common to see regression models that include interaction terms. It is also more common to see regression being used in experimental settings, since the tests of IVs (even those that are manipulated) will be more powerful if you have a quantitative measurement of the level of the IV than simply a categorical assignment.

- Creating code variables can become very tedious if you want to test for the influence of a number of different categorical IVs in the same analysis. Luckily, the **General Linear Model** procedure in SPSS can directly analyze the effect of one or more categorical variables using dummy codes without requiring you to manually create any code variables. Additionally, it will not only provide you with the tests of your individual coefficients, but will also automatically report the F-test for the joint ability of your code variables to independently predict variability in your DV. However, to use this procedure your categorical variables must already be represented in your data set by variables that use different numbers to indicate which category each case is in.

  To perform an analysis in SPSS using the General Linear Model

  - Choose **Analyze → General Linear Model → Univariate**.
  - Move your DV to the box labeled **Dependent Variable**.
  - Move any categorical IVs to the box labeled **Fixed Factor(s)**.
  - Move any continuous IVs to the box labeled **Covariate(s)**.
  - Click the **Options** button.
  - Check the box next to **Parameter estimates**.
  - Click the **Continue** button.
  - Click the **OK** button.

- The SPSS output from running an analysis using the General Linear Model contains the following sections.
    - **Between-Subjects Factors**. This table just lists out the different levels of any categorical variables included in your model.
    - **Tests of Between-Subjects Effects**. This table provides an F test of each IV, indicating whether or not it can independently account for a significant amount of variability in your DV. This provides the same results as testing the change in model R^2 that you get from the test of an IV set (where each IV is treated as its own set) described in section 5.3.
    - **Parameter Estimates**. This table provides the estimates and tests of the individual parameters in your model. The dummy code variables for categorical IVs are defined so that the group corresponding to the highest value in the IV is the reference group.

Chapter 9

Interactions involving Categorical IVs

9.1 Interactions among categorical IVs

- In Chapter 7 we discussed how interaction effects can be used to determine if the influence of one continuous IV on the DV depends on the value of a second continuous IV. An interaction between two continuous IVs indicates that the slope of the relation between the first IV and the DV is different at different levels of the second IV.
- We can conceive of a parallel interaction effect when working with categorical IVs. Let us consider a study that was designed to look at the effect of a training program and gender on the extent to which heart-attack victims comply with their doctor's orders following their release from the hospital. In this example we have two categorical IVs. The training program IV indicates whether or not participants were provided with health-care training during their hospital stay. The gender IV indicates whether the participants are male or female. The DV will be a scale designed to measure the extent to which the patients have changed their lifestyle in compliance with their doctor's orders six months following their release.

  In the last chapter we learned how we can use regression analysis to determine if either (or both) of these categorical IVs have an influence on the amount of compliance. In addition, we also might want to determine whether the effectiveness of the training program depends on whether the patient is male or female. This would represent an interaction effect between training and gender on compliance.
- If we have a main effect of a categorical IV, we expect that people in different groups will have different values on the DV. An interaction between two categorical variables would mean that the differences among the groups of the first IV vary across the different levels of the second IV.

  Considering the medical compliance example presented above, a main effect of the treatment program would indicate that those who receive the treatment program comply with their doctors' orders to a different degree than those who do not receive the treatment program (averaging over any effects of gender). A main effect of gender would indicate that there are overall differences between the compliance rates of men and women (averaging over any effects of the treatment program). An interaction between the treatment program and gender would mean that the difference between men who receive the treatment program and men who do not receive the treatment program is not the same as the difference between women who receive the treatment program and women who do not receive the treatment program. This means that the effectiveness of the treatment program depends on whether the patient is a man or a woman.
- We often refer to the *study design* when we conduct a study that only has categorical IVs. The study design is a shortcut notation that tells us how many categorical IVs that we have, and how many levels there are in each IV. To report the design of a study you go simply list out the levels for each of your factors with an "x" placed in between them. For example, if you conduct a study that has one IV with 3 levels, a second IV with 2 levels, and a third IV with 4 levels, you would report that the study has a "3x2x4 design." Orally you would say the word "by" wherever there is an "x," so the example we just gave would be a "three by two by four design." The total number of possible combinations of your IVs can be determined by simply multiplying the numbers in the design. So, there would be a total of 24 different possible combinations in a 3x2x4 design.

- In a 2x2 design, you can represent the main effects and the interaction between the variables as a specific set of contrasts comparing the cell means. An examination of these contrasts can help us understand how to interpret main effects and interactions between categorical variables. Consider Figure 9.1, which illustrates all of the possible combinations in such an example.

Figure 9.1: Combining the IVs in a 2x2 design.

	Variable A level 1	Variable A level 2
Variable B level 1	Cell 1	Cell 2
Variable B level 2	Cell 3	Cell 4

In this design, your effects could be tested by using the following contrasts.

  - Main effect of A: (Cell 1 + Cell 3) - (Cell 2 + Cell 4).
  - Main effect of B: (Cell 1 + Cell 2) - (Cell 3 + Cell 4).
  - AxB interaction: (Cell 1 - Cell 2) - (Cell 3 - Cell 4) = (Cell 1 - Cell 3) - (Cell 2 - Cell 4).

The main effect for A combines the cells under Variable A level 1 to see if they are different from the combination of the cells under Variable A level 2. Similarly, the main effect for B combines the cells under Variable B level 1 to see if they are different from the combination of the cells under Variable B level 2. The interaction effect looks to see if the difference between the two levels of Variable A under level 1 of Variable B is the same as the difference between the two levels of Variable A under level 2 of Variable B. Since interactions are symmetric, this comparison will always be the same as checking to see if the difference between the two levels of Variable B under level 1 of Variable A is the same as the difference between the two levels of Variable B under level 2 of Variable A.

- Our method for testing the interaction between two categorical variables can be derived from two earlier sections of these notes. As discussed in section 8.3, the effect of a categorical IV on a DV can be examined using linear regression if the IV is represented by a set of code variables. The overall effect of the categorical IV is tested by examining the contribution of the entire set of code variables. In section 7.5 we mention that the interaction between two sets of IVs can be represented by a third set of variables, where each variable in the third set is created by multiplying one of the variables in the first set by one of the variables in the second set. If a categorical variable is represented by a set of code variables, and the interaction between two sets can be represented by a third set containing the products of the variables in the first two sets, we represent the interaction between two categorical variables in regression by a third set of variables that are equal to the code variables for the first IV multiplied by the code variables for the second IV. Figure 9.2 provides an example of the codes that would be calculated to represent the interaction between a categorical IV designed to represent a participant's native language (Spanish, English, or Chinese) and the state they are from (Washington, Oregon, California, or Arizona).

  The interactive effect can be tested by determining whether this third set of code variables can account for a significant amount of variability in the DV above and beyond that accounted for by the main effects of those IVs. This means that you must include the original code variables from both of the categorical IVs in the model if you want to test the effect of their interaction.

- An interaction between two categorical IVs will be represented by a number of code variables equal to $(j-1)(k-1)$, where the first IV has j groups and the second IV has k groups. This makes sense, since multiplying each of the $j-1$ variables in the first set by each of the $k-1$ variables in the second set will produce a total of $(j-1)(k-1)$ new variables for our interaction set.

- It is possible to interpret the tests of the individual coefficients in your interaction set, but it is a bit more difficult to do so than with main effects. As we mentioned above, each interaction term is the product of two main effect terms. The coefficient measures the extent that the difference represented by the coefficient for the first main effect term depends on the difference represented by the coefficient for the second main effect term. Exactly what these differences are will depend on the specific method that you used to code your categorical IVs. For example, the test of the coefficient on interaction code

Figure 9.2: Codes for the interaction between two categorical IVs.

Type of person	Language codes		State codes		
	L1	L2	S1	S2	S3
Spanish speaker from Washington	1	0	1	0	0
Spanish speaker from Oregon	1	0	0	1	0
Spanish speaker from California	1	0	0	0	1
Spanish speaker from Arizona	1	0	-1	-1	-1
English speaker from Washington	0	1	1	0	0
English speaker from Oregon	0	1	0	1	0
English speaker from California	0	1	0	0	1
English speaker from Arizona	0	1	-1	-1	-1
Chinese speaker from Washington	-1	-1	1	0	0
Chinese speaker from Oregon	-1	-1	0	1	0
Chinese speaker from California	-1	-1	0	0	1
Chinese speaker from Arizona	-1	-1	-1	-1	-1

Type of person	Language by state interaction codes					
	L1S1	L1S2	L1S3	L2S1	L2S2	L2S3
Spanish speaker from Washington	1	0	0	0	0	0
Spanish speaker from Oregon	0	1	0	0	0	0
Spanish speaker from California	0	0	1	0	0	0
Spanish speaker from Arizona	-1	-1	-1	0	0	0
English speaker from Washington	0	0	0	1	0	0
English speaker from Oregon	0	0	0	0	1	0
English speaker from California	0	0	0	0	0	1
English speaker from Arizona	0	0	0	-1	-1	-1
Chinese speaker from Washington	-1	0	0	-1	0	0
Chinese speaker from Oregon	0	-1	0	0	-1	0
Chinese speaker from California	0	0	-1	0	0	-1
Chinese speaker from Arizona	1	1	1	1	1	1

L1S1 in Figure 9.2 tells you whether the difference between Spanish and Chinese speakers is the same for people from Washington and Arizona.

- You should not use dummy codes when you are coding a categorical IV for use in a model that will include interactions with that IV. The interaction sets created by multiplying together variables with dummy codes will be collinear with main effects of those sets. This will affect both the tests of the individual coefficients on the main effect code variables as well as the overall tests of the main effects. However, it will not influence the tests of the interaction term, its code variables, or the overall R^2 of the regression model. Researchers typically prefer unweighted effect codes for your categorical IVs in regression because the results parallel those obtained from performing a standard ANOVA.

- It is possible to have an interaction between more than two categorical variables. In this case the interaction will be represented by a number of code variables equal to the product of the number of code variables for the effects involved in the interaction. For example, the interaction between an IV with 2 groups, an IV with 3 groups, an IV with 4 groups, and an IV with 5 groups would be represented by a total of (2-1)(3-1)(4-1)(5-1) = 24 different code variables. Each of these 24 variables would be a product of four code variables, one from each IV.

  In order to test for a higher-order interaction, your model must contain all of the main effects for

the IVs involved in the interaction and all of the possible lower-order interactions that can be created between those main effects. In this case, the test of the interaction can be determined by testing whether the set of interaction code variables can account for a significant amount of variability in the DV above and beyond that accounted for by other terms in the model.

- Once you detect a significant interaction between a set of categorical IVs, you must then determine the exact nature of the interaction. One way to do this is to just examine the tests of the individual coefficients to see where the significant differences lie. An easier method would be to examine a chart of the mean values of the DV within different combinations of the categorical IVs.

  Using SPSS, you can easily obtain these means if you have categorical variables in the data set representing each of your IVs. In this case you can obtain the means within each combination of the IVs in your interaction by taking the following steps.

  - Choose **General Linear Model → Univariate**.
  - Move your DV to the box labeled **Dependent Variable**.
  - Move the categorical IVs to the box labeled **Fixed Factor(s)**.
  - Click the **Options** button.
  - Move the corresponding term for each interaction you want to examine from the **Factor(s) and Factor Interactions** box to the **Display Means for** box.
  - Click the **Continue** button.
  - Click the **OK** button.

  This procedure will provide you with the mean value of your DV within each possible combination of the levels of your categorical IVs. Looking at the pattern of means will then allow you to see how the differences among the groups of one IV may change at different levels of the other IVs.

9.2 Interactions between categorical and continuous IVs

- An interaction between a categorical IV and a continuous IV can arise if the slope of the relation between the continuous IV and the DV varies across the different groups of the categorical IV. Equivalently, we might think about the size of the difference between the mean value of the DV in two different groups might be related to the value of the continuous IV.

- The interaction between a continuous and a categorical IV is represented in regression by a set of code variables. These code variables will be the products of the continuous IV with each of the code variables for the categorical IV. The effect of the interaction can be determined by testing whether the set of interaction code variables predicts a significant amount of variability in the DV above and beyond that predicted by the main effects of the categorical and continuous IVs. This means that you must include the original categorical and continuous IVs in the model to test the effect of their interaction.

- You need to center your continuous IV before multiplying it by the code variables for the categorical IV to obtain the code variables for the interaction. In addition, your model testing the effect of the interaction should include the centered version of the continuous IV. The reasons for this parallel those for centering in models containing the interactions between continuous IVs: Centering reduces the collinearity between the main effect and interaction terms, and also provides a meaningful interpretation for the coefficient on the continuous IV.

- The interpretation of the coefficients on the interaction code variables depends on the coding system used for the categorical IV. Specifically, each coefficient represents the difference in the slope relating the continuous IV to the DV between the two groups compared in the code variable. For example, if you use dummy codes then the coefficient for each interaction code variable represents the difference in slopes between the group corresponding to the code variable and the reference group.

- The easiest way to investigate the nature of an interaction between a continuous and a categorical IV is to examine the slope between the continuous IV and the DV within each group. You can get these slopes from SPSS if you have the categorical IV represented in your data set by a single numeric variable. In this case you should take the following steps.

- Choose **Data → Split file**.
- Click the radio button next to **Organize output by groups**.
- Move the variable corresponding to the categorical IV to the box labeled **Groups based on**.
- Click the **OK** button.
- Perform a simple linear regression analysis predicting your DV from the continuous IV. The model should not contain either the categorical IV or the interaction between the continuous and the categorical IV.

- The output from this procedure will include a set of regression analyses, each predicting the DV from the continuous IV just from the cases in a single group. You can then compare the estimated slopes from these analyses to determine how the relation between the continuous IV and the DV varies across different levels of the categorical IV.

- It is possible to have an interaction between any number of categorical IVs and any number of continuous IVs. The interaction would be represented in a regression model by a number of code variables equal to the product of the number of code variables used to represent the categorical IVs in the interaction. For example, if you wanted to represent the interaction between a categorical IV with three levels, a categorical IV with four levels, and two continuous IVs, you would need a total of (3-1)(4-1) = 6 code variables. Each of these code variables would represent a product of one of the code variables from each of the categorical IVs with all of the continuous IVs.

  The test of such interaction would have to take place in a model that contains all of the main effects and lower-order interactions between the various IVs composing the interaction. To determine whether your interaction accounts for a significant amount of variability in the DV, you would test whether the model containing the interaction and all of the lower-order terms accounts for significantly more variance than a model containing all of the above terms except those for the interaction of interest.

  You must center any continuous IVs before creating the terms for any interactions with categorical IVs to provide meaning to the main effects and reduce collinearity among your terms. In addition, you should probably avoid using dummy codes when you want to include a test of an interaction including multiple categorical IVs so that your tests of lower-order categorical terms are not affected by collinearity.

- It is also possible to have an interaction between a categorical variable and the polynomial function of a continuous IV. Theoretically this would mean that the nature of the polynomial function would depend on what particular group a case is in. The interaction between the categorical IV and the polynomial function would be represented by a set of code variables, where each code is equal to the product of one of the codes for the categorical IV and one of the terms in the polynomial function. To fully represent the interaction you would therefore need a number of code variables equal to $(j-1)(p)$, where j is the number of groups in the categorical IV and p is the power of the highest order term in the polynomial function. You can determine whether the interaction between the categorical variable and the polynomial function accounts for a significant amount of the variability in your DV by testing whether the inclusion of the set of interaction code variables accounts for a significant amount of variability above that accounted for by the categorical IV and the polynomial function on their own. Tests of the coefficients on the individual interaction terms can tell you how the elements of the polynomial function vary between groups. Of course, the exact interpretation of each parameter will depend on the way you code your categorical IV.

- It can be very tedious to test for interactions with categorical IVs using standard regression procedures. You must create code variables for each categorical main effect and for each interaction you want to consider. However, in section 8.4 we discussed how you can get SPSS to automatically take care of generating the code variables for an analysis using a categorical IV by using the **General Linear Model** procedure. You can also use this same procedure to get SPSS to analyze the effect of interactions between categorical variables as well as the interactions between categorical and continuous variables.

- To analyze data using the general linear model, you must first have each of your categorical IVs represented by a single variable in your data set where different values of the variable correspond to different groups. You must also have a variable representing a centered version of each of your continuous IVs.

- To perform an analysis in SPSS using the general linear model
  - Choose **Analyze → General Linear Model → Univariate**.
  - Move your DV to the box labeled **Dependent Variable**.
  - Move any categorical IVs to the box labeled **Fixed Factor(s)**.
  - Move the centered versions of your continuous IVs to the box labeled **Covariate(s)**.
  - By default SPSS will include all possible interactions between your categorical IVs, but will only include the main effects of your continuous IVs. If this is not the model you want then you will need to define it by hand by taking the following steps.
    - Click the **Model** button.
    - Click the radio button next to **Custom**.
    - Add all of your main effects to the model by clicking all of the IVs in the box labeled **Factors and covariates**, setting the pull-down menu to **Main effects**, and clicking the arrow button.
    - Add each of the interaction terms to your model. You can do this one at a time by selecting the variables included in the interaction in the box labeled **Factors and covariates**, setting the pull-down menu to **Interaction**, and clicking the arrow button for each of your interactions. You can also use the setting on the pull-down menu to tell SPSS to add all possible 2-way, 3-way, 4-way, or 5-way interactions that can be made between the selected variables to your model.
    - Click the **Continue** button.
  - Click the **OK** button.
- The SPSS output from running an analysis using the General Linear Model contains the following sections.
  - **Between-Subjects Factors**. This table just lists out the different levels of any categorical variables included in your model.
  - **Tests of Between-Subjects Effects**. This table provides an F test of each main effect or interaction that you included in your model. It indicates whether or not the effect can independently account for a significant amount of variability in your DV. This provides the same results as testing the change in model R^2 that you get from the test of the set of terms representing the effect, as described in section 5.3.

  You can ask SPSS to produce the parameter estimates from your model by clicking the **Options** button in the variable selection menu. However, SPSS will always report the parameter estimates using dummy codes for your variables. Earlier we discussed how dummy coded IVs produce poor results when testing interactions. The tests that SPSS produces using the general linear model procedure actually correspond to the results you'd get if you used unweighted effect codes, even though it will report the parameter estimates for dummy codes. This makes the parameter estimates you get using this procedure of limited use when you are considering a model with interactions between categorical IVs.
- Using the general linear model, you can test for the effect of any possible combination of continuous and categorical IVs on a DV, as well as test for the effects of any possible interactions among those variables. The procedure itself is very easy to use, and it produces output that can be easily interpreted. You will likely find that this procedure will become the backbone for all of your univariate analyses except those where you want to obtain parameter estimates from effect-coded or contrast-coded categorical IVs.

Chapter 10

Outlier and Multicollinearity Diagnostics

10.1 Detecting outliers

- An *outlier* is an observation that is in some way substantially different from the other cases you have in a data set. It is important to always check your data set for outliers because they can strongly influence the regression equation computed using least squares estimation.
- The extent to which a case is an outlier can be assessed in two different ways. The *leverage* of a case represents the extent to which it has unusual values on the IVs, while the *discrepancy* of a case represents the extent to which it has an unusual value on the DV, given its values on the IVs.
- If you have a single IV (X), the leverage for case i can be determined using the formula

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2}, \tag{10.1}$$

  where n is the number of cases. From this formula we can see that leverage is directly related to the difference between its values on the IVs and the average values on the IVs. Cases that have extremely large or small values on the IVs will have greater leverage than those that have values closer to the average. We can also see that the minimum leverage value of $\frac{1}{n}$ is obtained where X_i is equal to $\bar{X}$.
- The average leverage value will always be equal to $\frac{k+1}{n}$, where k is the number of IVs in the model. The maximum possible leverage value is 1. An observation may be considered an outlier if its leverage is greater than $\frac{2(k+1)}{n}$ for large samples, or $\frac{3(k+1)}{n}$ for small samples. You might also graphically look at the distribution of leverages to see if there is a small number of observations that have considerably higher leverage than the other cases in the data set.
- Notice that the formula for the leverage is not a function of the DV. Your leverage values specifically measure the extremity of each observation on your IV, and so will stay the same if you use the same set of IVs to predict a different DV.
- The way that least squares estimation computes its equations makes it so that observations with large leverages have a greater influence on the estimation of the slope of the regression line. A single observation with an unusually large leverage can cause an otherwise significant relation to become nonsignificant, or an otherwise nonsignificant relation to become significant.
- This effect of extreme observations is called "leverage" because it is similar to how it becomes easier to move an object using a lever when you push down farther away from the fulcrum than if you push down close to the fulcrum.

- In addition to affecting the coefficients of the estimated regression line, the leverage of an observation also affects the contribution of the observation to the overall model R^2. Even if it does not affect the actual coefficients, a case with a high leverage will still have a strong influence on the *tests* of the regression coefficients. A case that has a high leverage but whose value of Y is consistent with the regression line estimated from the other data points will substantially increase the test statistic for the slope and the overall model R^2, even though it won't actually change the slope coefficient.

- You can obtain the leverage values from a regression analysis performed in SPSS by clicking the **Save** button in the variable selection window and then checking the box next to **Leverage.** This will add a new variable to your data set starting with the letters **lev** containing the leverage values for that analysis.

  Important note: SPSS does not actually provide you with the leverage values for your cases. Instead it reports the *centered leverages*, which are computed as

$$h_{ii}^* = h_{ii} - \frac{1}{n}. \tag{10.2}$$

  The centered leverages have a minimum value of 0, a maximum value of $(1 - \frac{1}{n})$, and an average of $\frac{k}{n}$. When using the centered leverages to detect outliers you should use breakpoints of $\frac{2k}{n}$ for large samples and $\frac{3k}{n}$ for small samples.

- The simplest measure of the discrepancy for a given case is its residual. As discussed in Chapter 2, the residual for a case is equal to its observed value on the DV minus the predicted value from the regression equation. Cases that have high residuals are those that have values of the DV that are inconsistent with the values found in other cases from the data set.

- Residuals from different DVs will necessarily have different ranges, so it is not possible to establish a general cutoff regarding what values of the residuals are expected and what values indicate outlying cases. Researchers will therefore often transform the residuals to put them on a standard scale to make it easier to locate outliers.

  - *Standardized residuals* are equal to

$$\frac{e_i}{\sqrt{\text{MSE}}}. \tag{10.3}$$

    Dividing by $\sqrt{\text{MSE}}$ scales the residuals so that they indicate the number of standard deviations that the observed values of the DV are from the regression line.

  - *Studentized residuals* (sometimes called *internally studentized residuals*) are equal to

$$\frac{e_i}{s\{e_i\}}. \tag{10.4}$$

    They are similar to standardized residuals except that they are equal to the raw residual divided by the standard error of a predicted value at that point on the regression line. Compared to standardized residuals, studentized residuals will have smaller values when you are less confident about your predicted values at a particular point on your regression line.

  Researchers typically consider cases with standardized or studentized residuals less than -3 or greater than 3 to be outliers.

- There is a problem, however, with using the residuals mentioned above. When an observation has very extreme values on the DV it can have an unusually strong influence on the estimated regression line. In this case, an observation might not have a large residual simply because it bent the estimated regression line toward itself. To solve this problem, people will often use one of the following two types of residuals.

  - *Deleted residuals* are equal to

$$\frac{Y_i - \hat{Y}_i^*}{\sqrt{\text{MSE}^*}}, \tag{10.5}$$

where $\hat{Y}_i^*$ is the predicted value from a regression line estimated from all of the values in the data set *except* the observation whose residual is being calculated, and MSE^* is the MSE from that same model.

- *Studentized deleted residuals* (sometimes called *externally studentized residuals*) are equal to

$$\frac{Y_i - \hat{Y}_i^*}{s\{Y_i - \hat{Y}_i^*\}}, \tag{10.6}$$

where $\hat{Y}_i^*$ is the predicted value from a regression line estimated from all of the values in the data set *except* the observation whose residual is being calculated, and $s\{Y_i - \hat{Y}_i^*\}$ is the standard error of a predicted value on the new regression line corresponding to the residual being calculated. These are the same as studentized residuals, except that you use the standard error calculated from a regression line that includes every case in the data set *except* the one for which you are calculating the studentized deleted residual.

Researchers typically consider cases with deleted or studentized deleted residuals less than -3 or greater than 3 to be outliers.

- You can obtain all of the residuals discussed above from SPSS by clicking the **Save** button in the variable selection window and then checking the box next to whichever type of residual you want. Checking one of these boxes will add a new variable to your data set containing the corresponding set of resisuals. The unstandardized, raw residuals are saved to a variable with a name beginning with **res**, standardized residuals are saved to a variable with a name beginning with **zre**, studentized residuals are saved to a variable with a name beginning with **sre**, deleted residuals are saved to a variable with a name beginning with **dre**, and studentized deleted residuals are saved to a variable with a name beginning with **sdr**.

- A final factor that you should consider when looking for outliers is the overall *influence* of each case on the estimated regression line. The influence for a given case is a direct function of that case's leverage and discrepancy, and measures the extent to which the coefficients of the estimated regression line would change if the given case was dropped from the analysis.

- The influence of each case on a single coefficient is measured by the *DFBETAS* statistic. For each case in your data set, you will have DFBETAS for each term in your regression model. The DFBETAS for single case i representing its influence on a single coefficient j can be computed using the formula

$$\text{DFBETAS}_{ij} = \frac{b_j - b_{j(i)}}{s\{b_{j(i)}\}}, \tag{10.7}$$

where b_j is the value of the coefficient obtained when case i is included in the analysis, $b_{j(i)}$ is the value of the coefficient when case i is excluded from the analysis, and $s\{b_{j(i)}\}$ is the standard error of the coefficient obtained when case i is excluded from the analysis.

DFBETAS can have either positive or negative values. The influence of an observation is determined by how far the DFBETAS is from zero. The simplest way to determine if you have any outliers is to examine the distribution of DFBETAS for each coefficient to see if there are any cases that seem to be much more influential than the others in the data set. As a general rule, cases from small or medium samples with DFBETAS values greater than 1 or less than -1 have a strong influence on the value of the corresponding coefficient and may be considered outliers. For large samples, cases with DFBETAS greater than $\frac{2}{\sqrt{n}}$ or less than $-\frac{2}{\sqrt{n}}$ may be considered outliers.

- The influence of each case on the entire set of coefficients from your regression model is measured by the *DFFITS* statistic. The DFFITS for case i can be calculated using the formula

$$\text{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}}, \tag{10.8}$$

where $\hat{Y}_i$ is the predicted value for observation i based on the regression equation derived from the full data set, $\hat{Y}_{i(i)}$ is the predicted value for observation i based on the regression equation derived from

the data set excluding case i, $\text{MSE}_{(i)}$ is the mean squared error from the regression model derived from the data set excluding case i, and h_{ii} is the leverage for case i.

As with DFBETAS, DFFITS can have positive or negative values, and the influence of a case is determined by how far the DFFITS value is from zero. The easiest way to detect influential observations is to examine the distribution of DFFITS to see if there are any cases that are much more influential than the others in the data set. As a general rule, cases from small or medium samples with DFFITS values greater than 1 or less than -1 have a strong influence on the estimated regression equation and may be considered outliers. For large samples, cases with DFFITS greater than $2\sqrt{\frac{k+1}{n}}$ or less than $-2\sqrt{\frac{k+1}{n}}$ may be considered outliers, where k is the number of predictor variables in your regression model and n is the number of cases.

- There is a second statistic that measures the overall influence of each case on your estimated regression equation called *Cook's D*. It serves exactly the same function as DFFITS but is scaled differently. The value of Cook's D for case i can be calculated using the formula

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{\text{MSE}(k+1)}, \tag{10.9}$$

  where the summation in the numerator is over all of the cases in the data set (including case i), $\hat{Y}_j$ is the predicted value for case j based on the regression equation estimated from the full data set, $\hat{Y}_{j(i)}$ is the predicted value for case j based on the regression equation estimated from data set excluding case i, MSE is the mean squared error from the regression model estimated from the full data set, and k is the number of IVs in the regression model.

  Cook's D has a minimum value of zero, and larger values of D indicate that a case has a greater influence on the estimated regression equation. The easiest way to detect influential observations is to examine the distribution of D values to see if there are any cases that are much more influential than the others in the data set. Alternatively, researchers often use a guideline that cases with values of D greater than the critical value from the F distribution at $\alpha = .5$ with $k+1$ numerator and $n-k-1$ denominator degrees of freedom may be considered outliers.

- You should use the DFBETAS to detect outliers whenever you have a specific IV that you are most interested in, and you want to detect any cases that are having a particularly strong influence on the test of its coefficient. If you value your IVs about equally, you are better off using DFFITS or Cook's D to detect outliers since each case in your data set will only have a single value for these statistics.

- To obtain the DFBETAS from a regression model in SPSS you would click the **Save** button in the variable selection window and check the box next to **Standardized Dfbeta(s)**. This will add a set of variables to your data set beginning with the letters **sdb** containing the DFBETAS for each of the terms in your model. Note that what SPSS labels as DFBETAS is *not* the same thing as the DFBETAS we have been talking about in these notes. SPSS's DFBETAS will provide you with the raw difference between the coefficients with and without including the given case in the data set. We prefer using the standardized values because it will allow you to use a general cutoff point when determining whether a given observation is influential or not.

  To obtain the DFFITS from a regression model in SPSS you would click the **Save** button in the variable selection window and check the box next to **Standardized Dffit(s)**. This will add a variable to your data set beginning with the letters **sdf** containing the DFFITS for your cases. Again, note that what SPSS refers to as DFFITS is not the same as we have been discussing in these notes.

  To obtain the values of Cook's D from a regression model in SPSS you would click the **Save** button in the variable selection window and click the button next to **Cook's** in the **Distances** section on the left-hand side. This will add a variable to your data set beginning with the letters **coo** containing the value of Cook's D for each of your cases.

10.2 Remedial measures for outliers

- The choice of how to deal with an outlier depends heavily on the reason why the case in question has unusual values. There are two major sources of outliers.
    - **Contaminated cases** are those where the the outlying value does not truly represent the construct that is to be measured on the variable. There are many potential sources of contamination in a study. For example, a manipulation may fail to work properly, a participant may be unduly influenced by factors outside of the study, coders may misinterpret the event they are recording, or a typist may simply press an incorrect key during data entry.
    - **Rare cases** are those where the outlying value does truly represent the construct that is to be measured. In this case, the variable has an unusual value simply because it is measuring an unusual person or object.

  To determine whether an outlier is a contaminated or a rare case you must go back and examine the characteristics of the outlying observation. You should compare the outlying value in your analysis with any primary materials, such as a recorded transcript or a written survey, to eliminate the possibility that an outlier is due to a data-entry error. Other sources of contamination can be more difficult to detect. You will typically have to rely on an examination of other measures you may have collected on the outlying case or on the personal observations of the experimenter to determine if there were any other factors that may have caused the outlying value. In the absence of any evidence that the outlier represents a contaminated case, you assume that it represents a rare case.

- If your outlier represents a contaminated case, you should first see if you can correct it. If the contamination was simply from a data-entry or a coding error, a simple re-examination of the original source will usually enable you to determine the correct value. If the contamination was due to some other source, you might try collecting data from the case at a later point in time when the contaminating influence is no longer present. However, when doing this you must be sure that this procedure itself is not going to act as a contaminant. If you can find no way to correct the data in a contaminated case, you should delete it from your data set.

- If your outlier represents a rare case, the issue becomes more complicated. On the one hand, the rare case represents valid data, suggesting that the outlier should be included in the analysis. On the other hand, the outlier can have a very strong effect on the results of your analysis so that your findings may not well represent the relations present in the rest of the data. The decision to keep or discard rare cases from your data set has to balance these two concerns.

  The first thing you might do when you have rare cases is to consider whether there are any transformations you could perform on your data so that the outlying values are no longer so extreme. In Chapter 6 we discussed how power transformations can remove skewness from your residuals. It is possible that the raw scores from an analysis that needs a power transformation would show outliers that would disappear after the appropriate transformation.

  If transformations will not help your outliers, one way to help resolve the issues surrounding rare cases is to determine whether the rare cases might be thought of as coming from a different theoretical population than the rest of your data. If you define your research question so that you are investigating the relations between your variables specifically within the population defined by the majority of your data points, then you should clearly discard the rare cases if they do not seem to fit the standard population.

- If you decide that you want to keep outlying observations in your data set, you might try analyzing your data using a statistical procedure that is less influenced by outliers than least squares regression. A number of different "robust regression" approaches that are less influenced by outliers are presented in pages 417-419 of the textbook. Among these are least absolute deviation, least trimmed squares, and M-estimation.

10.3 Detecting multicollinearity

- We have already discussed multicollinearity at length back in Chapter 3, when we were first considered how to interpret the coefficients obtained from least squares regression. You will recall that the test of

each coefficient is based on the relation between the part of the corresponding IV that is independent of the other IVs in the model with the part of the DV that is independent of the other IVs in the model. Multicollinearity refers to the fact that the coefficients we obtain from least squares regression are influenced by the relations among the IV. Specifically, any variability that is jointly predicted by two or more IVs does not contribute to the significance of any of the regression coefficients. However, the jointly predicted variability *is* considered in the test of the overall model.

- Multicollinearity is a problem because it can cause the tests of your coefficients to be nonsignificant even when they may have a strong relation with the DV. If you test a given coefficient and find that it is not significant, this could either be because there is no relation between the corresponding IV and the DV or it could be because the IV is collinear with other variables in the model. In this section we will discuss statistics that can tell you the extent to which the test of a given coefficient is being influenced by multicollinearity.

- The simplest (but least accurate) way to check for multicollinearity is to simply look at a correlation matrix of your IVs. The maximum amount of variance that two IVs could be jointly predicting in the DV will be equal to the square of the correlation between them. However, knowing this correlation will not tell you how much of their overlap contributes to the prediction of the DV.

- The most commonly used measure of multcollinearity is the *variance inflation factor* (VIF). The VIF for coefficient i can be calculated using the equation

$$\text{VIF}_i = \frac{1}{1 - R^2_{i.12\ldots(i)\ldots k}}, \tag{10.10}$$

  where $R^2_{i.12\ldots(i)\ldots k}$ is the multiple correlation obtained from predicting the IV corresponding to b_i from all of the other IVs in the regression model. VIFs have a minimum value of 1, and higher VIFs are associated with more multicollinearity. Conceptually, the VIF for a coefficient tells you how much the variance of your coefficient is increased due to multicollinearity. So, if your coefficient had a VIF of 2.5, then you would know that the variance of the coefficient in your model is 2.5 times larger than it would be if it wasn't correlated with any of the other IVs.

  Statisticians have established an arbitrary criterion stating that coefficients with VIFs of 10 or more have a serious problem with multicollinearity. However, multicollinearity can have substantial effects on tests of your IVs long before this point. Since the tests of your coefficients are directly related to their standard errors, the statistic testing an IV with a VIF of 4 will be half as large as it would have been in the absence of collinearity.

- Instead of reporting VIFs, some statistical software packages (such as SPSS) provide you with the *tolerances* of each of your coefficients. The tolerance for coefficient i can actually be directly calculated from its VIF using the equation

$$\text{Tolerance}_i = \frac{1}{\text{VIF}_i}. \tag{10.11}$$

  Tolerances range from 0 to 1, where lower values are associated with more multicollinearity. Conceptually, the tolerance for a coefficient measures the proportion of the variability in the corresponding IV that is independent of the other IVs in the regression model. If we were to follow the guideline presented above that VIFs over 10 indicate severe multicollinearity, we would correspondingly claim that coefficients with tolerances less than $\frac{1}{10} = .10$ have severe multicollinearity.

- To obtain the VIFs and tolerances of your coefficients in SPSS you can simply click the **Statistics** button in the variable selection window and check the button next to **Collinearity diagnostics**. SPSS will then add two columns to the **Coefficients** table in your results, one containing the VIFs and the other containing the tolerances of your coefficients.

10.4 Remedial measures for multicollinearity

- By definition, we use regression to find the independent influence of each IV on the DV. Therefore any method that you might use to analyze your data will have the same problems with multicollinearity as you find with least squares regression. The remedial measures for multicollinearity therefore all involve changing your IVs so that they are no longer correlated.

- The least invasive procedure you can try is to simply collect more data. Although this will not typically reduce the amount of collinearity among your IVs, the increase in power that you get from having a larger sample can at least somewhat counter the reduction in power from having collinear IVs. However, the benefit of increasing your sample size decreases exponentially as your sample size gets larger, so the collection of more data is only likely to have a substantial benefit if your original sample size is reasonably small.

- A second option is to change your regression model so that your IVs do not have as much overlap. You might do this by dropping those IVs that have high correlations with other variables in the model but are not of great theoretical interest. If you have specific sets of IVs that seem to be strongly related to each other, you might consider creating a composite measure (like the sum or the average of those variables) and replace the original variables in your model with the composite.

- A third option would be to perform a factor analysis or a principle components analysis on your IVs using an orthogonal rotation, and use the resulting factor or component scores as your predictors. While a thorough discussion of these techniques is beyond the scope of these notes, in general both factor analysis and principle components analysis are used to find the theoretical structure underlying a set of measures. By examining the correlations among the IVs, these analyses can determine a set of uncorrelated "factors" or "components" that theoretically are causing people to respond the way that they do. Instead of using the original variables, you can use the factor/component scores to predict your DV. The set of components will be able to predict just as much of the variability in the DV as the original variables, but it will do so using a set of uncorrelated variables.

  The main problem with using factor/component scores is that you are no longer predicting your DV from the same variables that you started off with. If the factor or principle component analysis provides you with a set of meaningful theoretical variables, then often examining those relations can be even more informative than the analysis on your original variables. However, little is gained if you analyze the relation between your DV and a set of unintelligible component scores.

- The final option that you have is to perform a *ridge regression*. Details about this procedure are provided in your text on pages 427-428. In general, this procedure involves adding a small amount of random error to each of your scores. While this random error itself will reduce the predictive ability of your IVs, it will also reduce their correlations with each other. If you have severe multicollinearity, the cost of adding in the error can actually be less than the benefit gained from the reduction in multicollinearity. Ridge regression is a somewhat complicated procedure, and is actually rarely done in practice. Very few statistical packages have built-in commands that will allow you to perform a ridge regression.

References

Aiken, L. S., & West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions.* Thousand Oaks, CA: Sage.

Aldrich, J. (2005). *Fisher and Regression.* Retrieved August 24, 2005 from http://www.economics.soton.ac.uk/staff/aldrich/aldrich.htm

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Mathods & Research, 16*, 78-117.

Carmer, S. G., & Swanson, M. R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal of the American Statistical Association, 68*, 66-74.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1336.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edition). Mahwah, NJ: Erlbaum.

Enders, W. (2004). *Applied Econometric Time Series.* Hoboken, NJ: Wiley.

Fisher, R. A. (1928). *Statistical Methods for Research Workers* (2nd ed.). London: Oliver & Boyd.

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics, 21*, 607-611.

Granger, C., & Newbold, P. (1974). Spurious regressions in econometrics. *Journal of Econometrics, 2*, 111-120.

Kenny, D. A., & Judd, C. M. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99*, 422-431.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19-40.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear Statistical Models* (4th edition). Chicago, IL: Irwin.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin, 118*, 155-164.

Rencher, A. C., & Scott, D. T. (1990). Assessing the contribution of individual variables following rejection of a multivariate hypothesis. *Communications in Statistics: Simulation and Computation, 19*, 535-553.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245-251.