**STAT14S: Exercise Using SPSS 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.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 SPSS to explore regression and also uses FREQUENCIES and SELECT CASES. A good reference on using SPSS is SPSS for Windows Version 23.0 A Basic Tutorial by Linda Fiddler, John Korey, Edward Nelson (Editor), and Elizabeth Nelson. The online version of the book is on the [**Social Science Research and Instructional Council's Website**](http://ssric.org/node/582).  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, the SPSS syntax necessary to carry out the exercise (SPSS syntax file), and the SPSS output for the exercise (SPSS output file).  Please contact the author for additional information.

I’m attaching the following files.

* [**Data subset**](http://ssric.org/files/gss14_subset_for_classes_STATISTICS.sav) (.sav format)
* [**Extended notes for instructors**](http://ssric.org/files/Extended_Notes_for_Instructors_for_STAT14S.docx) (MS Word; .docx format)
* [**Syntax file**](http://ssric.org/files/SPSS_Syntax_for_STAT14S.sps) (.sps format)
* [**Output file**](http://ssric.org/files/SPSS_Output_for_STAT14S.spv) (.spv format)
* [**This page**](http://ssric.org/files/STAT14S.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 SPSS.

**Part I – 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.sav.

In the previous exercise (STAT13S) 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 SPSS to get a frequency distribution for the variable tv1\_tvhours which is the number of hours that a respondent watches television per day. Before you click on “OK” click on the “Statistics” button and select “Skewness” and “Kurtosis.” Notice that there are only a few people (i.e., 14) who watch 14 or more hours of television per day. There’s even one person who says he or she watches television 24 hours per day. These are what we call outliers and they can affect the results of our statistical analysis.

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. (See Chapter 3, Using Select Cases in the online SPSS book mentioned on page 1.) Click on “Data” in the menu bar in SPSS and then click on “Select Cases” in the dropdown menu. Click on the circle next to “If condition is satisfied” and then click on the “If” button directly below it. Scroll down the list of variables on the left and select tv1\_tvhours and click on the arrow pointing to the right to move it into the box in the upper right. Then click on the <= button, the 1 button, and the 3 button so the expression in the box reads “tv1\_tvhours <= 13”. Finally click on “Continue” and then on “OK”. Use FREQUENCIES to get a frequency distribution for tv1\_tvhours after eliminating the 14 outliers. Remember to click on the “Statistics” button and ask for “Skewness” and “Kurtosis” (see STAT3S).

Now let’s 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/511/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/511/edit%22%20%5Cl%20%22_edn2%22%20%5Co%20%22)**

Create the scatterplot (SPSS calls this Scatter/Dot) for d1\_age and tv1\_tvhours putting your dependent variable (tv1\_tvhours) on the Y-Axis and your independent variable (d1\_age) on the X-Axis. (See STAT13S for more detailed instructions on how to create the scatterplot.) Now let’s have SPSS draw the regression line on the graph. Double click anywhere inside the scatterplot. This will open the Chart Editor. Click on “Elements” in the menu bar of the Chart Editor and then click on “Fit line at total” which will enter the best fitting regression line on the graph. Click anywhere outside the Chart Editor and your scatterplot will contain the regression line.

**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 SPSS to carry out the calculations.

Click on “Analyze” in the menu bar of SPSS 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 four output boxes.

* The first box lists the variables you entered and reminds you which is your dependent variable.
* The second box tells you the value of the Pearson Correlation Coefficient and the correlation coefficient squared. Age explains or accounts for 3.8% 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. 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 third 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 fourth box gives you the regression coefficients.
	+ The slope of the line (b) is equal to 0.023.
	+ The point at which the regression line crosses the Y-Axis is 1.667. 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 .023 units (i.e., number of hours watching television). So our regression equation is Y = 1.667 + .023X. Thus for a person that is 20 years old, the predicted number of hours that he or she watches television 1.667 + (.023) (20) or 1.667 + 0.46 or 2.127 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. Get the scatterplot, tell SPSS to plot the regression equation on the scatterplot, and ask SPSS to give you the regression equation.

* 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/511/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/511/edit%22%20%5Cl%20%22_ednref2%22%20%5Co%20%22)** When you square a value the result is always a positive number.