**STAT15S\_pspp: Exercise Using PSPP to Explore Multiple 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\_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 multiple linear regression and also uses FREQUENCIES, BIVARIATE CORRELATION, and SELECT CASES.  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_STAT15S_pspp.docx) (MS Word; docx format).
* [**PSPP syntax file**](http://ssric.org/files/PSPP_Syntax_for_STAT15S_pspp.sps) (.sps format)
* [**This page**](http://ssric.org/files/STAT15S_pspp.docx) (MS Word; docx format).

 **Goals of Exercise**

The goal of this exercise is to introduce multiple linear regression.  The exercise also gives you practice using LINEAR REGRESSION, FREQUENCIES, and SELECT CASES in PSPP.

**Part I – Linear Regression with Multiple Independent 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 the 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 expands the analysis to include multiple independent variables.  In the first part of this exercise we’re going to focus on two independent variables.  Then we’re going to add a third independent variable into the analysis.  An important assumption is that “the dependent variable is seen as a linear function of more than one independent variable.”  (Colin Lewis-Beck and Michael Lewis-Beck, *Applied Regression – An Introduction*, Sage Publications, 2015, p. 55)

In the last exercise we used tv1\_tvhours as our dependent variable which refers to the number of hours that the respondent watches television per day.  In other words, we want to understand why some people watch more television than others.  We found that age was positively related to television viewing and father’s education was negatively related.  Older respondents tended to watch more television and respondents whose fathers had more education tended to watch less television.

Let’s start by using FREQUENCIES to get the frequency distribution for tv1\_tvhours.  In the previous exercise 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.

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.

In bivariate linear regression we have one independent and one dependent variable.  So we are trying to find the straight line that best fits the data points in a two-dimensional space.  With two independent and one dependent variable we have a three-dimensional space.  So now we’re trying to find the plane that best fits the data points in this three-dimensional space.

With two independent variables our regression equation for predicting Y is a + b1X1 + b2X2 where a is the constant, b1 and b2 are the unstandardized multiple regression coefficients, and X1 and X2 are the independent variables.  As with bivariate linear regression we want to minimize error where error is the difference between the observed values and the predicted values based on our regression equation.  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.**[[1]](http://ssric.org/node/579/edit%22%20%5Cl%20%22_edn1%22%20%5Co%20%22)**

**Part II – Getting the Regression Coefficients**

The regression equation will contain the values of a, b1, and b2 that minimize the sum of the squared errors.  There are formulas for computing these coefficients but usually we leave it to PSPP to carry out the calculations.

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.  In the previous exercise we ran two bivariate linear regressions – one with tv1\_tvhours and d1\_age and a second with tv1\_tvhours and d24\_paeduc.  In this exercise we’re going to use both independent variables simultaneously.  Enter both d1\_age and d24\_paeduc as your independent variables 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 Multiple Correlation Coefficient (R) and the Squared Multiple Correlation (R2) which is usually referred to as the Coefficient of Determination.  How does this differ from the Pearson Correlation Coefficient (r) and r squared.  The Multiple R squared tells us that d1\_age and d24\_paeduc together explain or account for 8% of the total variation in the number of hours per day that respondents watch television.   In the previous exercise we saw that d1\_age by itself explained 4% of the variation in tv1\_tvhours and that d24\_paeduc by itself explained 5%.  Why can’t we just add 4% and 5% and say that 9% of the total variation is explained by these two variables together.  We see from the PSPP output that this is not true.  In fact, 8% of the total variation is explained by these two variables together.  Why is that?  It’s because the variation explained by these two independent variables overlap and because of this overlap they only account for 8% of the total variation in the dependent variable.
* The second box also gives us the Adjusted R squared which “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 Publishers, p. 293)  The standard error is an estimate of the amount of sampling error for this statistic.
* The third box is the analysis of variance table that tests the null hypothesis that the Squared Multiple R **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 and father’s education together explain more than 0 percent of the variation.
* Recall that the equation for predicting Y is a + b1X1 + b2X2.  The fourth box gives you more information.
	+ The constant (a) is 2.70.
	+ The unstandardized multiple regression coefficient (b1) for d1\_age is .02.  This means that an increase of one unit in d1\_age results in an average increase of .02 units in tv1\_tvhours after statistically adjusting for d24\_paeduc.  Or, to put this in more easily understood terms, an increase of one year in the respondent’s age results in an average increase of .02 hours of television viewing after statistically adjusting for father’s education.
	+ The unstandardized multiple regression coefficient (b2) for d24\_paeduc is -.08.  This means that an increase of one unit in d24\_paeduc results in an average decrease of -.08 units in tv1\_tvhours after statistically adjusting for d1\_age.  Or, to put it another way, an increase of one year in the father’s education results in an average decrease of .08 hours of television viewing after statistically adjusting for the respondent’s age.
	+ So what does it mean to statistically adjust for something?  Suffice it to say that it means that b1 tells us the effect of X1 on the dependent variable after taking into account the other independent variables (i.e., in this case X2).   The other regression coefficient, b2, would be similarly interpreted.
	+ The standard error of these coefficients which is an estimate of the amount of sampling error.
	+ The standardized multiple regression coefficients (often referred to as Beta).  You can’t compare the unstandardized multiple regression coefficients (b1 and b2) because they have different units of measurement.  One year of age is not the same thing as one year of education.  The standardized multiple regression coefficients (Beta) are directly comparable.  You can see that the Beta for d24\_paeduc is -.18 and for d1\_age is .17 which means that father’s education is slightly more important in predicting hours of television viewing than is age.
	+ The t test which tests the null hypotheses that the **population** constant and **population** multiple regression coefficients are equal to 0.
	+ The significance value for each test.  As you can see, in this example we reject all three null hypotheses.  However, we’re usually only interested in the t test for the population multiple regression coefficients.

So our multiple regression equation for predicting Y is 2.700 + .02X1 - .08X2.  Thus for a person that is 20 years old and whose father completed 12 years of school,  the predicted number of hours that he or she watches television 2.700 +   (.02) (20) - .08 (12) or 2.700 + 0.4 – 0.96 or 2.14 hours.

It’s **important** to keep in mind that everything we have done assumes that our dependent variable is a “linear function of more than one independent variable.”  (Colin Lewis-Beck and Michael Lewis-Beck, *Applied Regression – An Introduction*, Sage Publications, 2015, p. 55)

**Part III – It’s Your Turn Now**

Use the same dependent variable, tv1\_tvhours, but this time add d4\_educ to your list of independent variables.  Now you will have three independent variables – d1\_age, d24\_paeduc, and d4\_educ.  The variable d4\_educ is the years of school completed by the respondent.  Use PSPP to get the regression equation.

* Write out the regression equation.
* What do the unstandardized multiple regression coefficients (b1, b2, and b3) 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?

**Part IV – Do we have a problem?**

Multicollinearity occurs when the independent variables are highly correlated with each other.  If one of your independent variables is a perfect linear function of the other independent variables, then you would not be able to determine the regression coefficients.  But this is not typical.  What is more likely is that some of the independent variables might explain a large portion of the variation in another independent variable.  For example, in Part 3 what if father’s education and age explained a very large portion of the variation in respondent’s education?  Then you would have high multicollinearity.  The problem that multicollinearity creates is that it tends to make your regression coefficients less reliable.  The standard errors of the regression coefficients increase which makes it harder to reject the null hypothesis in your t tests.

There are several ways to determine if multicollinearity is a problem in your analysis.  You can start by looking at the Pearson Correlation matrix for your independent variables.  Use BIVARIATE CORRELATION in PSPP to get the bivariate Pearson Correlation matrix for the three independent variables you used in Part 3 – d1\_age, d24\_paeduc, and d4\_educ.  If any of these correlations is really high, then you would have a problem but in this example it doesn’t appear that is the case.  There are other ways to detect multicollinearity but for this exercise we’ll stop here.

**[[1]](http://ssric.org/node/579/edit%22%20%5Cl%20%22_ednref1%22%20%5Co%20%22)** When you square a value the result is always a positive number.