Stata Handout #3

The purpose of this lab is twofold. First, I want to give you some basic familiarity with regression analysis in Stata. Second, I want to give you the opportunity to generate and interpret some regression models using a real social science dataset. I should note that I am writing these descriptions as a Mac user with a Mac version of Stata. Some things may be slightly different on a Windows machine, but we will cross that bridge when and if we get to it.

Section 1a: Setting Up Some Data
Open Stata 14 (not the 64-bit version). From the command line, type the following:

```
cd N:/soc_30903/
```

If you haven't done so already, download the 2014 GSS dataset located in the "Handout Data" sub-folder under the "Stata Handouts" tab in Sakai. Load it up with the following command:

```
use data/GSS2014.dta, clear
```

You should see the GSS variables populate the "Variables" pane in the main Stata window. Once you have done that, start up your do-file for this session. Recall that you can either use the "Do-File" icon on the top ribbon, or you can simply type `-doedit-` from the command line (without the dashes).

You will turn in this do-file at the end of the lab, so go ahead and put your name on it. You should use this do-file to write and execute the remainder of your commands for the day.

Section Ib: Bivariate ("Simple") Regression
The basic syntax for a bivariate regression models is as follows:

```
reg DV IV
```

where `-DV-` is the interval-ratio dependent variable and `-IV-` is the interval-ratio or dichotomous independent variable.

Let's try an example. Let's say we want to know if, in the adult U.S. population, the age at which a person has their first child varies depending on the number of years of schooling they have completed.

> **Thought experiment #1a:** What are the dependent and independent variables? Have we made the necessary assumptions and met the proper requirements? What is the null hypothesis? What is the alternative hypothesis? What α-level do we want to use?

Now let's run the test:

```
reg agekdbrn educ
```

> **Thought experiment #1b:** What does our slope tell us, and what can we say about its statistical significance? What is the constant telling us? What is the F-statistic for the model telling us? What about R^2 and the root MSE? What is the conclusion of our test?

Now let’s try an example with a dichotomous independent variable. Let’s say we want to know if black Americans are more likely to have their first child at a younger age than non-black Americans.

> **Thought experiment #2a:** What are the dependent and independent variables? Have we made the necessary assumptions and met the proper requirements? What is the null hypothesis? What is the alternative hypothesis? What α-level do we want to use?

The “race” variable in the dataset consists of three categories (white, black, and other), so let’s create a new dichotomous variable where 0 = non-black and 1 = black:

```
gen black = .
replace black=0 if race==1 | race==3
replace black=1 if race==2
label def black 0 “non-black” 1 “black”
label val black black
```

The first line creates a new variable called “black” that is equal to nothing (well, technically a “.”—wwhich is used to indicate missing data). The second and third lines populate the new variable conditional on values of the original “race variable; specifically, we set “black” equal to 0 if “race” is equal to 1 or 3(the white and other categories, respectively) and equal to 1 if “race” is equal to 1 (the black category). The pipe symbol (|) means “or” in Stata (i.e., we set a case’s new variable value to 0 if the case’s original variable value was 1 or 3).

The fourth line gives the new variable a variable label. The fifth line defines a new value label where 0 is equal to “non-black” and 1 is equal to “black.” Value labels are what allow us to see qualitative names for the categories rather than the numerals that they represent. Finally, the sixth line “attaches” the value label to the new variable (we type “black black” because both of them have the same name).

Finally, we can run our bivariate regression. The Stata syntax looks just like the one from before:

```
reg agekdbrn black
```

Run the test.

> **Thought experiment #2b:** What does our slope tell us, and what can we say about its statistical significance? What is the constant telling us? What is the F-statistic for the model telling us? What about R^2 and the root MSE? What is the conclusion of our test?

Section Ic: Multiple Regression

The Stata syntax for multiple regression looks just like the syntax for bivariate regression. The only difference, as you might have guessed, is that we simply add one or more variables to the model. The basic syntax looks like this:

```
reg DV IV CV1 CV2 CV3
```

where, as before, `-DV-` is the interval-ratio dependent variable and `-IV-` is the independent variable (the latter of which can be either interval-ratio, dichotomous, or multinomial, at this point). Now we have added `-CV-`, which indicates our control variables.

Let's say we want to know if the effect of educational attainment on the age at which a person's first child is born holds after controlling for age and sex.

> **Thought experiment #4a:** What are the dependent and independent variables? Have we made the necessary assumptions and met the proper requirements? What is the null hypothesis? What is the alternative hypothesis? What α-level do we want to use?

Let's run the test:

```
reg agekdbrn educ age sex
```

> **Thought experiment #4b:** What do our slopes tell us, and what can we say about their statistical significance levels? What is the constant telling us? What is the F-statistic for the model telling us? What about R^2 and the root MSE? What is the conclusion of our test?

If we want to include a multinomial independent/control variable, we need dummy variables for each category with one left over as the reference category (see the multiple regression slides for a refresher on the concept of dummy variables).

Back in the day you would have to do this manually. For instance, if I wanted to include race-ethnicity as a control variable, it used to be the case that we would have to do something like the following:

```
recode race (1=1)(nonm=0), gen(white)
recode race (2=1)(nonm=0), gen(black)
recode race (3=1)(nonm=0), gen(other)
```

This is all well and good, but I prefer to not create a series of dummy variables every time I want to include a categorical independent/control variable in my models. Luckily, the fine people over at StataCorp share my laziness and have provided us with what they refer to as "factor variable notation." (A factor variable, by the way, is just another term for categorical variable—the logic being that there are multiple "factors" or "groups" that constitute the variable.)

Factor variable notation is incredibly easy to implement. For a categorical predictor, all you have to do is specify `i.` before the variable name. Stata then recognizes that you want a separate intercept for each level of the variable and it automatically sets the smallest value level as your reference category (or baseline, or omitted category). Let's give it a try:

```
reg agekdbrn educ age sex i.race
```

White Americans are our reference category, since they have the smallest value level (1).

Thought experiment #5: What do the slopes for our race-ethnicity variable tell us?

We can also change our reference category using `ibX.` as our prefix, where `X` is the value level for the category we want to use. So if we want black Americans to serve as our reference category:

```
reg agekdbrn educ age sex ib2.race
```

Or perhaps we want to use the category with the most observations—not an uncommon strategy in empirical research:

```
reg agekdbrn educ age sex ib(freq).race
```

Now we have all of the tools necessary to include almost any variable that we want as either independent or control variables. Let's add social class, religious fundamentalism, and annual earnings as control variables:

```
reg agekdbrn educ age sex i.race ib(freq).class ib2.fund conrinc
```

> **Thought experiment #6:** What do our slopes tell us, and what can we say about their statistical significance levels? What is the constant telling us? What is the F-statistic for the model telling us? What about R^2 and the root MSE? What is the conclusion of our test?

Section Id: Predicted Values

Stata also makes it easy to generate predicted values. To do this, we use the post-estimation `-margins-` command. We call it a "post-estimation" command because we use it after we have run our regression model. If we want to find the predictions at observed values (i.e., the mean of all of the predicted values we get when we use the observed values for the independent/control variables for each case), all we have to do is this:

```
reg agekdbrn educ age sex i.race ib(freq).class ib2.fund conrinc
margins
```

The predicted value is located under the `Margins` column.

We can also tell Stata to generate predicted values using the mean of each independent variable rather than the observed values:

```
margins, atmeans
```

As you can see, the predicted values do not change much between the observed values option and the means option.[1]

[1] Mathematically, the predictions at means are found with the following equation:

$$\hat{y} = a + \beta_1\bar{x}_1 + \beta_2\bar{x}_2 + \cdots + \beta_k\bar{x}_k$$

The real appeal of the `-margins-` command is that it allows us to specify values in the predictions. For instance, let's say I want to the predicted age at which someone had their first child for a person who has a high school diploma, is 35, female, white, middle class, a moderate religious fundamentalist, and makes an average annual income:

```
margins, at(educ=12 age=35 sex=2 race=1 class=3 fund=2) atmeans
```

The `-atmeans-` option tells Stata to set any unspecified independent/control variables to their mean for calculating the prediction. In this case, the only variable for which I didn't specify a value was income; therefore, Stata set income to its mean ($36,421.38).

We can also use the `-margins-` command to quickly produce a whole series of predicted values. For instance, let's say I wanted to take the same hypothetical person from above and alter only their education level—say, between a 10th grade education and up to 20 years of education (roughly a doctoral degree), going up in 2 year increments. In this way, I am generating predictions that isolate the effect of educational attainment on the dependent variable by allowing only the education variable to change; all other variables are held constant:

```
margins, at(educ=(12(2)20) age=35 sex=2 race=1 class=3 fund=2) ///
     atmeans
```

Including the three forward slashes is a way of telling Stata that I am "breaking" my code across two lines. Otherwise Stata would not recognize the `-atmeans-` portion of the syntax as part of the desired command.

> **Thought experiment #7:** What do the predicted values tell us about the effect of educational attainment on the age at which a person is likely to have their first child?

Section Ie: Creating a Regression Table

Lastly, I want to show you how to use Stata to create the regression tables that you will often see in published social science research. You can always make these manually, but Stata really can save you a lot of time and effort—especially when you start putting together large and complicated models.

You can make these tables using the user-written `-estout-` package. Since it is user-written, you need to install it. Luckily, installation is easy enough:

```
ssc install esout
```

where k is the number of independent/control variables. The predictions at observed values are found with the following:

$$\hat{y} = \frac{\sum_{i=1}^{n} a + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}{n}$$

In any case, the results will often be quite similar.

You can use the -esttab- function that comes with the package to make your tables. First, after each regression, you'll want to "store" the estimates so that you can create the table after running consecutive models:

```
reg agekdbrn educ
est store m1
reg agekdbrn educ age sex i.race
est store m2
reg agekdbrn educ age sex i.race ib(freq).class ib2.fund conrinc
est store m3
```

Now you are ready to create your table. Just do the following:

```
esttab m1 m2 m3, se
```

We specify the -se- option so that Stata knows to include the standard errors of the slopes in the parentheses instead of the t-statistics.

Section II: Go Forth and Explore

Now try running some regression analyses on your own. Use the -codebook-, -tab-, and -summarize- commands to find some variables that interest you. Go through the Thought Experiment exercises again with these new variables (and their interrelationships) in mind. Type your codes into a do-file rather than from the command line in the main window.

Save your dataset and do-file when you are finished (you have to save them separately). You can do this directly from your do-file:

```
save "data/GSS2014.dta", replace
save "do_files/stata_handout_2.do", replace
```

Once you are done, execute the following to close you Stata session and exit the program.

```
clear
exit
```