Chapter Seven: Correlation and Regression

Correlation and regression analysis (also called "least squares" or "ordinary least squares (OLS)" analysis) helps us examine relationships among interval or ratio variables. In this chapter, we'll explore techniques for doing correlation and bivariate regression. Chapter 8 will include a look at multiple correlation and regression.

To illustrate these techniques, we'll use the "COUNTRIES.sav" file, derived from several sources and containing data on the countries of the world. See Appendix B for a codebook with information on the variables included in this file. Open the file following the instructions in Chapter 1 under "Getting a Data File."

We'll begin by considering the relationship between perceived government corruption and Internet freedom. Our hypothesis will be that in countries where Internet freedom is high, people will have a greater sense that they can hold government accountable (the technical term for this sense is called "political efficacy") and they will tend to regard their system as less corrupt. We'll also add a measure of political rights (rights to "participate freely in the political process") to the mix, primarily for consideration in Chapter 8.

The Perceived Corruption Index (*corruption*) was devised by Transparency International. It uses a scale from 0 to 100 in which 100 represents the lowest possible level of perceived corruption. The Internet Freedom Index (*ifreedom*) is a measure developed by an organization called Freedom House and also uses a scale from 0 to 100 on which the higher the number, the more the restrictions on Internet use. (Unfortunately, while the countries file includes 195 countries, the Internet Freedom Index is available for only 47.) The Political Rights Index (*polrights*), also developed by Freedom House, is on a scale from 1 to 7, in which 1 indicates the highest level of political rights in a country, and 7 the lowest.

For our purposes, the way that Transparency International and Freedom House have coded these variables is rather counter-intuitive. Keep in mind that **higher** values indicate **lower** levels of perceived corruption, Internet freedom, and political rights**.**

Correlation

How close are the relationships among Internet freedom, political rights, and perceived corruption? To find out, click on Analyze, Correlate, and Bivariate. A dialog box will appear on your screen. Click on *corruption* and then click the arrow to move it into the box. Do the same with *ifreedom* and *polrights*. The dialog box should look like Figure 7–1.

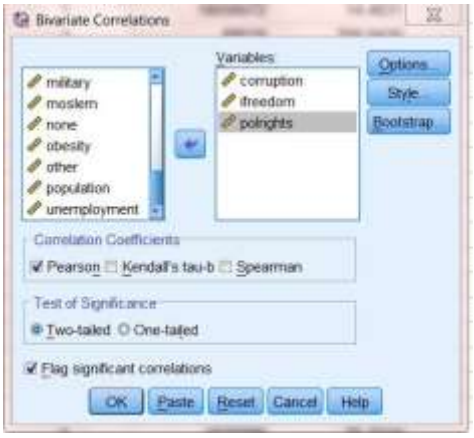

Figure 7-1

The most widely used bivariate test is the Pearson's r correlation coefficient. It is intended to be used when both variables are measured at either the interval or ratio level and each variable is normally distributed. However, sometimes we do violate these assumptions. If you do histograms of our three variables (see Chapters 4 and 9), you will notice that none are actually normally distributed. Furthermore, *polrights* should probably be considered an ordinal, not an interval, measure. We'll use the Pearson's r, but will need to proceed with caution. IBM SPSS includes another correlation test, Spearman's rho, which is designed to analyze variables that are not normally distributed, or are ranked. We will conduct both tests to see how much the results differ depending on the test used—in other words, whether those who use Pearson's r for these variables are seriously off base.

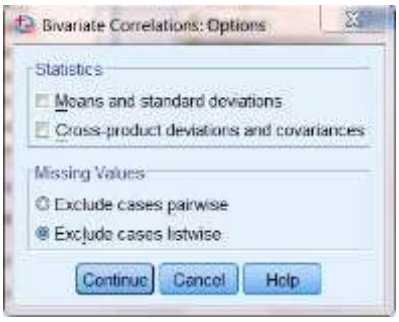

Figure 7-2

In the dialog box, click on Options and, in the resulting box, on Exclude cases listwise. The result should look like Figure 7–2). The reason for doing this is that, as we've noted, *ifreedom* is based on many fewer cases than the other two variables, and we want to be able to make "apples to apples" comparisons. Click on Continue.

The box next to Pearson is already checked, as this is the default. Click in the box next to Spearman. Click the button next to One-tailed test of significance. (This is because we will be testing "directional" hypotheses, that is, not just the idea that two variables are related but, for example, that the **lower** the value of the *ifreedom* index, the **higher** the value of the perceived *corruption* index. Remember how these variables are coded: we are hypothesizing that more Internet freedom is associated with less perceived corruption.) Therefore, we would expect the correlation to be negative. Your dialog box should now look like the one in Figure 7–3. Click OK to run the tests.

Figure 7-3

Your output screen will show two tables (called matrices): one for Pearson's r and one for Spearman's rho. The Pearson's correlation matrix should look like the one in Figure 7–4. The cells of the table show the Pearson's r correlation between each variable and each other variable, the level of statistical significance of the relationship (that is, the likelihood that it could have occurred by chance), and the number of cases on which the correlation is based.

The correlation coefficient may range from -1 to 1, where -1 or 1 indicates a "perfect" relationship. The further the coefficient is from 0, regardless of whether it is positive or negative, the stronger the relationship between the two variables. Thus, a coefficient of .467 is exactly as strong as a coefficient of -.467. Positive coefficients

Correlations[b]

		corruption Perceived Corruption Index	ifreedom Internet Freedom Index	polrights Political Rights Index
corruption Perceived Corruption Index	Pearson Correlation	1	-.467**	-.601**
	Sig. (1-tailed)		.001	.000
ifreedom Internet Freedom Index	Pearson Correlation	-.467**	1	.830**
	Sig. (1-tailed)	.001		.000
polrights Political Rights Index	Pearson Correlation	-.601**	.830**	1
	Sig. (1-tailed)	.000	.000	

**. Correlation is significant at the 0.01 level (1-tailed).

b. Listwise N=46

Figure 7-4

tell us there is a direct relationship: when one variable increases, the other increases. Negative coefficients tell us that there is an inverse relationship: when one variable increases, the other decreases. Notice that the Pearson's r for the relationship between Internet freedom and perceived corruption is -.467. This tells us that, just as we predicted, as Internet freedom increases, perceived corruption decreases. But should we consider the relationship strong? We'll revisit this question later in the chapter.

The correlation matrix also gives the probability that the relationship we have found could have occurred just by chance. (Labeled as Sig. [1-tailed]). The probability value is .001, which is well below the conventional threshold of $p \leq .05$. Thus, our hypothesis is supported. There is a relationship (the coefficient is not 0), it is in the predicted direction (negative), and is statistically significant.

Recall that we had some concerns about using the Pearson's r coefficient. Figure 7–5 shows the results using Spearman's rho. Notice that the coefficient for the relationship between *ifreedom* and *corruption* is -.414, or about the same as the value of Pearson's r for this relationship. Similarly, the other values of Spearman's rho are similar to those for Pearson's r. This is reassuring.

Correlations[b]

			corruption Perceived Corruption Index	ifreedom Internet Freedom Index	polrights Political Rights Index
Spearman's rho	corruption Perceived Corruption Index	Correlation Coefficient	1.000	-.414**	-.539**
		Sig. (1-tailed)	.	.002	.000
	ifreedom Internet Freedom Index	Correlation Coefficient	-.414**	1.000	.851**
		Sig. (1-tailed)	.002	.	.000
	polrights Political Rights Index	Correlation Coefficient	-.539**	.851**	1.000
		Sig. (1-tailed)	.000	.000	.

**. Correlation is significant at the 0.01 level (1-tailed).

b. Listwise N = 46

Figure 7-5

Regression

Let's look more closely at the relationship between *ifreedom* and *corruption* graphically by creating a scatterplot. Click on Graphs, Chart Builder. This will open up the dialog box shown in Figure 7-6. (If you get a message telling you to be sure that the measurement levels of each variable have been set properly, click on **OK**, since this has already been done for you for the "COUNTRIES.sav" file.)

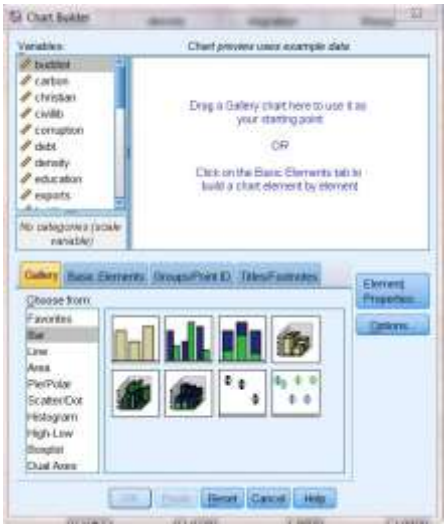

Figure 7-6

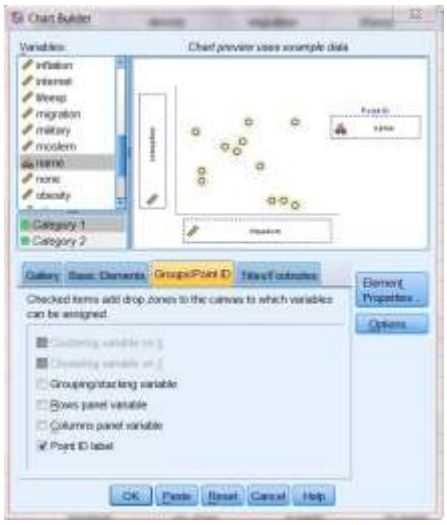

Figure 7-7

Next, in the "Choose from," list at the lower left, click on Scatter/Dot. Then, shift your attention to the sample graph patterns, and click on the first one (upper left). Holding down the mouse button, drag the sample to the large chart preview window. Then, add variables to the chart preview window. From the list of variables, click on *ifreedom* and drag it to the box located on the horizontal (X) axis (because it is the independent variable in our hypothesis and the independent variable belongs on the horizontal axis). Next, click on *corruption* and drag it into the box located on the vertical (Y) axis. Finally, add data labels: from the menu in the middle of the Chart Builder, click on Groups/Point ID**,** select Point ID label and, from the list of variables, click on *name* and drag it to the box on the chart called "Point Label Variable?" (Note: Point ID labels aren't a good idea if you have a large number of cases, but will work well here.) Your dialog box should now look like the one in Figure 7–7. Then, click OK. What you see is a plot of the Perceived Corruption Index for each country included in the chart by each country's Internet Freedom Index. Your scatterplot should look like the one in Figure 7–8.

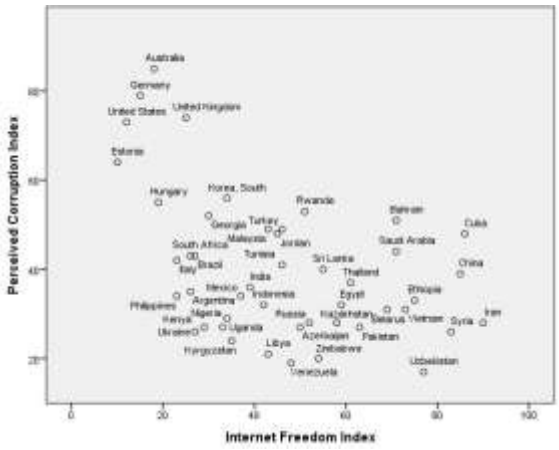

Figure 7-8

You can edit your graph to make it easier to interpret. First, double-click anywhere in the graph. This will cause the graph to open in its own window. On the menu bar, click on Elements, then Fit Line at Total. You will get a dialog box that looks like the one in Figure 7–9. In the Fit Method section, click on Linear (it is the default) and then click on Apply and close the box. (If the **Apply** button is not active, select a different **Fit Method, then change back to Linear** before clicking on **Apply.** If your graph doesn't show country names, click on Elements again, then on Show Data Labels.) Your graph now looks like the one in Figure 7–10. Notice the line variously known as the "least squares line," the "line of best fit," or the "regression line"—we'll go with the last of these— that is now drawn on the graph. Regression and correlation analyze linear relationships between variables, finding the regression line that best fits the data (that is, keeps the errors, the squared distances of each point from the line, to a minimum). Also notice the formula (y=55.95+-.35*x), called the "regression equation," superimposed on the line, and the R-square Linear

Figure 7-9

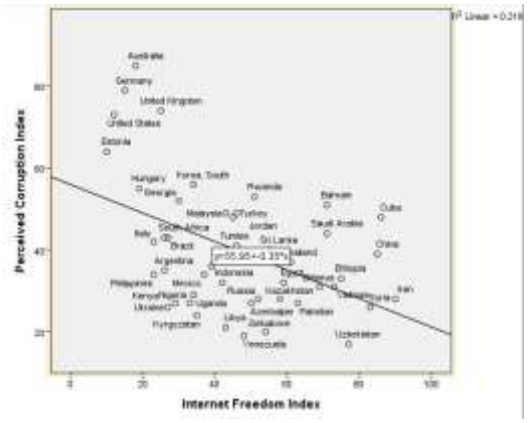

Figure 7-10

statistic (.218) to the right of the graph. We'll return a bit later to the regression equation and the R-square Linear statistic (usually just called "r^2").

In general, countries to the left on the graph (that is, those that have freer Internet access) tend also to be higher on the graph (that is, have less perceived corruption). This is just what we hypothesized. We can now do some "deviant case analysis." Countries that appear above the regression line are those with less perceived corruption than we would expect given their level of Internet freedom, while those below the line have more.

Some countries are pretty much where we'd expect (in that they are close to the line), while some others are well above or below. Can you think of any other factors that might explain the "deviant" cases? We'll return to this question in Chapter 8.

Multiplied by 100, r^2 tells us the percentage of the variation in the dependent variable (*corruption*, on the Y-axis) that is explained by the scores on the independent variable (*ifreedom*, on the X-axis). Thus, Internet freedom explains 21.8% of the variation in perceived corruption. Recall that the Pearson's r coefficient was -.467. If you take the negative square root of .218, you get -.467, the same as the value of r. (If the relationship were positive, you'd take the positive square root.) Though the r statistic is the one most commonly reported, r^2 is extremely useful, since it tells us the "proportional reduction in error" we achieve in "predicting" the value of the dependent variable by knowing that of the independent variable.

How strong a relationship is this? There's no firm answer to this question. One scholar (Karl Deutsch) once suggested that, if you can explain at least 10% of the variance of a variable, you have something worth talking about. If your r^2 exceeds .5 (that is, it explains over 50% of variance), then your knowledge exceeds your ignorance! We would probably consider anything between an r^2 of .1 and .5 (or an r between about ±.3 and ±.7) to be a moderately strong relationship.

Doing a regression analysis can help us to understand the regression line in more detail. Close the IBM SPSS Chart Editor. Click on Analyze, Regression, and Linear. This opens up the dialog box shown in Figure 7-11. Move *corruption* to the Dependent box, and *ifreedom* to the Independent(s) box. Click OK. The results should look like those shown in Figure 7-12.The first table just shows the variables that have been included in the analysis. The second table, "Model Summary," shows the R-square statistic, which is .218 (Where have you seen this before? What does it mean?) (Note: the "Adjusted R Square," .200, is slightly lower because it takes into account the "degrees of freedom" in the equation.)

Figure 7-11

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	ifreedom Internet Freedom Index[b]	.	Enter

a. Dependent Variable: corruption Perceived Corruption Index

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.467[a]	.218	.200	14.410

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

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2544.374	1	2544.374	12.253	.001[b]
	Residual	9136.431	44	207.646		
	Total	11680.804	45			

a. Dependent Variable: corruption Perceived Corruption Index

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	55.953	5.045		11.091	.000
	ifreedom Internet Freedom Index	-.348	.099	-.467	-3.500	.001

a. Dependent Variable: corruption Perceived Corruption Index

Figure 7-12

The third table, ANOVA, gives you information about the model as a whole. ANOVA is discussed briefly in Chapter 6. Note that if you take the Regression Sum of Squares (the variance explained by the relationship) and divide by the Total Sum of Squares, the result is equal to R^2. The final table, Coefficients, gives results of the regression analysis that are not available using only correlation techniques. Look at the "Unstandardized Coefficients" column. Two statistics are reported: "B," which is the regression coefficient, and the standard error. Notice that there are two statistics reported under B, one labeled as "(Constant)," the other labeled as "ifreedom Internet Freedom Index". The latter is the regression coefficient, which is the slope of the line that you saw on the scatterplot. (Note that in scholarly reports, it is conventional to refer to the regression coefficient using the lower case, "b.") The one labeled as

(Constant) is not actually a regression coefficient, but is the Y-intercept (IBM SPSS reports it in this column for convenience only).

What do these numbers mean? You may recall from your statistics course that the formula for a straight line is:

$Y = a + bX$

Y refers to the value of the dependent variable for a given case, a is the Y-intercept (the point where the line crosses the Y-axis, listed as Constant on your output), b is the slope of the line which describes the relationship between the independent and dependent variables, and X is the value of the independent variable for a given case.

We know that the linear relationship between X and Y (*ifreedom* and *corruption*) is not perfect. The correlation coefficient was not 1 (or –1), and the scatterplot showed plenty of cases that did not fall directly on the line. Thus, it is clear to us that knowing a country's level of Internet freedom will not tell us without fail its level of perceived corruption. It is clear that there is some error built into our findings. This is the reason that the regression line is also called the "Best Fit Line." For these reasons, it is conventional to write the formula for the line as:

$\hat{Y} = a + bX + e$, where "e" refers to error. $\hat{Y}$ ("Y hat") indicates the value of Y predicted by the equation for a given case. We could also write it as "Y'" (Y prime) or "Y^c" (the calculated value of Y).

What can we do with this formula? One thing we can do is make predictions about particular values of the dependent variable, using just a little arithmetic. All we have to do is plug the values from our output into the formula for a line. For now, we will ignore the error terms ("e"), but will come back to them shortly. Plugging the numbers from Figure 7–12 into the formula for a straight line, we obtain $\hat{Y}$=55.953+-.348*X, the same equation we saw earlier in Figure 7–10, except that, here, numbers have been carried out to three decimal places. We can then plug in the value of X (*ifreedom*) for any given country, multiply by .348, and subtract that from 55.953. The result will be the predicted value of the *corruption* variable for that country.

For example, looking at the file in Data View mode (see Chapter 2), we see that South Africa, the United Kingdom, and Ukraine all have similar *ifreedom* scores (26, 25, and 27 respectively). Plugging these values into the equation we obtain:

- For South Africa, $\hat{Y}$=55.953+-.348*26=**46.905**.
- For the United Kingdom, $\hat{Y}$=55.953+-.348*25=**47.253.**
- For Ukraine, $\hat{Y}$=55.953+-.348*27=**46.557.**

These numbers represent the predicted values of *corruption* for these three countries, that is, what the values would be if all three countries fell right on the regression line. In other words, we would predict that, since all three countries have similar *ifreedom* scores, they will also have similar *corruption* scores. Going back to Data View, however, we see that the actual scores are 43, 26, and 74 respectively. If we subtract the predicted scores from the actual scores ($Y-\hat{Y}$), we

obtain the "residual," which is a measure of the error in our prediction for a given case. In this example, the residuals are:

- For South Africa, Y-Ŷ=43-46.905=**-3.905**.
- For the United Kingdom Y-Ŷ=74-47.253=**26.747**.
- For Kyrgyzstan, Y-Ŷ=26-46.557=**-20.557**.

In other words, as can be seen in Figure 7-10 above, perceived corruption in South Africa is about what we would expect, whereas it is much less than predicted in the United Kingdom, and much higher than predicted in Ukraine.

We won't go into it here, but you can, for all cases, add the predicted values of the dependent variable and the residuals as additional variables in the data file. To do this, click on SAVE in the regression dialog box, and select Unstandardized Predicted Values and Unstandardized Residuals.

Chapter Seven Exercises

1. Can you think of any other variables included in the codebook in Appendix B that might help explain levels of perceived corruption among countries? Repeat the analysis presented in this chapter, but substitute your variable for *ifreedom*.

2. Pick another variable from the codebook (for example, adult obesity rate). Pick another variable that you think might help explain why some countries have a much higher rate than others. Repeat the analysis presented in this chapter, but substitute your variables for *ifreedom* and *corruption*.

3. The variables in the General Social Survey are mostly nominal or ordinal, but there are some exceptions. In this exercise, we'll use the data set GSS14A.sav and work with two of these variables, the number of hours per week a respondent reports watching television (*tvhours*), and the respondent's age (*age*). Be sure to weight cases (using the weight variable, *wtss*).

   a. It is likely that people of different ages watch different amounts of television. How do you think these may be related? Write a hypothesis that predicts the direction of the relationship between *age* and *tvhours*.

   b. Do a Pearson correlation to test your hypothesis. Was your hypothesis supported? Explain. Remember that whether or not your hypothesis is supported depends on three things: whether or not the coefficient was 0, whether your prediction of the hypothesized direction of the relationship (+ or -) was correct, and the significance (the probability that you will be wrong if you generalize your finding to the population from which the sample was drawn). Be sure to discuss all three in your explanation.

c. Discuss the strength of the relationship between *age* and *tvhours*. Then, speculate about a second factor that might also influence the amount of television that people watch.

d. How much of the variance in *tvhours* is explained by *age*? Tell how you found out.

e. Do a regression analysis of the relationship between *age* and *tvhours*. Be sure to place your variables into their proper boxes (in other words, correctly identify the independent and dependent variable). If you were writing a scholarly report, how would you describe the relationship between *age* and *tvhours* based on your results? (Hint: If it is small, IBM SPSS may have expressed your regression coefficient in scientific notation in order to save space. If you see something like 2.035E-2 on your IBM SPSS output, that is scientific notation. The E-2 is telling you to move the decimal point two places to the left. Thus, 2.035E-2 becomes .02035. If you don't want to move the decimal yourself, click rapidly several times on the coefficient in the output screen and IBM SPSS will show you the actual value of the coefficient.)

f. Do the results of the regression analysis suggest that your hypothesis is supported? Be sure to discuss the magnitude of the regression coefficient, the direction (+ or -), and the probability.

g. How many hours of television does your model predict that people aged 21 tend to watch each day? People aged 42? Show how you calculated these predicted scores.

4. Repeat exercise 3, but this time use *income* as the dependent variable, and *educ* as the independent variable