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 from Chapter 5 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. We might begin this first step by hypothesizing that women are more strongly religious than men, and that African Americans and Hispanics are more strongly religious than Anglos.

The 2016 General Social Survey provides data that we can use to test these hypotheses. The measure of *sex* (or gender) is relatively straightforward. A variable we can use to measure religiosity (*religten*) was obtained by asking respondents about the strength of their religious affiliation ("strong," "somewhat strong," "not very strong," or "no religion"). Finally, the variable *ethnicity* was created by combining a question asking respondents to identify their race with one asking whether the respondent was Hispanic (which can be of any race). This yields four categories: Anglo (the term used by the U.S. Census in 2010 was "non-Hispanic whites"). African American ("non-Hispanic blacks"), Hispanic, and non-Hispanic other.

Open GSS16A.sav and select all respondents except "non-Hispanic other"[1] for analysis. (Review the procedures described in Chapter 3 for selecting cases.)[2]

Following the instructions in chapter 5, crosstabulate *religten* with *sex* and with *ethnicity* selecting column percentages for the cells. You'll obtain the results shown in Figures 8–1 and 8–2[3]. (We've left out the "case processing summary.")

---

[1] Because there are relatively few cases in this category, and because it combines people who may have little in common in terms of their ethnicity, we are not including them in this analysis.

[2] It's important to weight the cases so they better represent the population from which the sample is selected. Our data set – GSS16A.sav – has already been weighted so you don't need to weight it again.

[3] Note that, since you have elected to exclude "non-Hispanic other" respondents, they will be excluded from all tables, including those crosstabulating *religten* and *sex*. This has the advantage of basing all tables on the same respondents, but at the price of eliminating some you might have wanted to include in your comparison of males and females.

reliten STRENGTH OF RELIGIOUS AFFILIATION * sex RESPONDENT'S SEX Crosstabulation

			sex RESPONDENT'S SEX		
			1 MALE	2 FEMALE	Total
reliten STRENGTH OF RELIGIOUS AFFILIATION	1 STRONG	Count	405	589	994
		% within sex RESPONDENT'S SEX	33.2%	39.9%	36.9%
	2 SOMEWHAT STRONG	Count	52	92	144
		% within sex RESPONDENT'S SEX	4.3%	6.2%	5.3%
	3 NOT VERY STRONG	Count	441	533	974
		% within sex RESPONDENT'S SEX	36.2%	36.1%	36.1%
	4 NO RELIGION	Count	321	264	585
		% within sex RESPONDENT'S SEX	26.3%	17.9%	21.7%
Total		Count	1219	1478	2697
		% within sex RESPONDENT'S SEX	100.0%	100.0%	100.0%

Figure 8-1

reliten STRENGTH OF RELIGIOUS AFFILIATION * ethnicity RACE ETHNICITY Crosstabulation

			ethnicity RACE ETHNICITY			
			1 NON-HISPANIC WHITE	2 NON-HISPANIC BLACK	3 HISPANIC	Total
reliten STRENGTH OF RELIGIOUS AFFILIATION	1 STRONG	Count	643	229	122	994
		% within ethnicity RACE ETHNICITY	34.7%	52.6%	29.9%	36.9%
	2 SOMEWHAT STRONG	Count	98	25	22	143
		% within ethnicity RACE ETHNICITY	5.2%	5.7%	5.4%	5.3%
	3 NOT VERY STRONG	Count	670	108	195	973
		% within ethnicity RACE ETHNICITY	36.2%	24.8%	47.7%	36.1%
	4 NO RELIGION	Count	442	73	70	585
		% within ethnicity RACE ETHNICITY	23.9%	16.8%	17.1%	21.7%
Total		Count	1851	435	409	2695
		% within ethnicity RACE ETHNICITY	100.0%	100.0%	100.0%	100.0%

Figure 8-2

As the results show, women are more likely than men to report a strong or somewhat strong religious affiliation, and are less likely to report that they have no religious affiliation. Differences between African American and Anglo respondents are even greater, with over half of African American respondents, but only about a third of Anglos, reporting a strong religious affiliation, while larger proportions of Anglos than African Americans say they are not very strong in their religion or have no religious affiliation. Hispanics are least likely to report having a strong religious affiliation, but are also less likely than Anglos and about the same as African Americans to report no relgious affilation at all. (In the interest of conserving space, we haven't carried out measures of association or statistical significance, but you may wish to do so yourself.)

This one-step method of hypothesis testing is, however, very limited. It does not, for example, tell us whether African American men differ from African American women in religious intensity, whether there are differences in this regard between Anglo men and Anglo women or between Hispanic men and Hispanic women.

To answer this question, we will do a multivariate cross tabulation, also called an elaboration analysis.

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 three categories of the *ethnicity* variable; thus, we should expect to get three contingency tables, each one showing the relationship between *sex* and *reliten* for Anglos, for African Americans, and for Hispanics.

Open up the crosstabulation dialog box you used for Figures 8–1 and 8–2, but this time adding *ethnicity* in the third box on the right under "Layer 1 of 1." To make the table more compact, click on cells and unselect "Count." The dialog box should now look like Figure 8–3. Click OK.

Figure 8-3

Your results should look like the table shown in Figure 8-4.

reliten STRENGTH OF RELIGIOUS AFFILIATION * sex RESPONDENT'S SEX * ethnicity RACE ETHNICITY Crosstabulation

% within sex RESPONDENT'S SEX

ethnicity RACE ETHNICITY			sex RESPONDENT'S SEX 1 MALE	2 FEMALE	Total
1 NON-HISPANIC WHITE	reliten STRENGTH OF RELIGIOUS AFFILIATION	1 STRONG	31.9%	37.1%	34.7%
		2 SOMEWHAT STRONG	3.8%	6.3%	5.2%
		3 NOT VERY STRONG	35.6%	36.7%	36.2%
		4 NO RELIGION	28.7%	19.9%	23.9%
	Total		100.0%	100.0%	100.0%
2 NON-HISPANIC BLACK	reliten STRENGTH OF RELIGIOUS AFFILIATION	1 STRONG	48.0%	56.5%	52.6%
		2 SOMEWHAT STRONG	4.1%	7.1%	5.7%
		3 NOT VERY STRONG	25.5%	24.3%	24.8%
		4 NO RELIGION	22.4%	12.1%	16.8%
	Total		100.0%	100.0%	100.0%
3 HISPANIC	reliten STRENGTH OF RELIGIOUS AFFILIATION	1 STRONG	23.8%	34.7%	29.8%
		2 SOMEWHAT STRONG	6.5%	4.9%	5.6%
		3 NOT VERY STRONG	49.7%	45.8%	47.6%
		4 NO RELIGION	20.0%	14.7%	17.1%
	Total		100.0%	100.0%	100.0%
Total	reliten STRENGTH OF RELIGIOUS AFFILIATION	1 STRONG	33.3%	38.9%	36.9%
		2 SOMEWHAT STRONG	4.3%	6.2%	5.3%
		3 NOT VERY STRONG	36.1%	36.1%	36.1%
		4 NO RELIGION	26.4%	17.9%	21.7%
	Total		100.0%	100.0%	100.0%

Figure 8-4

Notice that the relationship between *reliten* and *sex* is roughly the same within each ethnic group.

Try other variables as a control (i.e., in place of *ethnicity*) to see what happens. As a general rule, here is how to interpret what you find from this elaboration analysis:

- If the relationship between the independent and dependent variables shown in the partial tables is similar to that shown in the zero-order (original bivariate) table you have *replicated* your original findings, which means that in spite of the introduction of a particular control variable, the original relationship persists. This is indeed the case here: the differences between men and women shown in the partial tables of Figure 8–4 are similar to those shown in Figure 8–1.

- If the difference shown in <u>all</u> 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 differences you see in the partial tables are less than you saw in the original table AND IF your control variable is intervening (that is, the control variable occurs in time <u>after</u> 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.

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 researchers sometimes use multiple regression with ordinal variables as well. Multiple regression also assumes that there is a linear relationship between each independent variable and the dependent variable, and that the distribution of values in your variables follows a normal distribution.

Recall from Chapter 7 that we investigated the impact that Internet freedom had on perceived corruption, and found evidence consistent with our hypothesis that high levels of Internet freedom seem to increase people's sense that they can hold government accountable, thus leading to perceptions of less government corruption. It may be, however, that holding government accountable requires more than the ability to publicize corrupt activities, but also requires the ability to exercise political rights, such as the right to vote in contested elections. In recent years, for example, protesters in some countries have used the Internet to help bring down corrupt regimes, but the absence of effective means to participate in ordinary political institutions has sometimes led to the emergence of new leaders as corrupt as those they replaced.

To test this, open the COUNTRIES.sav file and add the variable *polrights* to the regression equation we ran in Chapter 7. From the menu, click Analyze, Regression, Linear. Click on *honestgov* and move it into the Dependent box at the top of the dialog box. Click on *ifreedom* and *polrights* and move them into the Independent(s) box. The dialog box should look like the one shown in Figure 8-5. Click OK.

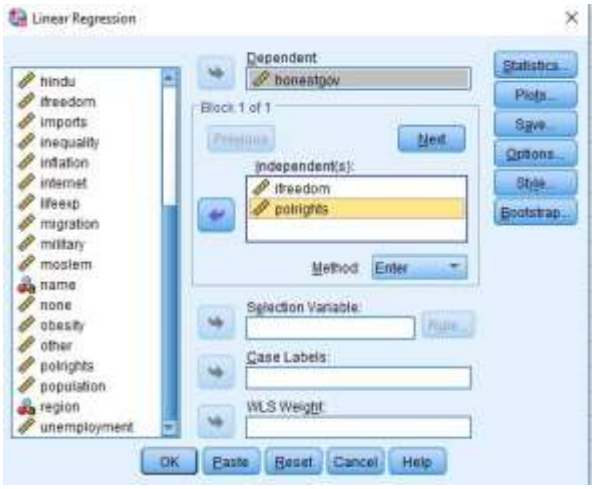

Figure 8-5

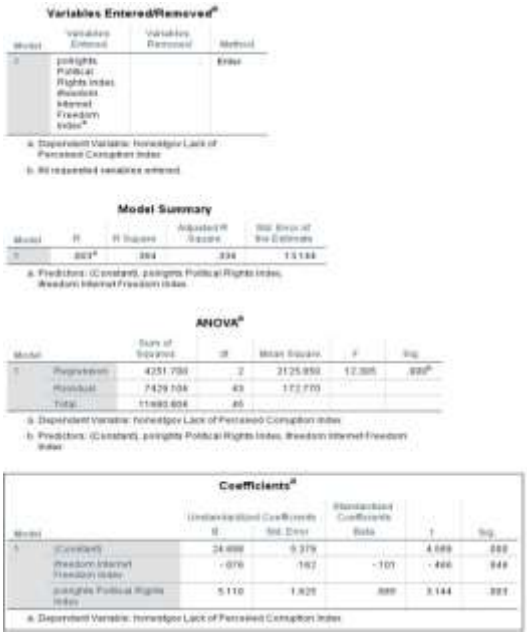

Figure 8-6

Your results should look like those shown in Figure 8-6. Looking first at the Model Summary table, you will see that the adjusted R-squared value is .334. As you recall from Chapter 7, this means that 33.4% of the variation in the dependent variable (perceived honesty in government) is explained by knowing a country's level of Internet freedom and political rights. The ANOVA table shows that the overall model is highly statistically significant. Next, we need to look at the Coefficients table. If you look at the B coefficient for *ifreedom*, you will see that it is -.076. How do we interpret this coefficient? Recall the discussion in Chapter 7: a one unit change in the independent variable (*ifreedom*) is associated with a change in the dependent variable (*honestgov*) equal to the value of B. So, if we increase the value of *ifreedom* by 1, on average, we get a change of -.076 units in *perceived honest in government*. Since the higher the level of the Internet freedom variable, the **lower** the level of perceived honesty in government, the results

are actually in the opposite direction than we had hypothesized. However, the regression coefficient is not statistically significant so we cannot conclude that *ifreedom* is related to *honestgov*. On the other hand, the value of B for *polrights* is 5.110, meaning that an increase of one unit on the political rights index is associated with an increase of 5.11 points on perceived honesty in government. The result is in the hypothesized direction.

However, one problem with interpreting the B coefficients is that the units of measurement we are using are quite different for different variables. Internet freedom and lack of perceived corruption are measured on scales of 0 to 100, whereas political rights are measured on a scale of 1 to 7. We're comparing apples to oranges.

To address this problem, look at the standardized (Beta) coefficients, which we've ignored to this point. Beta coefficients in effect convert all variables to standard scores (with means of 0 and standard deviations of 1). The Beta coefficient for *polrights* (.685) has an absolute value almost seven times as large as that for *ifreedom* (.101). In other words, when each independent variable is controlled for the other, an increase of one standard deviation in *polrights* has an impact on *corruption* that is much greater than that of the same increase in the *ifreedom* measure. Finally, note that the *polrights* is highly statistically significant (p = .003) while *ifreedom* is not at all statistically significant (p=.644).

If we convert the information in the Coefficients table to standard algebraic form (but leaving out the error terms) we get, for the unstandardized equation:

$\hat{Y}=24.688+.076*X_1-5.110*X_2$ where

X_1=*ifreedom* and
X_2=polrights.

The standardized equation looks like:

$\hat{Y}=.101*X_1-.685*X_2$.

The reason why the constant has dropped out of this equation is that, with variables converted to standard scores, it is equal to zero by definition.

Finally, note that the model as a whole only explains about a third of the variance among countries in perceived corruption. Does the dataset include any other variables that you think might explain some of the rest? Add these variables to the equation and see if they help.

Chapter Eight Exercises

Use GSS16A.sav for exercises 1 through 3.

1, Repeat the crosstabs we ran earlier in this chapter, but this time use *race* as the independent variable and *sex* as the control variable.

2. 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.

3. Choose three independent variables from the General Social Survey subset 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.

Use COUNTRIES.sav for exercises 4 and 5.

4. Using the unstandardized regression equation for predicting *honestgov* based on *ifreedom* and *polrights*, calculate the residuals for South Africa, the United Kingdom, and Ukraine. You can either do this manually or, when running the regression analysis, click on "Save" and save the unstandardized residuals as an additional variable, then go to DATA VIEW to find the values of this new variable (which SPSS will call "RES_1") for these countries. (Note that this new variable is calculated only for those countries for which there is no missing data for any of the variables in the equation.) Are the residuals for the United Kingdom and Ukraine less than those we calculated in Chapter 7? Are there other variables that, if added to the equation, might reduce them further?

5. From Appendix B select three variables that you think might help explain inequality of income distribution (*inequality*). Using the COUNTRIES.sav file, run a multiple regression analysis. Which of the three independent variables is the best predictor of *inequality*? How much of the variance among countries in inequality is explained by the model as a whole?