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

**Note to the Instructor:** This exercise uses the 2014 General Social Survey (GSS) and SDA to explore bivariate linear regression.  SDA (Survey Documentation and Analysis) is an online statistical package written by the Survey Methods Program at UC Berkeley and is available without cost wherever one has an internet connection.  The 2014 Cumulative Data File (1972 to 2014) is also available without cost by clicking [here](http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14).  For this exercise we will only be using the 2014 General Social Survey.  A weight variable is automatically applied to the data set so it better represents the population from which the sample was selected.  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 exercise itself.  Please contact the author for additional information.

I’m attaching the following files.

* [Extended notes for instructors](http://ssric.org/files/Extended_Notes_for_Instructors_for_STAT14S_SDA.docx) (MS Word; .docx format).
* [This page](http://ssric.org/files/STAT14S_SDA.docx) (MS Word; .docx format).

I’m attaching the following files.

## Goals of Exercise

The goal of this exercise is to introduce bivariate linear regression. The exercise also gives you practice using REGRESSION in SDA.

### 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 the 2014 GSS.  To access the GSS cumulative data file in SDA format click [here](http://sda.berkeley.edu/sdaweb/analysis/?dataset=gss14).  The cumulative data file contains all the data from each GSS survey conducted from 1972 through 2014.  We want to use only the data that was collected in 2014.  To select out the 2014 data, enter year(2014) in the Selection Filter(s) box.  Your screen should look like Figure 14\_1.  This tells SDA to select out the 2014 data from the cumulative file.

**Figure 14-1**

Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.  Notice also that in the SAMPLE DESIGN line SRS has been selected.

The GSS is an example of a social survey.  The investigators selected a sample from the population of all adults in the United States.  This particular survey was conducted in 2014 and is a relatively large sample of approximately 2,500 adults.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.  Often we want to describe respondents in terms of social characteristics such as marital status, education, and age.  These are all variables in the GSS.

In the previous exercises ([STAT13.1S\_SDA](http://www.ssric.org/node/608) and [STAT13.2S\_SDA](http://www.ssric.org/node/609)) 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 SDA to get a frequency distribution for the variable tvhours which is the number of hours that a respondent watches television per day. Click on OUTPUT OPTIONS and check the box for SUMMARY STATISTICS so you will get the skewness and kurtosis measures.  Click also on CHART OPTIONS and then click on TYPE OF CHART and select BAR CHART.  Now click on RUN THE TABLE to produce the frequency distribution and the summary statistics.  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 only those cases for which tvhours is less than 14. That way the outliners will be excluded from the analysis. To do this add tvhours(0-13) to the SELECTION FILTER(S) box.  Be sure to separate year(2014) and tvhours(0-13) with a space or a comma.  This will tell SDA to select out only those cases for which year is equal to 2014 **and** tvhours is less than 14.  Rerun FREQUENCIES in SDA to get a frequency distribution for tvhours after eliminating the outliers and check to make sure that you did it correctly.

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. (See exercise [STAT3S\_SDA](http://www.ssric.org/node/600) for a discussion of skewness and kurtosis.)

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]](" \l "_ftn1" \o ")

Let’s think about 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]](" \l "_ftn2" \o ")

### 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 SDA to carry out the calculations.

Click on REGRESSION at the top of the SDA page and enter your dependent variable (tvhours) in the DEPENDENT box.  We’re going to use age as our independent variable so enter age into the INDEPENDENT box.  Make sure that the WEIGHT and SELECTION FILTER(S) boxes are filled in appropriately for both year and age (see instructions in Part I) and that you have selected SRS in the SAMPLE DESIGN line.  Under OTHER STATISTICS uncheck the box for GLOBAL TESTS.  We’ll use this in the next exercise.  Under MATRICES TO DISPLAY, check the box for CORRELATION MATRIX.  Your screen should look like Figure 14-2.  Now click RUN REGRESSION to produce the regression analysis.

**Figure 14-2**

The first three boxes in your output are what you want to look at.

* The first box lists the variables you entered as your dependent, independent, weight, and filter variables.
* The second 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 value of the Multiple R and Multiple R-Squared.  With only one independent variable this is the same as the Pearson Correlation and the Pearson Correlation squared. Age explains or accounts for 3.8% of the variation in 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\_SDA](http://www.ssric.org/node/611)) 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 shows the Pearson Correlation for our two variables. Note that it is the same as the Beta value when you have only one independent variable.  They look different only because Beta is carried out to three decimal places and the Pearson Correlation is only carried out to two decimals.

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 is 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, tvhours, but this time use paeduc as your independent variable. This refers to the years of school completed by the respondent’s father. Tell SDA to give you the regression equation using the same specifications we used in Part II.

* 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]](" \l "_ftnref1" \o ") For example, a unit could be a year in age or a year in education depending on the variable we are talking about.

[[2]](" \l "_ftnref2" \o ") When you square a value the result is always a positive number.