**STAT16S\_pspp: Exercise Using PSPP to Explore Dummy Variable 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 dummy variable regression and also uses FREQUENCIES, SELECT CASES, and COMPUTE.  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 exercise. 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_STAT16S_pspp.docx) (MS Word; .docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT16S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT16S_pspp.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 LINEAR REGRESSION, FREQUENCIES, SELECT CASES, and COMPUTE in PSPP.

**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 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 a previous exercise (STAT14S\_pspp) we considered linear regression for one independent and one dependent variable which is often referred to as bivariate linear regression.  Multiple linear regression (STAT15S\_pspp) expands the analysis to include multiple independent variables.  In both these exercises the variables in the regression analysis were interval or ratio (STAT1S\_pspp).  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.  The variable d5\_sex has two categories – 1 for males and 2 for females.  What we do is to create two dummy variables – one for males and the other for females.  Here’s how we do it:

* d5\_sex\_male = 1 if male and 0 if female, and
* d5\_sex\_female = 1 if female and 0 if male.

If there are k categories, then you use k – 1 of the dummy variables in your regression analysis.  The category that you omit becomes your comparison group.  We’re going to enter d5\_sex\_male into the analysis and omit d5\_sex\_female.  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 (d25\_region) has nine categories.  So you would create nine dummy variables and omit one of them.  Actually, you wouldn’t need to create all nine dummy variables since you’re going to omit one.  If we decide to omit the category for the Pacific region (value 9), then you would create eight dummy variables, one for each of the other categories, and the Pacific region would be our comparison group.

Neither of these two variables – d5\_sex and d25\_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.

Let’s use tv1\_tvhours as our dependent variable as we did in the previous two exercises (STAT14S\_pspp and STAT15S\_pspp).  Run FREQUENCIES to get the frequency distribution for tv1\_tvhours.  In the previous exercises we discussed outliers and noted that there are a few individuals 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.

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.

To remove these outliers you will have to create a PSPP syntax file and then execute the file.  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.

SELECT IF tv1\_tvhours <= 13.

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.

Now use FREQUENCIES again to get the frequency distribution for tv1\_tvhours and make sure that you correctly removed the outlines.  You should not see any cases with more than 13 hours.

To create the dummy variable for males (d5\_sex\_males) click on Transform in the menu bar for PSPP and then click on “Compute Variable.”  Enter the variable name, d5\_sex\_males, in the target variable box and enter 0 in the “Numeric Expression” box.  Then click on “OK.”  This will assign the value 0 to all cases.

Now you want to change the value for d5\_sex\_males to 1 for all the males in the sample.  To this you will have to create a PSPP syntax file and then execute the file.  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.  Remember that 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.”

IF (d5\_sex = 1) d5\_sex\_males=1.

**Part II – Regression with Dummy Variables**

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 the dummy variable for males (d5\_sex\_males) as your independent variable.  Remember that you are omitting the dummy variable for females (d5\_sex\_females) so this becomes your comparison group.

Let’s look at the output box that contains your unstandardized regression coefficient.  From this you can see that your regression equation for predicting tv1\_tvhours is 2.68 + .13X where X is your dummy variable.  Remember that your dummy variable, d5\_sex\_males, 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 2.68 + .13 (1) or 2.81.  For females the predicted number of hours is 2.68 + .13 (0) or 2.68.  Since we left the dummy variable for females (d5\_sex\_females) out of the regression equation, females become our comparison group.  The unstandardized regression coefficient (0.13) is the mean number of hours that males watch television (2.81) minus the mean for females (2.68) which is 0.13.

PSPP 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 is 1.300 and the significance value is .194.  This means that we can’t reject the null hypotheses.  In others words, we have no basis for asserting that the population slope is significantly different from zero.  The Pearson Correlation Coefficient Squared (Coefficient of Determination) 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 PSPP to get the frequency distribution for d6\_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:

* d6\_race\_white – equals 1 if the person is white and 0 if the person is non-white, and
* d6\_race\_nonwhite – equals 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 d6\_race\_nonwhite out of the regression equation.  Use COMPUTE to create the d6\_race\_white dummy variable following the procedures we used in Part 2.

Now run the regression analysis with tv1\_tvhours as your dependent variable and d6\_race\_white as your independent variable.

* Write out the regression equation.
* What does the unstandardized multiple regression coefficient 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 you ran a regression analysis with tv1\_tvhours as your dependent variable and d1\_age, d24\_paeduc, and d4\_educ as your independent variables.  This time we’re going to add d5\_sex\_males into the analysis.  Use PSPP to carry out the regression analysis for this model.

The regression equation for predicting tv1\_tvhours is 3.37 + .02 (d1\_age) - .05 (d24\_paeduc) - .09 (d4\_educ) + .22 (d5\_sex\_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 (Coefficient of Determination) tells us that together the independent variables explain or account for 10% of the variation in television viewing.  The Adjusted Squared Correlation Coefficient adjusts for the number of independent variables and is slightly lower (9%).  The F test in the analysis of variance table is also statistically significant meaning that we can reject the null hypothesis that our independent variables explain none of the variation in the dependent variable.  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 d5\_sex\_males into the analysis this time add the dummy variable you created in Part 3 (d6\_race\_whites).  This means you will have four independent variables -- d1\_age, d24\_paeduc, d4\_educ, and d6\_race\_whites.

* 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?