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 measures of correlation.  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_STAT13.1S_SDA.docx) (MS Word; .docx format).
* [This page](http://ssric.org/files/STAT13.1S_SDA.docx) (MS Word; .docx format).

## Goals of Exercise

The goal of this exercise is to introduce measures of correlation. The exercise also gives you practice using CORRELATION MATRIX in SDA.

### Part I – Getting Started

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 13.1-1.  This tells SDA to select out the 2014 data from the cumulative file.


**Figure 13.1-1**

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.

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 a previous exercise ([STAT11S\_SDA](http://www.ssric.org/node/606)) we considered different measures of association that can be used to determine the strength of the relationship between two variables that have nominal or ordinal level measurement (see exercise [STAT1S\_SDA](http://www.ssric.org/node/598)). In this exercise we’re going to look at two different measures that are appropriate for interval and ratio level variables. The terminology also changes in the sense that we’ll refer to these measures as correlations rather than measures of association.

### Part II - Pearson Correlation Coefficient

The Pearson Correlation Coefficient (r) is a numerical value that tells us how strongly related two variables are. It varies between -1 and +1. The sign indicates the direction of the relationship. A positive value means that as one variable increases, the other variable also increases while a negative value means that as one variable increases, the other variable decreases. The closer the value is to 1, the stronger the linear relationship and the closer it is to 0, the weaker the linear relationship.

The usual way to interpret the Pearson Coefficient is to square its value. In other words, if r equals .5, then we square .5 which gives us .25. This is often called the Coefficient of Determination. This means that one of the variables explains 25% of the variation of the other variable. Since the Pearson Correlation is a symmetric measure in the sense that neither variable is designated as independent or dependent we could say that 25% of the variation in the first variable is explained by the second variable or reverse this and say that 25% of the variation in the second variable is explained by the first variable. It’s important not to read causality into this statement. We’re not saying that one variable causes the other variable. We’re just saying that 25% of the variation in one of the variables can be accounted for by the other variable.

The Pearson Correlation Coefficient assumes that the relationship between the two variables is linear. This means that the relationship can be represented by a straight line. In geometric terms, this means that the slope of the line is the same for every point on that line. Here are some examples of a positive and a negative linear relationship and an example of the lack of any relationship.



Pearson r would be positive and close to 1 in the left-hand example, negative and close to -1 in the middle example, and closer to 0 in the right-hand example. You can search for “free images of a positive linear relationship” to see more examples of linear relationships.

But what if the relationship is not linear? Search for “free images of a curvilinear relationship” and you’ll see examples that look like this.



Here the relationship can’t be represented by a straight line. We would need a line with a bend in it to capture this relationship. While there clearly is a relationship between these two variables, Pearson r would be closer to 0. Pearson r does not measure the strength of a curvilinear relationship; it only measures the strength of linear relationships.

Another way to think of correlation is to say that the Pearson Correlation Coefficient measures the fit of the line to the data points. If r was equal to +1, then all the data points would fit on the line that has a positive slope (i.e., starts in the lower left and ends in the upper right). If r was equal to -1, then all the data points would fit on the line that has a negative slope (i.e., starts in the upper left and ends in the lower right).

Let’s get the Pearson Coefficient for two variables that measure the amount of education of the respondent’s father and mother.  These variables are named maeduc and paeduc in the GSS.  Click on CORRELATION MATRIX at the top of the SDA screen and enter the two variables in the dialog box.  It doesn’t matter which you enter first.  Notice that the SELECTION FILTER(S) and the WEIGHT boxes are filled in.  Also notice that SDA has checked the box for the Pearson Correlation which is what we want.  Listwise has been selected in the MISSING-DATA EXCLUSION box.  That means that any case with missing data for either of these two variables will be excluded from analysis.  Your screen should look like Figure 13.1-2.  Now click on RUN CORRELATIONS to produce the Pearson Correlations.


**Figure 13.1-2**

You should see four correlations. The correlations in the upper left and lower right will be 1 since the correlation of any variable with itself will always be 1. The correlation in the upper right and lower left will both be 0.71. That’s because the correlation of variable X with variable Y is the same as the correlation of variable Y with variable X. Pearson r is a symmetric measure (see exercise [STAT11S\_SDA](http://www.ssric.org/node/606)) meaning that we don’t designate one of the variables as the dependent variable and the other as the independent variable. You don’t see r’s that big very often. That’s telling us that the linear regression line that we’re going to talk about in [STAT14S\_SDA](http://www.ssric.org/node/610) fits the data points reasonably well.

### Part III – Now it’s Your Turn Again

Use CORRELATION MATRIX in SDA to get the Pearson Correlation Coefficient for the years of school completed by the respondent (educ) and the spouse’s years of school completed (speduc). What does this Pearson Correlation Coefficient tell you about the relationship between these two variables?

### Part IV – Correlation Matrices

What if you wanted to see the values of r for a set of variables? Let’s think of the four variables in Parts 2 and 3 as a set. That means that we want to see the values of r for each pair of variables. This time move all four of the variables into the VARIABLES TO CORRELATE box (i.e., educ, maeduc, paeduc, and speduc) and click on RUN CORRELATIONS.   That would mean we would calculate six coefficients. (Make sure you can list all six.)

What did we learn from these correlations? First, the correlation of any variable with itself is 1. Second, the correlations above the 1’s are the same as the correlations below the 1’s. They’re just the mirror image of each other. That’s because r is a symmetric measure. Third, all the correlations are fairly large. Fourth, the largest correlations are between father’s and mother’s education and between the respondent’s education and the spouse’s education.

### Part V – Eta-Squared

The Pearson Correlation Coefficient assumes that both variables are interval or ratio variables (see [STAT1S\_SDA](http://www.ssric.org/node/598)). But what if one of the variables was nominal or ordinal and the other variable was interval or ratio? This leads us back to one-way analysis of variance which we discussed in exercise [STAT8S\_SDA](http://www.ssric.org/node/603).

Click on MEANS in the menu bar at the top of SDA and enter the variable tvhours in the DEPENDENT box and degree in the ROW box.  The SELECTION FILTER(S) and the WEIGHT boxes should still be filled in.  In STAT8S\_SDA we wanted to determine if the differences between levels of education are statistically significant so we carried out a one-way analysis of variance.  Make sure that the MEAN STATISTIC TO DISPLAY box says MEANS.  This is the default so it should.  Then click on OUTPUT OPTIONS and check the box for ANOVA STATS under OTHER OPTIONS.  Be sure to uncheck the box for COMPLEX STANDARD ERRORS and check the box for SRS STANDARD ERRORS.  Finally, click RUN THE TABLE to carry out the procedure.

The F test in the one-way analysis of variance tells us to reject the null hypothesis that all the population means are all equal. So we know that at least one pair of population means are not equal. But that doesn’t tell us how strongly related these two variables are. The SDA output tells us that eta-squared is equal to .051. You should be able to locate this in the Analysis of Variance output table.  This tells us that 5.1% of the variation in the dependent variable, number of hours the respondent watches television, can be explained or accounted for by the independent variable, highest education degree. This doesn’t seem like much but it’s not an atypical outcome for many research findings.

### Part VI – Your Turn

In exercise STAT8S\_SDA you computed the mean number of hours that respondents watched television (tvhours) for each of the nine regions of the country (region). Then you determined if these differences were statistically significant by carrying out a one-way analysis of variance. Repeat the one-way analysis of variance but this time focus on eta-squared. What percent of the variation in television viewing can be explained by the region of the country in which the respondent lived?