Chapter Eight: Multivariate Analysis

Up until now, we have covered univariate ('one variable') analysis and bivariate ('two variables') analysis. We can also measure the simultaneous effects of two or more independent variables on a dependent variable. This allows us to estimate the effects of each independent variable on the dependent variable, while controlling for the effects of one or more other independent variables. This is called multivariate ('multiple variables') analysis. In this chapter we review two ways to do that by using techniques that you have already used: crosstabs and regression analysis.

Crosstabs Revisited

Recall that the crosstabs procedure is used when variables are nominal (or ordinal). Simple crosstabs, which examine the influence of one variable on another, should be only the first step in the analysis of social science data (refer to Chapter Five). One might hypothesize that people who identify as Republicans are more likely to oppose abortion than those who identify as Democrats, then run the crosstabs, and then conclude you were right. However, this one•step method of hypothesis testing is very limited. What if most of the Republicans in your sample are strongly religious and most of the Democrats are not strongly religious? Is it the political party that best explains your findings, or is it religious orientation?

Let's explore this issue. First, we will test whether political conservatism is related to attitudes toward abortion. Before we run the Crosstabulation, we will recode the independent variable (PARTYID) to make our interpretation easier. Refer back to Chapter 3 to review how to recode into a new variable. Naming your new variable PARTYIDR (for partyid recode), recode PARTYID into three categories as follows:

Values for PARTYID	Values for PARTYIDR	Value Label for PARTYIDR
0-2	1	Democrat
3	2	Independent
4-6	3	Republican
All other values	System-missing	

Then, run Crosstabs using PARTYIDR as the independent variable (in the column position), and ABANY as the dependent variable (in the row position). You may want to review Chapter 5 for direction on how to do this. You should get results that look like the table shown in Figure 8.1:

As the results show, 43.4% of all respondents think that abortion should be allowed for any reason. However, political party identification does appear to be related to attitudes toward abortion. Half (50.1%) of respondents who identify as Democrats are in favor of allowing abortion for any reason, compared to only about a third of those identifying as Republicans. So, our hypothesis appears to

be supported. But we are still left with the question of whether the strength of religious beliefs actually accounts for this relationship.

We will do a multivariate crosstabulation to test this idea, also called an elaboration analysis. If the relationship between party identification and attitutes toward abortion persists no matter what the strenght of religious beliefs, then we would conclude that party identification is the better predictor of these attitudes. If the relationship between party identification and attitudes disappears when the strength of religious beliefs is controlled (that is, nearly the same percentage of Democrats and Republicans approve of abortion within each level of religiosity), then we would conclude that the strength of religious beliefs is the better predictor. But what if the relationship between political party identification is different, depending on the strength of religious beliefs? If that is the case, then we would say that there is an interaction between the two independent variables; they work together in shaping attitudes toward abortion.

Recall that your original crosstabs procedure produces one contingency table, with as many rows as there are categories (or values) of the dependent variable, and as many columns as there are categories of the independent variable. When you start using control (sometimes called test) variables, you will get as many separate tables as there are categories of the control variable. There are four categories in the RELITEN variable, the measure of strength of religious beliefs. Thus, we should expect to get four contingency tables, each one showing the relationship between PARTYIDR and ABANY for a different category of RELITEN.

To do the crosstabs, click **Analyze**, **Descriptive Statistics**, **Crosstabs**, just as you did for testing the bivariate relationship between PARTYIDR and ABANY. This time, click on RELITEN and move it into the box called **Layer 1 of 1**, The dialog box should look like the one shown in Figure 8-2. Click **OK**. Your results should look like the table shown in Figure 8-3.

Notice that the relationship between PARTYIDR and ABANY is different depending on the different categories of RELITEN. What does that tell you?
Try other variables as a control (i.e. in place of RELITEN) to see what happens. As a general rule, here is how to interpret what you find from this elaboration analysis:

- If the partial tables are similar to the zero•order table (this is the original bivariate table shown in Figure 8-1), you have *replicated* your original findings, which means that in spite of the introduction of a particular control variable, the original relationship persists. The only way to convince us that this is indeed a strong, or even causal, relationship is if you control for all the other logical independent variables you can think of, and still find essentially no differences between the zero•order tables and their partials.

- If the difference shown in all the partial tables (the separate tables for each category of the control variable) are significantly smaller than those found in the original AND IF your

control variable is antecedent (occurs prior in time) to both the other variables, you have found a *spurious* relationship and explained away the original. In other words, the original relationship was due to the influence of that control variable, not the one you first hypothesized.

- If the partials are less AND IF your control variable is intervening (that is, the control variable occurs in time after the original independent variable), you have *interpreted* the relationship. If the time sequence between the independent and control variable is not determinable (or otherwise unclear), then you don't know whether you have explanation or interpretation, but you do know that the control variable is important.

- If one or more of the differences shown in the partial tables is stronger than in the original and one or more is weaker, you have discovered the conditions under which the original relationship is strongest. This is referred to as *specification*, or the interaction effect.

- If the zero order table showed weak association between the variables, you might still find strong associations in the partials (which is a good argument for keeping on with your initial analysis of the data even if you didn't "find" anything with bivariate analysis). The addition of your control variable showed it to have been acting as a *suppressor* in the original table.

- Last, if a zero order table shows only a weak or moderate association, the partials might show the opposite relationship, due to the presence of a *distorter* variable.

Try some of your own three•way (or higher) tables using some of the variables in the GSS02A data set. Recall that for this procedure, there should be few categories for each variable, particularly your control variables (so you might need to recode), and you are limited to variables measured at, or recoded to, nominal or ordinal levels.

Multiple Regression

Another statistical technique estimating the effects of two or more independent variables on a dependent variable is multiple regression analysis. This technique is appropriate when your variables are measured at the interval or ratio level, although independent variables with only two categories can be treated as interval level (these are generally called “dummy variables” in regression analysis). Although technically designed for interval/ratio variables, it is common for researchers to use multiple regression with ordinal variables as well.

You may remember from a statistics course that when we use regression analysis, we make several assumptions about our data: that there is a linear relationship between each independent variable and the dependent variable, that the distribution of values in your variables follows a normal distribution, and the, and variables are measured at interval or ratio levels.

The variable HRSRELAX indicates the amount of leisure time respondents have. Since it is measured in the number of hours per day, it is an interval level variable. There is much existing research that finds that women have less leisure time than men do. Let's see if this pattern holds for our data as well. Since we are only working with two variables at this stage, we will first do a bivariate regression analysis, just as we did in Chapter 7. But first, we must recode the sex variable. The mathematics on which regression analysis is based requires that variables having two categories be coded as 0 and 1. In the GSS02A data, sex is coded as 1 (male) and 2 (female). So, you must first recode sex. If necessary, review the directions for recoding into a new variable in Chapter 3. Name the new variable SEXR (sex recoded), and code male as 0 and female as 1. Do a frequency distribution of the new variable to make sure you did it correctly. It should look like the one shown in Figure 8-4. Don't forget to add value labels to the new variable.

Now we are ready to analyze the data. We will start first by examining the relationship between sex and leisure time. On the menu, click **Analyze**, **Regression**, **Linear**. Click on HRSRELXS and move it into the **Dependent** box at the top of the dialog box. Click on SEXR and move it into the **Independent(s)** box. It should look like the one shown in Figure 8-5. Click **OK**.

Your results should look like those shown in Figure 8-6. Looking first at the Model Summary table, you will see that the R-squared value is .016. As you recall from Chapter 7, this means that 1.6% of the variation in the dependent variable (hours of relaxation time) is explained by knowing the sex of the respondent. This doesn't seem like very much. However, our goal was not to explain all of the factors that influence how much time people have to relax. Our goal was to find out whether women report having less relaxation time than men do. For this, we need to look at the Coefficients table. If you look at the B coefficient for SEXR, you will see that it is -.722. If there was absolutely no relationship between SEXR and HRSRELAX, the coefficient would be zero. How do we interpret this coefficient? Recall the discussion in Chapter 7: a one unit change in the independent variable (SEXR) is associated with a change in the dependent variable (HRSRELAX) equal to the value of B. So, if we increase the value of SEXR by 1, on average, we get a change of -.722 in HRSRELAX. Since SEXR is a dichotomous variable, increasing the value by 1 simply means going from "male" to "female." Here is what one might say in a report: Women report .722 fewer hours than men in their average amount of relaxation time per day. (It is fewer, because the coefficient is negative; so, when X, the independent variable, goes up, from 0 to 1, Y, the dependent variable, goes down. Notice that the t is statistically significant, with a value of .000. This means that since the sample was randomly selected, the pattern you see in the data is very likely an accurate reflection of the average experiences in the population from which the sample was drawn.

Now, that analysis tells us that women report fewer hours of relaxation per week, but it doesn't tell us why. Two variables that might be relevant are the number of children that people have, and their income level. Because most women spend more time doing child care and housework than most

men do, that might be one reason that they have fewer hours of relaxation time. So, we would expect to see a negative relationship between the number of children and the amount of leisure time. Another might be income; the higher the family income, the more money there is available to hire others to do some of the work, like house cleaning, babysitting, gardening, etc, making more time for leisure. If that is the case, then we would expect to find a positive relationship between income and leisure time. Finally, if these two variables explain why women have less leisure time than men do, then controlling for them should result in a smaller B coefficient for SEXR. Lets test this out using multiple regression analysis.

Once again, click **Analyze**, **Regression**, **Linear**. Click on HRSRELAX and move it into the **Dependent** box at the top of the dialog box. Click on SEXR and move it into the **Independent (s)** box. Then, move CHILDS and INCOME98 into **Independent(s)** box as well. It should look like the one shown in Figure 8a-7. Click **OK**.

Look at the B values in the Coefficients in the table Figure 8b-7. None of our additional hypotheses are supported by the data. Looking at the coefficient for CHILDS, we see that the value is -.034. It is not 0, so there appears to be some relationship between number of children and hours of relaxation time in the data, but you see that the t statistic is not statistically significant, which means that we cannot generalize from the sample to the population from which it is drawn. Thus, although there is a non-zero coefficient, we cannot draw any inferences based on this result. In contrast, the coefficient for family income is statistically significant, however we had hypothesized that as family income increases, so does leisure time. But the B coefficient is negative. The coefficient tells us that with each additional movement of one unit upward in income (review a frequency distribution of INCOME98 to see what a unit increase means), leisure time goes down by .052 hours. Finally, we hypothesized that controlling for the number of children and family income would "explain" some of the gender difference in the number of hours of relaxation time, which would result in a smaller B coefficient. Instead, notice that the coefficient is actually larger! How might we interpret this finding?

Chapter Eight Exercises

How would you hypothesize the relationship between FEAR (Afraid to walk at night in neighborhood) and SEX?

a. Write out your hypothesis.
b. Run a crosstabs to test your hypothesis and report your results.
c. Now, do a second crosstabs, this time controlling for CLASS. Report your results.
d. Now run FEAR and SEX but control for TRUST. Report your results.

Choose three independent variables for the GSS02A data that you think influence the number of hours people watch television (TVHOURS, the dependent variable).

a. Write up your hypotheses (how and why each independent variable is associated with the dependent variable.
b. Run a multivariate regression to test your hypotheses, and report your results.

New Exercise - Not in the text

Does education (educ), religiosity (measured by church attendance), and sex influence the number of children that someone has? Write a hypothesis for the relationship between each IV and the DV (you should have three hypotheses). Is each hypothesis supported? How does education influence the age of childbearing? How much change in age of childbearing occurs for a one year increase in education? Do people with higher degrees of religiosity have kids at younger or older ages? How do you know? Do women have children at a younger age than men? (Refer to a statistics text book for help with these questions) [Question contributed by Pat Jennings]