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 are more strongly religious than whites.

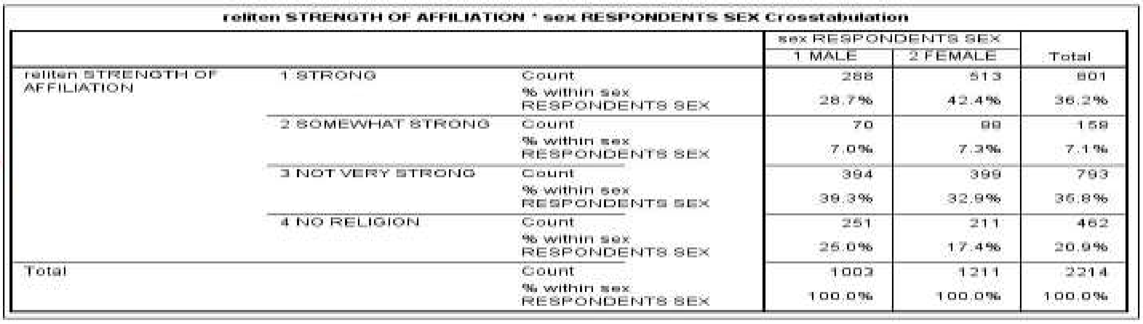
The 2014 General Social Survey provides data that we can use to test these hypotheses. The measure of sex (or gender) is straightforward. A variable we can use to measure religiosity was obtained by asking respondents about the strength of their religious affiliation ("strong," "somewhat strong," "not very strong," or "no religion").

The measure of race is a little more complicated. The codes for this measure are 1) white, 2) black, and 3) other. The number of respondents in this last category is relatively small, and lumps together very different groups (Asian Americans, Native Americans, etc.). For this reason, we will limit our analysis of race to blacks and whites. Another problem with the race variable, and one we'll live with for purposes of this exercise, is that it measures only race and not, as is more common in analysis, race and ethnicity. In particular, it does not distinguish between Hispanic and non-Hispanic white or black respondents.

Open GSS14A.sav and select only white and black respondents for analysis.[1] (Review the procedures described in Chapter 3 for selecting cases.)

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

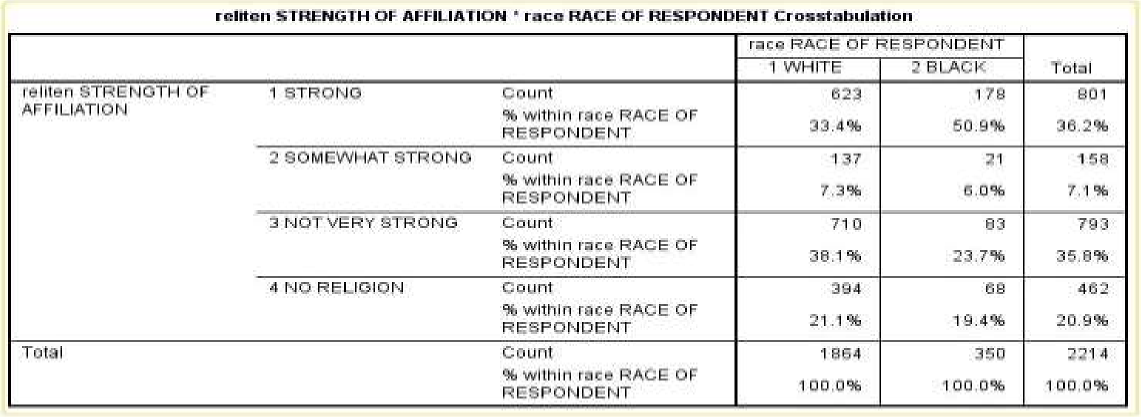
[1] It's important to weight the cases so they better represent the population from which the sample is selected. Our data set – GSS14A,sav – has already been weighted so you don't need to weight it again.



reliten STRENGTH OF AFFILIATION * sex RESPONDENTS SEX Crosstabulation

			sex RESPONDENTS SEX		Total
			1 MALE	2 FEMALE	
reliten STRENGTH OF AFFILIATION	1 STRONG	Count	288	513	801
		% within sex RESPONDENTS SEX	28.7%	42.4%	36.2%
	2 SOMEWHAT STRONG	Count	70	88	158
		% within sex RESPONDENTS SEX	7.0%	7.3%	7.1%
	3 NOT VERY STRONG	Count	394	399	793
		% within sex RESPONDENTS SEX	39.3%	32.9%	35.8%
	4 NO RELIGION	Count	251	211	462
		% within sex RESPONDENTS SEX	25.0%	17.4%	20.9%
Total		Count	1003	1211	2214
		% within sex RESPONDENTS SEX	100.0%	100.0%	100.0%

Figure 8-1



reliten STRENGTH OF AFFILIATION * race RACE OF RESPONDENT Crosstabulation

			race RACE OF RESPONDENT		Total
			1 WHITE	2 BLACK	
reliten STRENGTH OF AFFILIATION	1 STRONG	Count	623	178	801
		% within race RACE OF RESPONDENT	33.4%	50.9%	36.2%
	2 SOMEWHAT STRONG	Count	137	21	158
		% within race RACE OF RESPONDENT	7.3%	6.0%	7.1%
	3 NOT VERY STRONG	Count	710	83	793
		% within race RACE OF RESPONDENT	38.1%	23.7%	35.8%
	4 NO RELIGION	Count	394	68	462
		% within race RACE OF RESPONDENT	21.1%	19.4%	20.9%
Total		Count	1864	350	2214
		% within race RACE OF RESPONDENT	100.0%	100.0%	100.0%

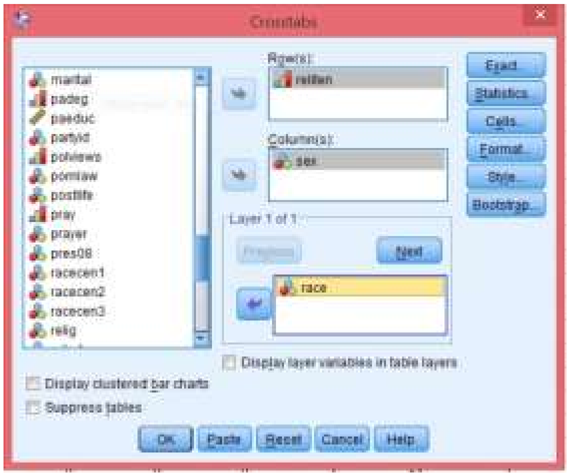
Figure 8-2

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 their affilation is not very strong or that they have no relious affiliation. Differences between black and white respondents are even greater, with over half of black respondents, but only about a third of whites, reporting a strong religious affiliation, while larger proportions of whites than blacks fall into each of the other categories. (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 white men and white 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



Crosstabs

Row(s): reliten

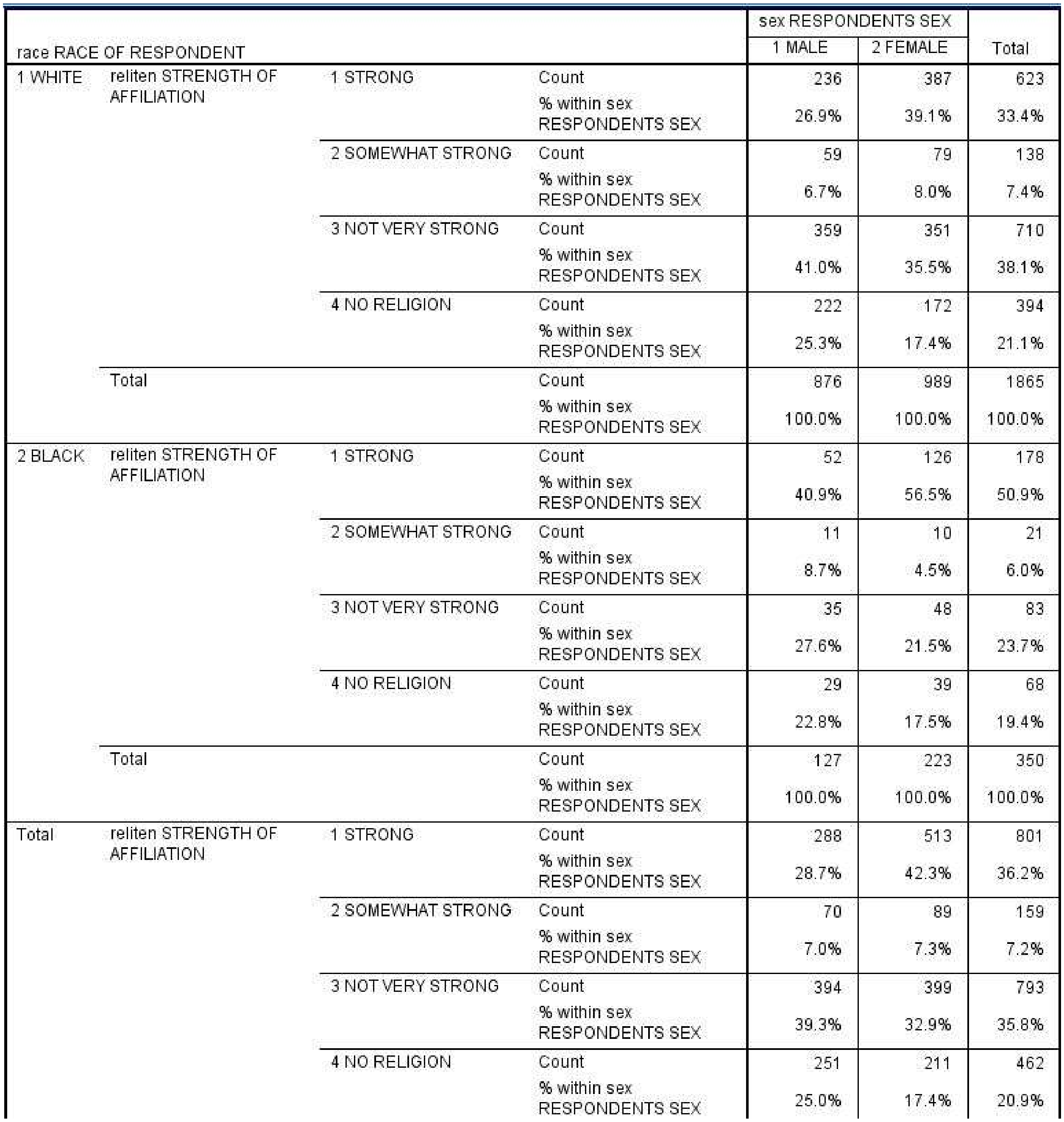
Column(s): sex

Layer 1 of 1: race

Figure 8-3

as many separate tables as there are categories of the control variable. There are two categories of the *race* variable; thus, we should expect to get two contingency tables, each one showing the relationship between *sex* and *reliten* for whites, the other for African Americans.

Open up the crosstabulation dialog box you used for Figures 8–1 and 8–2, but this time adding *race* in the third box on the right under "Layer 1 of 1." The dialog box should now look like Figure 8–3. Click OK. Your results should look like the table shown in Figure 8-4.



				sex RESPONDENTS SEX		Total
race RACE OF RESPONDENT				1 MALE	2 FEMALE	
1 WHITE	reliten STRENGTH OF AFFILIATION	1 STRONG	Count	236	387	623
			% within sex RESPONDENTS SEX	26.9%	39.1%	33.4%
		2 SOMEWHAT STRONG	Count	59	79	138
			% within sex RESPONDENTS SEX	6.7%	8.0%	7.4%
		3 NOT VERY STRONG	Count	359	351	710
			% within sex RESPONDENTS SEX	41.0%	35.5%	38.1%
		4 NO RELIGION	Count	222	172	394
			% within sex RESPONDENTS SEX	25.3%	17.4%	21.1%
	Total		Count	876	989	1865
			% within sex RESPONDENTS SEX	100.0%	100.0%	100.0%
2 BLACK	reliten STRENGTH OF AFFILIATION	1 STRONG	Count	52	126	178
			% within sex RESPONDENTS SEX	40.9%	56.5%	50.9%
		2 SOMEWHAT STRONG	Count	11	10	21
			% within sex RESPONDENTS SEX	8.7%	4.5%	6.0%
		3 NOT VERY STRONG	Count	35	48	83
			% within sex RESPONDENTS SEX	27.6%	21.5%	23.7%
		4 NO RELIGION	Count	29	39	68
			% within sex RESPONDENTS SEX	22.8%	17.5%	19.4%
	Total		Count	127	223	350
			% within sex RESPONDENTS SEX	100.0%	100.0%	100.0%
Total	reliten STRENGTH OF AFFILIATION	1 STRONG	Count	288	513	801
			% within sex RESPONDENTS SEX	28.7%	42.3%	36.2%
		2 SOMEWHAT STRONG	Count	70	89	159
			% within sex RESPONDENTS SEX	7.0%	7.3%	7.2%
		3 NOT VERY STRONG	Count	394	399	793
			% within sex RESPONDENTS SEX	39.3%	32.9%	35.8%
		4 NO RELIGION	Count	251	211	462
			% within sex RESPONDENTS SEX	25.0%	17.4%	20.9%

Figure 8-4

Notice that the relationship between *reliten* and *sex* is about the same for whites and African Americans.

Try other variables as a control (i.e., in place of *race*) 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 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 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 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.

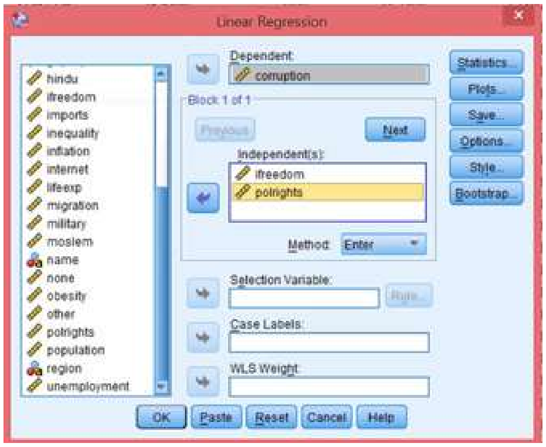
- Last, if a zero order table shows only a weak or moderate association, the partials might show the opposite relationship, due to the presence of a *distorter* variable.

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 *corruption* 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.



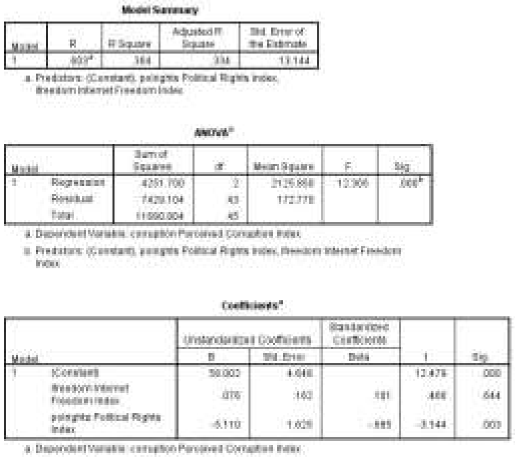
Dependent: corruption

Block 1 of 1

Independent(s): ifreedom, polrights

Method: Enter

Figure 8-5



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.603a	.364	.334	13.144

a. Predictors: (Constant), polrights Political Rights Index, ifreedom Internet Freedom Index

ANOVAb

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4251.700	2	2125.850	12.306	.000a
	Residual	7429.104	43	172.770		
	Total	11680.804	45			

a. Dependent Variable: corruption Perceived Corruption Index

b. Predictors: (Constant), polrights Political Rights Index, ifreedom Internet Freedom Index

Coefficientsa

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	58.003	4.648		12.479	.000
	ifreedom Internet Freedom Index	.076	.162	.081	.468	.644
	polrights Political Rights Index	-5.110	1.625	-.483	-3.144	.003

a. Dependent Variable: corruption Perceived Corruption Index

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 corruption) 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 (*corruption*) 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 *corruption*. Since the higher the level of the *corruption* variable, the **lower** the level of perceived corruption, the results are actually in the opposite direction that we had hypothesized. However, the regression coefficient is not statistically significant so we cannot conclude that *ifreedom* is related to *corruption*. 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 a decrease of 5.11 points on the Perceived Corruption Index. 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 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:

$\hat{Y}=58.002+.076*X_1-5.110*X_2$ 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 GSS14A.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 *corruption* 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, save the residuals as an additional variable, then go to DATA VIEW to find the values of RES_1 for these countries. 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. 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?