**STAT14S\_pspp: Exercise Using PSPP to Explore Bivariate Linear Regression**

Author: Ed Nelson
Department of Sociology M/S SS97
California State University, Fresno
Fresno, CA 93740
Email: ednelson@csufresno.edu

Note to the Instructor: The data set used in this exercise is gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav which is a subset of the 2014 General Social Survey. Some of the variables in the GSS have been recoded to make them easier to use and some new variables have been created. The data have been weighted according to the instructions from the National Opinion Research Center. This exercise uses LINEAR REGRESSION in PSPP to explore regression and also uses FREQUENCIES and SELECT CASES. I prepared two documents to help you with PSPP – “[**Notes on Using PSPP**](http://www.ssric.org/files/notes_on_using_PSPP.docx)” and “[**Differences between PSPP and SPSS**](http://www.ssric.org/files/differences_between_PSPP_and_SPSS.docx)” which should answer many of your questions about PSPP. You have permission to use this exercise and to revise it to fit your needs. Please send a copy of any revision to the author. Included with this exercise (as separate files) are more detailed notes to the instructors and the PSPP syntax necessary to carry out the exercises. These, of course, will need to be removed as you prepare the exercise for your students. Please contact the author for additional information.

I’m attaching the following files.

* [**Data subset**](http://ssric.org/files/gss14_subset_for_classes_STATISTICS_pspp.sav) (.sav format)
* [**Extended notes for instructors**](http://ssric.org/files/Extended_Notes_for_Instructors_for_STAT14S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT14S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT14S_pspp.docx) (MS Word; docx format).

**Goals of Exercise**

The goal of this exercise is to introduce bivariate linear regression. The exercise also gives you practice using LINEAR REGRESSION, FREQUENCIES, and SELECT CASES in PSPP.

**Part 1 – Finding the Best Fitting Line to a Scatterplot**

We’re going to use the General Social Survey (GSS) for this exercise. The GSS is a national probability sample of adults in the United States conducted by the National Opinion Research Center (NORC). The GSS started in 1972 and has been an annual or biannual survey ever since. For this exercise we’re going to use a subset of the 2014 GSS. Your instructor will tell you how to access this data set which is called gss14\_subset\_for\_classes\_STATISTICS\_pspp.sav.

In the previous exercise (STAT13S\_pspp) we considered the Pearson Correlation Coefficient which is a measure of the strength of the linear relationship between two interval or ratio variables. In this exercise we’re going to look at linear regression for two interval or ratio variables. An important assumption is that there is a linear relationship between the two variables.

Before we look at these measures let’s talk about outliers. Use FREQUENCIES in PSPP to get a frequency distribution for the variable tv1\_tvhours which is the number of hours that a respondent watches television per day. Look in the “Statistics” box and check the boxes for  “Skewness” and “Kurtosis.” Notice that there are only a few people who watch 14 or more hours of television per day. There are even some who say they watch television 24 hours per day. These are what we call outliers and they can affect the results of our statistical analysis.

PSPP will list the variables and you will select those variables you want to use.  PSPP lists the variables using the variable labels.  However, it’s easier to find the variables if they are listed by variable names.  You can change the way PSPP lists the variables by right clicking anywhere on the list of variables and selecting “Prefer variable labels” and that will list the variables by name.  However, you will have to do this each time you encounter a list of variables.  There is no way to do this permanently.

Let’s exclude these individuals by selecting out only those cases for which tv1\_tvhours is less than or equal to 13. That way the outliners will be excluded from the analysis. To do this you will have to create a PSPP syntax file and then execute the command.  Click on “File” in the menu bar and then on “New” and then on “Syntax.”  This will open a blank syntax file.  In the syntax file enter the following command.  You can do this by cutting and pasting this command into the PSPP syntax file.  Once you have done this click on “Run” in the menu bar and then click on “All.”  Note that you have permanently removed these cases from your data file for this exercise.  So when you complete this exercise do **NOT** save the data file because you will want to use the entire data set for future exercises.  To see your output click on the PSPP icon at the bottom of your screen (i.e., looks like a red circle with a blue cutout at the top).  This will open the output window where you will see your results.

SELECT IF tv1\_tvhours <= 13.

Now use FREQUENCIES to get a frequency distribution for tv1\_tvhours. Remember to ask for “Skewness” and “Kurtosis” (see STAT3S) by checking these boxes in the “Statistics” box.

Compare the frequency distribution before we eliminated the outliers with the distribution after eliminating them. Notice that the skewness and kurtosis values are considerably lower for the distribution after eliminating the outliers than they were before the outliers were dropped. This is because outliers affect our statistical analysis.

Now we’re ready to find the straight line that best fits the data points. The equation for a straight line is Y = a + bX where a is the point where the line crosses the Y-Axis, b is the slope of the line, and Y is the predicted value of Y. Think of the slope as the average change in Y that occurs when X increases by one unit.**[[1]](http://ssric.org/node/578/edit%22%20%5Cl%20%22_edn1%22%20%5Co%20%22)**

Let’s think how we’re going to do that. The best fitting line will be the line that minimizes error where error is the difference between the observed values and the predicted values based on our regression equation. So if our regression equation is Y = 10 + 2X we can compute the predicted value of Y by substituting any value of X into the equation. If X = 5, then the predicted value of Y is 10 + (2)(5) or 20. It turns out that minimizing the sum of the error terms doesn’t work since positive error will cancel out negative error so we minimize the sum of the squared error terms.**[[2]](http://ssric.org/node/578/edit%22%20%5Cl%20%22_edn2%22%20%5Co%20%22)**

**Part II – Getting the Regression Coefficients**

The regression equation will be the values of a and b that minimize the sum of the squared errors. There are formulas for computing a and b but usually we leave it to PSPP to carry out the calculations by running the REGRESSION command.

Click on “Analyze” in the menu bar of PSPP and then click on “Regression” which will open another dropdown menu. Click on “Linear” in the menu. Your dependent variable will be tv1\_tvhours. Enter d1\_age as your independent variable and click on “OK.”

You should see three output boxes.

* The first box tells you the value of the Pearson Correlation Coefficient (R and the correlation coefficient squared (R2)). Age explains or accounts for 4.0% of the variation in tv1\_tvhours. The Adjusted R Square “takes into account the number of independent variables relative to the number of observations.” (George W. Bohrnstedt and David Knoke, Statistics for Social Data Analysis, 1994, F.E. Peacock, p. 293) The standard error is an estimate of the amount of sampling error in this statistic. By the way, notice the output refers to R square. In our example with only one independent variable this is the same as r square. But in the next exercise (STAT15S) we’ll talk about multivariate linear regression where we have two or more independent variables and we’ll explain why this is called R square and not r square.
* The second box is the analysis of variance table that tests the null hypothesis that the Pearson Correlation Coefficient Squared in the population is 0. In this example we reject the null hypothesis since the significance value is less than .05 (or whatever level of significance you’re using which is usually .05 or .01 or .001). This means that age explains more than 0 percent of the variation.
* The third box gives you the regression coefficients.
	+ The slope of the line (b) is equal to 0.02.
	+ The point at which the regression line crosses the Y-Axis is 1.67. This is often referred to as the constant since it always stays the same regardless of which value of X you are using to predict Y.
	+ The standard error of these coefficients which is an estimate of the amount of sampling error.
	+ The standardized regression coefficient (often referred to as Beta). We’ll have more to stay about this in the next exercise but with one independent variable Beta always equals the Pearson Correlation Coefficient (r).
	+ The t test which tests the null hypotheses that the population constant and population slope are equal to 0.
	+ The significance value for each test. As you can see, in this example we reject both null hypotheses. However, we’re usually only interested in the t test for the slope.

The slope is what really interests us. The slope or b tells us that for each increase of one unit in the independent variable (i.e., one year of age) the value of Y increases by an average of .02 units (i.e., number of hours watching television). So our regression equation is Y = 1.67 + .02X. Thus for a person that is 20 years old, the predicted number of hours that he or she watches television 1.67 + (.02) (20) or 1.67 + 0.4 or 2.07 hours.

It’s really important to keep in mind that everything we have done assumes that there is a linear relationship between the two variables. If the relationship isn’t linear, then this is all meaningless.

**Part III – It’s Your Turn Now**

Use the same dependent variable, tv1\_tvhours but this time use d24\_paeduc as your independent variable. This refers to the years of school completed by the respondent’s father.

* Write out the regression equation.
* What do the constant and the slope tell you?
* What are the values of r and r2 and what do they tell you?
* What are the different tests of significance that you can carry out and what do they tell you?

**[[1]](http://ssric.org/node/578/edit%22%20%5Cl%20%22_ednref1%22%20%5Co%20%22)** For example, a unit could be a year in age or a year in education depending on the variable we are talking about.

**[[2]](http://ssric.org/node/578/edit%22%20%5Cl%20%22_ednref2%22%20%5Co%20%22)** When you square a value the result is always a positive number.