SOC 385L – 4th Lab Session
Lab Assignment 2

Simple Regression and Basic Multiple Regression:

In the last lab session, you practiced basic coding to transform variables and began to explore some initial "descriptive statistics" including frequencies and percentages for variables by various groups.

In this lab session, you'll jump right into some simple bivariate (two variable) regressions and basic multivariate (multiple variable) regressions. But first, you will get some practice exploring the data further in looking for some measures of central tendency (mean, etc.) and some measures of the distribution (skewness and kurtosis). You will also continue the recode-style transformations you did last time by adding some functional form (log) transformations.

The data this time are taken from the 2010 GSS (General Social Survey).

> *"The GSS contains a standard 'core' of demographic, behavioral, and attitudinal questions, plus topics of special interest. Many of the core questions have remained unchanged since 1972 to facilitate time-trend studies as well as replication of earlier findings. The GSS takes the pulse of America, and is a unique and valuable resources. It has tracked the opinions of Americans over the last four decades."* (http://www3.norc.org/GSS+Website/)

In this exercise you will perform an OLS analysis of attitudes toward abortion. The dependent variable is a count of whether a respondent believes that a woman who wants an abortion for a variety of reasons (e.g., doesn't want more children, when her health is threatened by the pregnancy) should be allowed to get one.

Part One: Recoding, Exploring, and Simple Regression

1. Let's begin by finding the abortion variables. Lookfor "ab" and scan the reduced list of variables in Stata's output for those related to abortion. How many variables relate to abortion? Do any of these seem different?

2. Let's get more info to see whether or not this variable belongs. Go to the GSS website above and let's look at the codebook, or even easier "browse GSS Variables" and look at some of the ones in our dataset. Should we include this other variable with the others?

3. Find and recode the seven abortion attitudes variables (hint: they all begin with "ab") so that 1=yes and 0=no.

4. Add them all. Do it two different ways-- using egen to create (abcount1) and using gen to create (abcount2) – what is the difference? If this is a problem, can you find something in the egen help file that will help fix this? Try creating a third variable, abcount3 using this option and compare this one. Which one do we want to use? From now on, when we refer to "abcount" we will use the one you choose here.

5. Examine the distribution of the count variable you chose above and note what you find.

6. Create a scatterplot with a fitted line for the abortion count variable and education. Comment on the relationship between these two variables.

7. Regress this count variable on education. Interpret the results.

<u>Part Two: Multiple Regression and Functional Forms</u>

Next, let's move on to multiple regression.

8. All variables are in the 2010 General Social Survey (codebook and data in the lab assignment and data set folder on Blackboard). Code religion, gender, and race as dummy variables so that Catholic=1 and 0 for other religions; female=1 and 0 for males; and black=1,0; otherrace=1,0 and whites as the omitted category.

9. Create a new variable "income_ind" from rincom06 (respondent's income) and recode these values to their midpoints. Add $25,000 to the last category to set its value. Use this code to also create a new variable "income_fam" from income06 (family income) and code it the same way.

10. Run a multiple regression to determine the influence of education, individual income, religion, gender, and race on this attitude. Interpret this regression output briefly. How does our R-squared and the coefficient of education compare to the simple regression part one? If you need to run the simple regression again for comparison, go ahead.

11. Now run the same regression above again, substituting family income for individual income. Does it make a difference to consider family income instead of respondent's income? Look at the coefficient for female. What is going on here?

12. Look at the distribution of the explanatory variables in terms of normality, skewness, kurtosis. Do you want to transform any of them? (a functional form transformation). Experiment with this.

13. Log your newly created individual income variable, to create a variable called lnincome_ind, and run a multiple regression of your abcount variable on education and lnincome_ind.

14. Repeat this to create a logged variable for family income, lnincome_fam. Run a regression of your abcount variable with both lnincome_fam (log family income) and education as independent variables. What is going on here?

Finally, think about what else you might include in this model.

15. Run a regression of abcount one more time with education, individual income (logged), religion (catholic), gender and race (black and otherrace). Note the R-squared score.

16. Are there any other variables that you would include in the model as explanatory variables? Skim through the GSS dataset and select at least one other variable. Transform it as necessary. Check its distribution. Perform a functional transformation if needed. And finally, add this as an explanatory variable in the model above.

17. Interpret your results and comment on the effect of your chosen variable on abcount.

Some Key Commands:

Un-ivariate (one variable) commands:

tab variable	Creates a table of frequencies (the ", row" or ", col" option gives percentages)
hist variable	Create a histogram to view the distribution
stem variable	Creates a stem-and-leaf plot similar to the histogram
sum variable, d	Provides a set of summary statistics including mean and median. (The ", d" option gives details including skewness (ideally zero), and kurtosis (ideally 3)).
graph box variable	Creates a box plot where you can see the spread of a variable and check for outliers.
pnorm variable	Creates a scatterplot with lines comparing the variable to the normal distribution, to let us know if we are close.

Bi/multi-variate (two variable) commands:

tab var1 var2	Creates a two-way table of frequencies. Remember the row and col options for percentages.
graph twoway (scatter var1 var2) (lfit var1 var2)	Creates a two-way scatterplot of two variables. Lfit gives the best fitted line based on the given points of data.
regress var1 var2	Performs a simple regression of the two variables, providing output on the estimates of the coefficients for the slope, intercept and error among other things.

Other commands:

preserve	Takes a snapshot of dataset as it is.
restore	Reverts to the last version of the dataset that was preserved.
gen lnvar = ln(var)	Creates a log of a variable
lookfor text	Looks for variables with the given text in name, label, or values.