

Chapter Seven: Correlation and Regression

Correlation and regression analysis (also called "least squares" analysis) helps us examine relationships among interval or ratio variables. As you will see, results of these two tests tell us slightly different things about the relationship between two variables. In this chapter, we'll explore techniques for doing correlation and bivariate regression. Chapter 8 will include a look at multiple regression and correlation.

Correlation

How does education influence the types of occupations that people enter? One way to think about occupations is in terms of "occupational prestige." Your data set includes a variable, PRESTG80, in which a prestige score was assigned to respondents' occupations. Higher numbers indicate greater prestige. (To get more information about how the occupational prestige scale was constructed, you can go to <http://www.csub.edu/~jgross/projects/spss/>)

Let's hypothesize that as education increases, the level of prestige of one's occupation also increases. To test this hypothesis, click on **Analyze, Correlate, and Bivariate**. The dialog box shown in [Figure 7•1](#) will appear on your screen. Click on EDUC, and then click the arrow to move it into the box. Do the same with PRESTG80.

The most widely used bivariate test is the Pearson correlation. 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 a histogram of both EDUC and PRESTG80, you will notice that neither is actually normally distributed. Furthermore, if you noted that PRESTG80 is really an ordinal measure, not an interval, you would be correct. Nevertheless, most analysts would use the Pearson correlation because the variables are close to being normally distributed, the ordinal variable has many ranks, and because the Pearson correlation is the one they are used to. SPSS includes another correlation test, Spearman's rho, that is designed to analyze variables that are not normally distributed, or are ranked, as is PRESTG80. We will conduct both tests to see if our hypothesis is supported, and also to see how much the results differ depending on the test used – in other words, whether those who use the Pearson correlation on these types of variables are seriously off base.

In the dialog box, the box next to **Pearson** is already checked, as this is the default. Click in the box next to **Spearman**. Your dialog box should now look like the one in [Figure 7-2](#). Click **OK** to run the tests.

Your output screen will show two tables: one for the Pearson correlation, and one for the

Spearman's rho. The results of the Pearson's correlation, which is called a correlation matrix, should look like the one in [Figure 7-3](#):

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 $.453$ is exactly as strong as a coefficient of $-.453$. Positive coefficients 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 one decreases. Notice that the Pearson coefficient for the relationship between education and occupational prestige is $.496$, and it is positive. This tells us that, just as we predicted, as education increases, occupational prestige increases. But should we consider the relationship strong? At $.496$, the coefficient is only about half as large as is possible. It should not surprise us, however, that the relationship is not "perfect" (a coefficient of 1). Education appears to be an important predictor of occupational prestige, but no doubt you can think of other reasons why people might enter a particular occupation. For example, someone with a college degree may decide that they really wanted to be a cheese-maker, which has an occupational prestige score of only 29 , while a high-school dropout may one day become an owner of a bowling alley, which has a prestige score of 44 . Given the variety of factors that may influence one's occupational choice, a coefficient of $.496$ suggests that the relationship between education and occupational prestige is actually quite strong.

The correlation matrix also gives the probability of being wrong if we assume that the relationship we find in our sample accurately reflects the relationship between education and occupational prestige that exists in the total population from which the sample was drawn (labeled as Sig. (2-tailed)). The probability value is $.000$ (remember that the value is rounded to three digits), 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 (positive), and we can generalize the results to the population ($p \leq .05$).

Recall that we had some concerns about using the Pearson coefficient, given that PRESTG80 is measured as an ordinal variable. [Figure 7-4](#) shows the results using Spearman's rho. Notice that the coefficient, $.513$, is nearly identical to coefficient obtained using the Pearson correlation. What do you conclude?

Regression

We can also analyze the relationship between education and occupational prestige using regression analysis. But first, let's look at the relationship graphically by creating a scatterplot. Click on **Graphs, Scatter/Dot** and **Define** (we will use the default format, Simple). This will

open up the dialog box shown in [Figure 7•5](#). In the box on the left, click on EDUC then on the arrow key that is pointing toward the box labeled **X Axis** (because it is the independent variable in our hypothesis). Next, click on PRESTG80 and move it into the box labeled **Y Axis** (because it is the dependent variable). Your dialog box should look like the one in [Figure 7-5](#). Then, click **OK**.

What you see is a plot of the number of years of education by the occupational prestige score for persons in the data set who have a job. Your scatterplot should look like the one in [Figure 7-6](#):

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. Double-click on the **X** in the menu bar. A dialog box will open. Click the **Scale** tab at the top of the box. Then, click the check-box next to **Major Increment**. You will see that the box to the right, in the Custom column, will no longer be shaded. Enter the number 2. When you finish, it should look like the one in [Figure 7-7](#). Then, click on **Apply** and then close the box.

Now, 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 7-8](#). In the **Fit Method** section, click on Linear (it is the default) and then click on **Apply** and close the box.

Your graph now looks like the one in [Figure 7-9](#). Notice the Fit Line that is now drawn on the graph. Regression (and correlation) analyze linear relationships between variables, finding the line that “best fits” the data (i.e. it keeps the errors, distances of points from the line, to a minimum). The Fit Line shows you the line that describes the linear relationship. Also notice the R-square statistic listed to the right of the graph. Multiplied by 100, this statistic tells us the percentage of the variation in the dependent variable (PRESTG80, on the Y-axis) that is explained by the scores on the independent variable (EDUC, on the X-axis). Thus, years of education predicts 24.6% of the variation in occupational prestige in our sample. Recall that the Pearson coefficient was .496. If you square the Pearson coefficient ($.496 \times .496$), you get .246 – the same as the R-square! Thus, by knowing the correlation coefficient, you can also know the amount of variance in one variable (dependent) that is explained by the other variable (independent) in a bivariate analysis.

Doing a regression analysis can help us to understand the Fit Line in more detail. Close the SPSS Chart Editor. Click on **Analyze, Regression, and Linear**. This opens up the dialog box shown in [Figure 7•10](#). Move PRESTG80 to the Dependent box, and EDUC to the Independent(s) box. Click **OK**. The results should look like those shown in [Figure 7•11](#).

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 .246. Where have you seen this before? What does it mean?

The third table, ANOVA, gives you information about the model as a whole. ANOVA is discussed briefly in chapter 6. 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 EDUC. The statistic labeled as EDUC 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 (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 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 (B for EDUC), and X is the value of the independent variable for a given case.

We know that the linear relationship between X and Y (EDUC and PRESTG80) 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 someone’s education will not tell us without fail what their occupational prestige is, and furthermore, we are only analyzing a sample of cases and not the whole population to which we want to generalize our findings. It is clear that there is some error built into our findings (the reason that the Fit Line is usually called the “Best Fit Line”). For these reasons, it is conventional to write the formula for the line as

$$\bullet = a + bX + e, \text{ where } e \text{ refers to error.}$$

What can we do with this formula? One thing we can do is make predictions about particular values of the independent 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 our purposes, we will ignore the “e”):

$$\bullet = 12.557 + 2.334X$$

12.557, the Y-intercept (or Constant), is interpreted as the average occupational prestige score (our dependent, or Y variable), holding constant the effects of education (our independent, or X variable). 2.334 is the slope of the line. That is, if you refer back to the scatterplot, if you move one unit to the right on the X-axis, then move 2.565 units upward, you will intersect with the regression line. (It is possible to have a negative coefficient. In that case, to intersect with the line, you would move one unit to the right, and then B units downward.)

What occupational prestige score would our results predict for a person who completed high school, but no higher education? Since X refers to the value of the independent variable, and EDUC is our independent variable, all we have to do is enter 12 (as in twelve years of education) into our equation as follows:

- = $12.557 + 2.334(12)$
- = 40.565

We find that having 12 years of education is associated with an occupational prestige score of 40.565. But what of the error? We know that not every high school graduate has this exact prestige score. We acknowledge this when we discuss results by stating that *on average*, those with 12 years of education will have occupations with prestige scores of 40.565. This language points out to our readers that it is likely that some of those respondents scored higher and some lower, but that 40.62 represents a central point. In sum, the error tells us about the distance from actual values of Y (the answers that the GSS survey respondents gave) and predicted values of Y (the one's you calculate based on the GSS respondent's information in the "X" variable). Thus, the error is the difference between a predicted value of Y for a given case and the actual value of Y for a given case ($\bullet - Y$).

More generally, though, when we discuss regression results, we rarely compute predicted scores for particular values of the independent variable. Instead, in scholarly reports, we usually point out the general process at work. In our case, we would say that "each additional year of education is associated with a 2.334 increase on the occupational prestige scale." Note that we refer to "an additional year of education" because our independent variable was measured as years of school completed. Thus, the "unit" of measurement is years. We say there was a 2.334 increase in prestige with a unit increase in education, because that is the distance we have to move to intersect with the Y-axis, which represents occupational prestige.

Chapter Seven Exercises

1. 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.

2. 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.
3. 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.
4. How much of the variance in TVHOURS is explained by AGE? Tell how you found out.
5. 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, 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 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 SPSS will show you the actual value of the coefficient.]
6. 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.
7. 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.