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

**Note to the Instructor:** This exercise uses the 2014 General Social Survey (GSS) and SDA to explore dummy variable 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_STAT16S_SDA.docx) (MS Word; .docx format).
* [This page](http://ssric.org/files/STAT16S_SDA.docx) (MS Word; .docx format).

## ****Goals of Exercise****

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

### ****Part I –Dummy Variables****

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 16\_1.  This tells SDA to select out the 2014 data from the cumulative file.

**Figure 16-1**

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 a previous exercise [STAT14S\_SDA](http://www.ssric.org/node/610) we considered linear regression for one independent and one dependent variable which is often referred to as bivariate linear regression.  Multiple linear regression ([STAT15S\_SDA](http://www.ssric.org/node/611) expands the analysis to include multiple independent variables.  In both these exercises the variables in the regression analysis were interval or ratio ([STAT1S\_SDA](http://www.ssric.org/node/598)). What do you do if you want to include a nominal or ordinal variable as one of your independent variables in the regression?

The answer is to create dummy variables.  Consider the respondent’s sex.  Since the variable sex has two categories (i.e., 1 for males and 2 for females), you could create two dummy variables – one for each value.  Your dummy variables would look like this.

* The dummy variable for males will equal 1 if male and 0 if female.
* The dummy variable for females will equal 1 if female and 0 if male.

In other words, you could create as many dummy variables as there are categories of your variable.

If there are k categories, then you would use k – 1 of the dummy variables in your regression analysis.  The category that you omit becomes your comparison group.  We could enter the dummy variable for males into the regression analysis and omit the dummy variable for females.  That means that females will be the comparison group.

What if you had more than two categories?  For example, the region where the respondent lives (region) has nine categories.  So you could create nine dummy variables and omit one of them.  If we decide to omit the category for the Pacific region (value 9), then you would use eight dummy variables in your regression analysis, one for each of the other categories, and the Pacific region would be our comparison group.

Neither of these two variables – sex and region – have missing data but if the variable for which you are creating dummy variables has missing data you need to be careful to exclude those cases with missing data from the analysis.  You want to be careful not to include them in one of your dummy variables.  SDA will do this automatically.

Let’s use tvhours as our dependent variable as we did in the previous two exercises (STAT14S\_SDA and STAT15S\_SDA).  Run FREQUENCIES in SDA to get the frequency distribution for tvhours.  In the previous exercises we discussed outliers and noted that there are a few individuals (i.e., 14) who watched a lot of television which we defined as 14 or more hours per day.  We also noted that outliers can affect our statistical analysis so we decided to remove these outliers from our 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.  Run FREQUENCIES in SDA to get a frequency distribution for tvhours after eliminating the outliers and check to make sure that there are no values greater than 13.

Click on REGRESSION at the top of the SDA page. To create the dummy variables you enter the variable name in the INDEPENDENT box.[[1]](" \l "_ftn1" \o ")  For example, enter sex in the INDEPENDENT box.  Following the variable name, enter (m:X) where X stands for the code or value of the category that you want to omit.  For example, if you entered sex(m:2) that would mean that you are omitting the dummy variable for code or value 2 (i.e., females).  So females becomes your comparison group.

### ****Part II – Regression with Dummy Variables****

Click on REGRESSION at the top of the SDA page and enter your dependent variable (tvhours) in the DEPENDENT box.  Enter the instructions for creating the dummy variable in the INDEPENDENT box.  In our case, we would enter sex(m:2).  (Don’t enter the period.)  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 MATRICES TO DISPLAY, check the box for CORRELATION MATRIX.  Under OTHER STATISTICS, uncheck the GLOBAL TESTS box which we won’t need in this analysis.  Your screen should look like Figure 16-2.  Now click RUN REGRESSION to produce the regression analysis.

**Figure 16-2**

Let’s look at the output that contains your unstandardized regression coefficients.  From this you can see that your regression equation for predicting tvhours is 2.679 + .129 X where X is your dummy variable.  Remember that your dummy variable equals 1 if the person is male and 0 if the person is female.  So for males the predicted number of hours watching television is 2.679 + .129 (1) or 2.808.  For females the predicted number of hours is 2.679 + .129 (0) or 2.679.  Since we left the dummy variable for females out of the regression equation, females become our comparison group.  The unstandardized regression coefficient (0.129) is the mean number of hours that males watch television (2.81) minus the mean for females (2.68) which is 0.13.[[2]](" \l "_ftn2" \o ")

SDA will also calculate t tests to test the null hypotheses that the regression coefficients **in the population** are equal to 0.  Normally we’re only interested in the slope.  The t value for B is 1.301 and the significance value is .194.  This means that we can’t reject the null hypothesis.  In others words, we have no basis for asserting that the population slope is significantly different from zero.  The Pearson Correlation Coefficient Squared tells us that the dummy variable for sex explains virtually none of the variation in the dependent variable.

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

Use SDA to get the frequency distribution for race.  There are three categories for this variable – white (value 1), black (value 2), and other (value 3).  We want to compare whites with non-whites.  This means that there will be two dummy variables:

* The dummy variable for whites will equal 1 if the person is white and 0 if the person is non-white.
* The dummy variable for non-whites will equal 1 if the person is nonwhite and 0 if the person is white.

Let’s make non-whites our comparison group so that means that we’ll leave the dummy variable for non-whites out of the regression equation.  To do this in SDA, enter the following in the INDEPENDENT box – race(m:2-3).  (Don’t enter the period.)  This means that the values 2 and 3 will form the comparison group.

Now run the regression analysis with tvhours as your dependent variable and race(m:2-3) as your independent variable.

* Write out the regression equation.
* What do the unstandardized multiple regression coefficients 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?

### ****Part IV – Multiple Regression with Dummy Variables****

In STAT15S\_SDA you did a regression analysis with tvhours as your dependent variable and age, paeduc, and educ as your independent variables.  This time we’re going to add the dummy variable for males into the analysis.  Use SDA to carry out the regression analysis for this model.

The regression equation for predicting tvhours is 3.372 + .021 (age) - .055 (paeduc) - .086 (educ) + .223 (dummy variable for males).  The unstandardized regression coefficients show the average change in the dependent variable when the independent variable increases by one unit after statistically adjusting for the other independent variables in the equation.  As age increases, television viewing increases but as the respondent’s education and father’s education increase, television viewing goes down.  Males watch more television that females.  The t tests show that all the unstandardized coefficients are statistically significant meaning that that we can reject the null hypotheses that they are zero in the population.  The Pearson Multiple Correlation Coefficient Squared tells us that together the independent variables explain or account for 9.6% of the variation in television viewing.  The Adjusted Squared Correlation Coefficient adjusts for the number of independent variables and is slightly lower (9.3%).  The Beta values show the relative importance of the independent variables in predicting television viewing and tell us that age is the most important and sex the least important with both respondent’s education and father’s education in between.

### ****Part V – Now it’s Your Turn Again****

Repeat the regression analysis you did in Part 4 but instead of adding the dummy variable for males into the analysis, this time add the dummy variable you created in Part 3 for race.  This means you will have four independent variables -- age, paeduc, educ, and the dummy variable for race.

* Write out the regression equation.
* What do the unstandardized multiple regression coefficients tell you?
* What do the standardized regression coefficients (Beta) 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 ") We only use dummy variables for the independent variables and not for the dependent variable.

[[2]](" \l "_ftnref2" \o ") The difference between .129 and 0.13 is due to the fact that SDA calculated the means to two decimal points and the regression coefficient to three decimal points.