**Week 1 Lab Handout- STATA & PSID PA 5032 – Applied Regression** January 22nd, 2021

Wenchen Wang

**PART A: INTRODUCTION ~ 5min PART B: APPSTOGO ~ 15min**

**PART C: PSID- Panel Study for Income Dynamics ~35min PART D: STATA REVIEW ~ 20min**

# PART A: INTRODUCTION

Welcome to Applied Regression Lab! Each Friday we will be covering topics that will aid you in your independent quantitative project. We will lecture on new commands, expand upon items discussed in class, and give you all time to practice new skills.

Labs are meant to be an interactive experience, so please ask questions if you feel confused or lost. It is expected that you attend all labs. If you must miss due to illness, a family emergency, or other conflict, please email us ahead of time so that we can make sure to get you up to speed on new material covered.

# PART B: APPSTOGO

If you have not already, make sure you know how to access AppsToGo on your personal computers. We will be using AppsToGo during lab to access STATA, and it allows you to use STATA outside the University.

This link walks you through the three steps necessary to access AppstoGo: <https://it.umn.edu/self-help-guide/appstogo-use-umn-apps-your-personal>. Refer to the section “Access AppsToGo”.

## PART C: PSID- Panel Study for Income Dynamics (PSID) and General Social Survey (GSS)

*PSID*

“The [Panel Study for Income Dynamics](https://psidonline.isr.umich.edu/) (PSID) is the longest running longitudinal household survey in the world. The study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. Information on these individuals and their descendants has been collected continuously, including data covering employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy,

education, and numerous other topics. The PSID is directed by faculty at the University of Michigan, and the data are available on this website without cost to researchers and analysts.”

*GSS*

“The [General Social Survey](https://gss.norc.org/) (GSS) is a nationally representative survey of adults in the United States conducted since 1972. The GSS collects data on contemporary American society in order to monitor and explain trends in opinions, attitudes and behaviors. The GSS has adopted questions from earlier surveys which allows researchers to conduct comparisons for up to 80 years.

The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, intergroup tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.”

There are two videos on the Canvas page under “Media Gallery” that were created by PA5031 TA’s last semester. Feel free to reference them for a step-by-step tutorial on downloading and opening data from each. If you are still having trouble, please contact one of the TAs.

# PART D: STATA REVIEW

**Directories and log files**

***DO files*** show your code and your comments

***LOG files*** show your code and your comments, and also the output generated from your code.

\*\*\*You will turn in a *text format* log file for every assignment in this course!\*\*\*

*Creating a log file:*

In your .do file, write the code for opening a log file after your header:

*cap log close* This ensures that any previously open log files are closed

*cd [folder address]* This sets the directory for the log file (where it will be saved). Use either your O: drive or your H: drive. You might want to make a folder specifically for PA 5044.

*log using lab1\_statareview, text replace*

This starts the log. The most important part of this command is “*text replace*”

*, text* : ensures the log is saved as a .log file rather than a

.smcl file.

*, replace* : allows you to re-run your .do file and generate an updated log file as you make edits to your code

You may also want to use *, text append* if you are adding on to a previously existing log!

Example Code: **clear**

**set more off cap log close**

**cd "O:\boesc011\Regression\TA" use "PSID\_Practice\_Clean.dta"**

**log using Boesch\_Lab1\_log, text replace**

**log close**

**Commenting**

Commenting in your do-file can be useful to remind yourself why you did something in Stata or to point out important information. There are a couple of ways to comment in your do-file.

\*For one line of text, you can use an asterisk at the beginning of a line.

/\* If your comment will be more than one line long, use a forward slash (/) and then an asterisk.

You can even create a new line and the comment will continue until you close the comment with another asterisk and slash. \*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* You can put in lots of stars to make things stand out, just start and end your comment with a slash. \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**Cleaning Data**

A lot of times you will encounter “messy” data that has missing observations, strange variable names, or less than useful coding. Luckily, STATA makes it super easy to take care of these issues so that we have clean usable data.

With PSID Data, the variables will have a name that does not relate to what it is actually showing, but the labels will tell us what the variables are. Our first step can be to “rename” the variables.

Other times, we may want to generate a dummy variable that subsets another variable in our data set. For example, we may want to divide total family income between less than $100,000 and greater than or equal to $100,000. The first step would be to “generate” a new variable that is equal to our old variable. Then, we would recode our new total family income variable so that households with a family income of less than $100,000 are assigned a value of 1, and families with incomes of

$100,000 or greater are assigned a value of 0.

*rename [old variable name] [new variable name]*

This gives a variable a new name. Be sure to list the old variable first, followed by the new variable.

## ex. rename ER71426 family\_income

*generate [new variable name] = [old variable name]*

*recode [variable name] [values to be recoded]*

Creates a new variable with the same data as another variable that may then be manipulated

## ex. generate family\_incomeunder100k = family\_income

Redefines different possible values (responses) within a variable to enable data to be categorized in new ways

## ex. recode family\_incomeunder100k(0/99999=1)

**Summarizing the data**

Now we are going to do some review of basic Stata commands to summarize your data.

If we want to see what is in our data (number of observations, number and descriptions of variables):

*describe*

For a summary of all of our variables:

*summarize*

Let’s take a look at the variable we just created, *family\_incomeunder100k* (a dummy variable noting whether a household’s total family income is less than $100,000 or not):

*sum family\_incomeunder100k tab family\_incomeunder100k*

Now, let’s get an understanding of both family income and the age of the reference person. Create a crosstab looking at family\_incomeunder100kincome and age\_50 (1=reference person is 50 years or older)

*tab family\_incomeunder100k age\_50*

Look at family income under a certain condition

*tab family\_incomeunder100k if age > 40*

Examine a "row, column, cell" table with family\_incomeunder100k and age\_50

*tab family\_incomeunder100k age\_50, row col cell*

For descriptives for age of reference person

*sum reference\_age, d*

Total number of observations where age\_50=0

*count if age\_50==0 count if age\_50!=0*

Break down the summary for age\_50 into two categories: those whose family income is less than

$100,000, and those whose family income is greater than or equal to $100,000

*tab family\_incomeunder100k, sum(age\_50)*

The help command can tell you what a command does or how to obtain a particular type of output

*help*

*help scatterplot help twoway scatter*

-