Basic Stata Tutorial

By Brandon Heck

Downloading Stata

To obtain Stata, select your country of residence and click "Go." Then, assuming you are a student, click "New Educational" then click "Students." The capacity of the different versions of Stata increase with price. Unless you know for a fact you are only going to work with very small data sets, I would highly recommend Stata/IC or better. Once you buy it, you should receive an electronic key and be able to install Stata.

Do-Files

Once you download and open Stata, click the "New Do-file Editor" button just below the standard menu. A do-file is the script for Stata, where we write all of the code that we want to execute. We can run all of the code that we want to execute directly from the do-file.

Example Analysis of a Data Set

For this example, we are going to use the April 2014 wave of the Current Population Survey, which is a cross-sectional survey of households in the United States. I have already taken a .7% sample of the original data set so that it can be used with small stata.

Importing, Exporting, and Using Data

Import Data

First, we need to import the data from a comma-separated values file, which we do using the command "insheet." The "clear" option is used to clear the current data set, since Stata can load into memory only one data set at a time.

```
insheet using "C:\Stata Workshop\cps_sample.csv", clear
```

Save data set in Stata format

Save the data set in Stata format. The replace option causes the current data set to override a data set of the same name (this is mainly for when the code is run multiple times, in order to avoid errors). Note that the extension for Stata files is .dta.

```
save "C:\Stata Workshop\cps_sample.dta", replace
```

Use the Stata data set

Once we save the data in Stata format, or if we obtained a Stata data set, we can simply use the data set to load it into memory.

```
use "C:\Stata Workshop\cps_sample.dta", clear
```

Export the data to Excel

This step is not needed for this analysis, but we are able to export Stata files into Excel or many other statistical packages. This is useful if we perform data cleaning in Stata and then want to export to another program. The code below exports a Stata file to Excel.

```
export excel "C:\Stata Workshop\cps.xlsx", replace
```

Data Cleaning

In order to make the data set more user friendly and eliminate odd values to allow the data to be used for proper analysis, we perform several data cleaning steps.

Keeping Variables

The first step of data cleaning is to identify the variables we are going to use and drop all others. We perform this with the "keep" statement. We may also use "drop" to drop certain variables. In this case, we are keeping sex, age, family income, and education.

```
keep pesex prtage hefaminc peeduca
```

Renaming Variables

Since the variables have somewhat odd names, we want to rename them to have names that make more sense to us. If we want to rename multiple variables, we use parantheses where the nth variable in the first set of parantheses is renamed to the nth variable in the second set of parantheses.

```
rename (pesex prtage hefaminc peeduca) (sex age famincome educ)
```

Recode missing values for all 4 variables

In the raw data, the Census codes missing values as "-1." In Stata, missing value numbers are denoted as "." So, we need to recode the -1's as dots for each variable. We do this using the "replace" command which replaces observations. In Stata, unlike many other packages, the "if" qualifier comes after the command. Note that a single "=" is for variable assignment whereas "==" is used as a comparison operator (that is, a test for equality). For example, the first statement below will replace all observations where the variable "sex" is equal to "-1" with a missing value.

```
replace sex = . if sex == -1
```

replace age = . if age == -1

replace famincome = . if famincome == -1

replace educ = . if educ == -1

Create Binary variables out of the education variable

The education variable is categorized in the following way:

31 - Less than 1st grade

32 - 1st, 2nd, 3rd or 4th grade

33 - 5th or 6th grade

34 - 7th or 8th grade

35 - 9th grade

36 - 10th grade

37 - 11th grade

38 - 12th grade no diploma

39 - High School Grad Diploma

40 - Some College But No Degree

41 - Associate Degree-Occupational/Vocation

42 - Associate Degree

43 - Bachelor's Degree

44 - Master's Degree

45 - Professional School Degree

46 - Doctorate Degree

In order to use this variable in analysis, it is likely much more useful and informative to create several binary variables of interest out of this variable. We will create a variable for high school graduate and college graduate. The variable for high school graduate will simply take on a value of one if "educ" is greater than or equal to 39 and 0 otherwise. The college graduate variable will take on a value of one if "educ" is greater than or equal to 43 and 0 otherwise. A very important note with the following code is that Stata treats missing values as essentially every number, so "educ >= 39" will also include missing values, which we do not want, so we must specify that "educ != ."

generate hsgrad = 1 if educ >= 39 & educ != .

replace hsgrad = 0 if educ < 39 & educ != .

generate collegegrad = 1 if educ >= 43 & educ != .

replace collegegrad = 0 if educ < 43 & educ != .

Income Categories for Reference

We are not going to convert the family income variable (although you probably would want to for research), but for reference, the variable is specified in broad categories in the following fashion:

1 - Less than 5,000

2 - 5,000 to 7,499

3 - 7,500 to 9,999

4 - 10,000 to 12,499

5 - 12,500 to 14,999

6 - 15,000 to 19,999

7 - 20,000 to 24,999

8 - 25,000 to 29,999

9 - 30,000 to 34,999

10 - 35,000 to 39,999

11 - 40,000 to 49,999

12 - 50,000 to 59,999

13 - 60,000 to 74,999

14 - 75,000 to 99,999

15 - 100,000 to 149,999

16 - 150,000+

Summary Statistics

Describe the data

"describe" produces a basic summary of the data set currently loaded into memory, including the number of observations, variables, and a list of the storage types and display formats of each variable.

describe

Provide basic summary statistics of the data

"summarize" provides the number of observations, mean, standard deviation, min, and max of the specified variables. If no variables are specified, these summary statistics are provided for the entire data set. Adding the option detail will provide even more summary statistics such as percentile values, skewness and kurtosis.

summarize age

summarize age, detail

summarize

Tabulate

"tabulate" will give frequency counts of each value of a given variable. If two variables are specified, "tabulate" performs a cross-tabulation, complete with subtotals for each value of both variables. For example, in the cross-tabulation below, we can see that 11,730 individuals graduated high school but not college (this will vary slightly depending on the random sample taken).

tabulate hsgrad

tabulate hsgrad collegegrad

Labeling Variables

In order to label variables for convenience, we can simply use the option "label variable." This label then shows up next to the variable on the right-hand side of Stata. In this case, we will label the variables with more meaningful descriptions.

label variable age "Age"

label variable sex "Sex (1 - Male 2 - Female)"

label variable educ "Education"

label variable hsgrad "High School Graduate"

label variable collegegrad "College Graduate"

label variable famincome "Family Income"

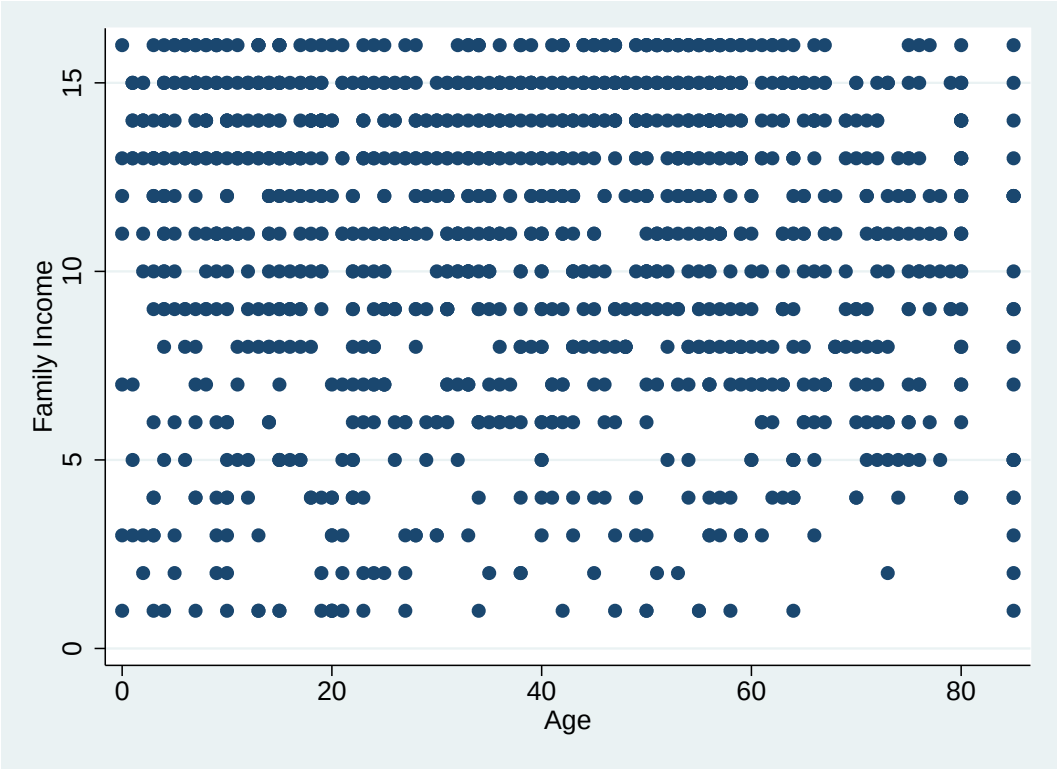
Plotting Data, Collapsing

Basic Scatter Plot and Exporting

In order to produce a two-way plot, we use the command "graph twoway." There are many different two-way plots as you can see with "help graph twoway", but we will just use "scatter". To export the figure as seen below, we can simply use the "graph export" command, and export to several graphics extensions, including .pdf.

graph twoway (scatter famincome age)

graph export "C:\Stata Workshop\scatter.pdf", replace



This graph is essentially meaningless, so we need to aggregate the data by age to create a meaningful plot.

Collapsing Data

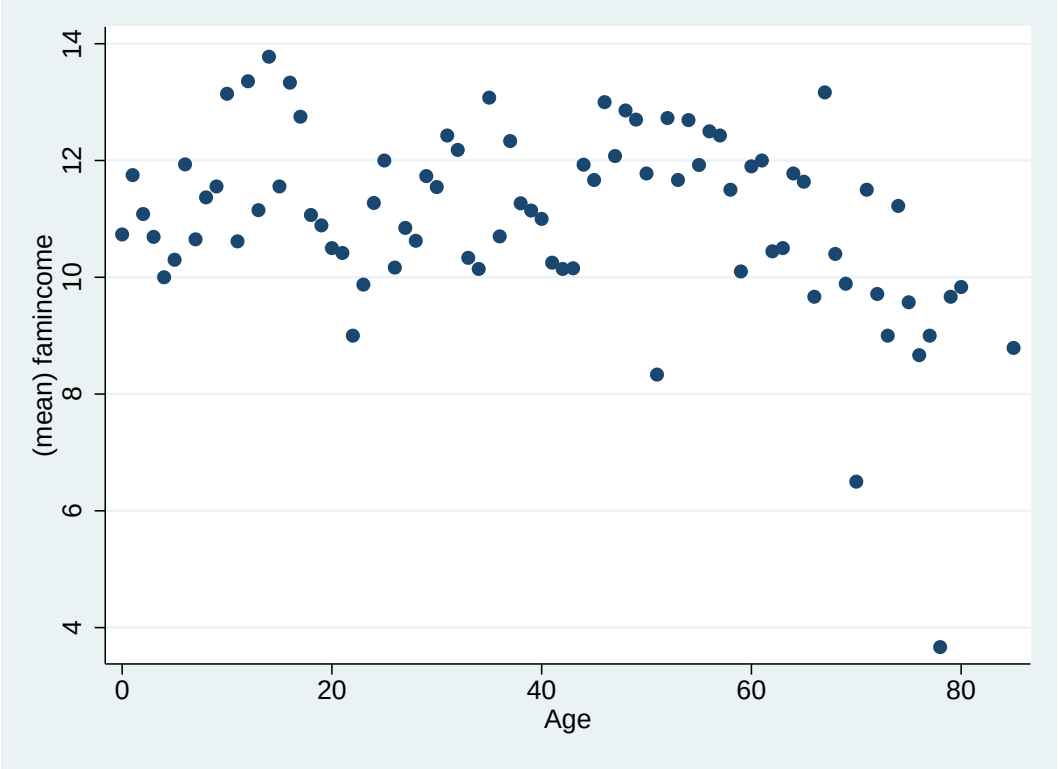
To aggregate the data so that we are able to produce a meaningful plot, we must aggregate the data by age. Particularly, we use the "collapse" statement, which aggregates the specified variable by the variable or variables in the "by" option. "mean" in parantheses specifies that we want to take the mean of "famincome" for each value of "age". There are also many other possible statistics that we could collapse by, including median, observation count, and sum.

collapse (mean) famincome, by(age)

Plot again to see the difference

To now see the difference, construct the same plot again. It looks much better now!

graph twoway (scatter famincome age)



Plot with Options and Delimiting

Our plot now has some meaning, but there are many things we may want to change to make it look nicer. Before we do that, we must understand delimiting. By default, Stata is delimited as carriage return, which means that each line indicates a separate command. Alternatively, if we delimit by semicolon (;), the command does not end until a semicolon is used, even if it spans multiple lines. To switch between the two, we use "#delimit ;" and "delimit cr". In order to create a graph with many options and make the code look clean, we delimit by semicolon and then revert to carriage return after.

To add options to the graph, we specify them after the comma just like always. "title" specifies the title of the graph, which is at the top of the plot region by default. "xtitle" specifies the title of the x-axis. "ytitle" specifies the title of the y-axis. "xlabel" specifies exactly how the x-axis should be labeled. "ylabel" specifies exactly how the y-axis should be labeled. In this case, I do not change the y-axis labels, but I do specify "angle(horizontal)" to flip the y-axis labels and "nogrid" to get rid of the lines. "graphregion(style(none) color(white))" makes the background white instead of blue. Note that there are many options to make the graph look exactly how you want which you can find by using "help graph twoway" and clicking on "twoway_options" in the Syntax specification.

#delimit ;

graph twoway (scatter famincome age),

title(Family Income Distribution Across Ages)

xtitle(Age)

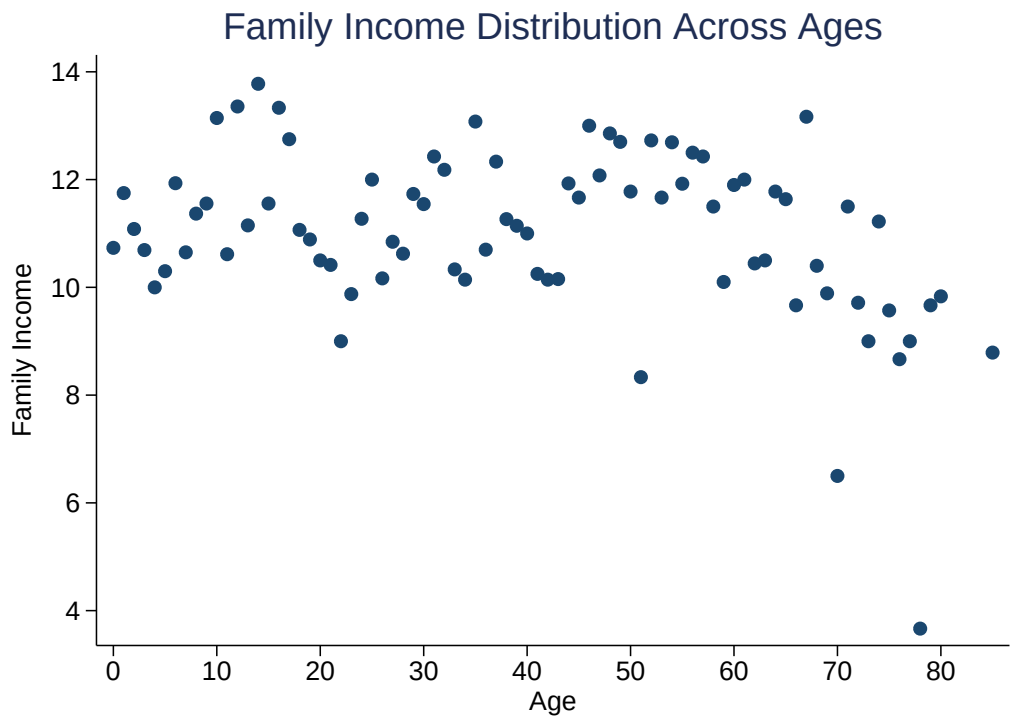
ytitle(Family Income)

xlabel(0 10 20 30 40 50 60 70 80)

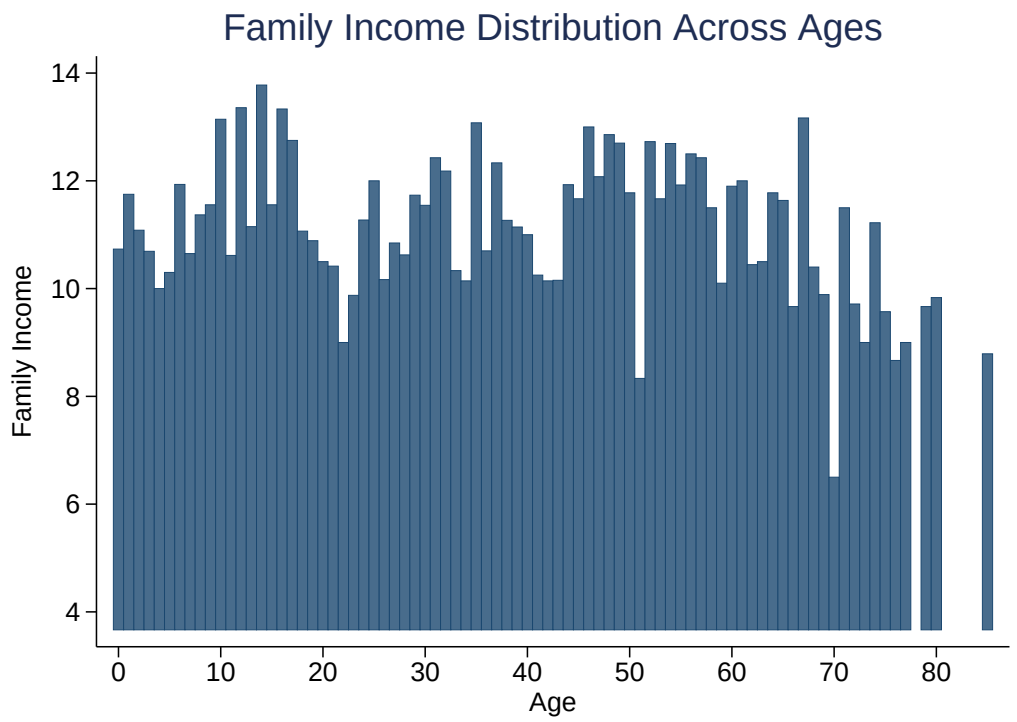
ylabel(, angle(horizontal) nogrid)

graphregion(style(none) color(white)) ;

#delimit cr



Alternatively, we can change "scatter" to "bar" and represent the data as a bar chart.



Analysis

One Sample T-test

The command "ttest" allows us to perform one sample, and two-sample (paired and unparied) t-tests. For a one sample t-test, we simply compare a variable to a value, which tests whether the mean of the variable is statistically the same as the value. The output provides the t-stat and p-values for one and two sample tests. The level option specifies the confidence level of the t-test.

ttest collegegrad == .4, level(95)

Two-Sample Unpaired T-test

"ttest" can perform a two-sample unpaired t-test by comparing the means of a given variable depending on the values of a variable. In the example below, we compare family income for individuals over 22 years old depending on college graduation status. Specifically, for individuals over 22 years old, the mean family income for college graduates is compared to the mean family income for college graduates.

ttest famincome if age >= 22, by(collegegrad)

Linear Regression

To perform a linear regression, we use the command "regress". The first variable is the dependent variable and all others are the independent variables. One of the nice features of

Stata is that options such as robust standard errors (to correct for heteroskedasticity) and clustered standard errors (accouting for correlations within clusters) are built in and very easy to use. Here, we will simply regress "famincome" on "collegegrad" with robust standard errors. Since "collegegrad" is binary, the coefficient tells us the difference between the mean of family income for college graduates and the mean of family income for non-college graduates.

regress famincome collegegrad, robust

User Installed Programs

Perhaps the most useful feature of Stata is that there is a very active community of skilled programmers that have written and continue to write commands that perform many of the procedures that we want to implement. For example, in economics, there are user-written commands to perform tasks such as instrumental variables regressions,and regression discontinuity regressions and plots (and selecting optimal bandwidths). These commands are stored on an internet database and we can install these programs very easily using "ssc install".

ssc install outreg2

The above program is a simple but very useful program which allows us to export regression results to a file in order to create results tables. After running a regression, we run "outreg2". The "excel" option specifies that the output file is an Excel spreadsheet (can also output to tex, word, or ASCII). The dec(3) option automatically rounds results to 3 decimal places. The "replace" option specifies that the command will replace all results on the output file (causing that regression to be the only output on the file), while "append" adds the results of that regression in another column on the file. The code below performs two regressions and exports the results of each to the same Excel file.

regress famincome hsgrad, robust

outreg2 using "C:\Stata Workshop\outreg.xls", excel dec(3) replace

regress famincome collegegrad, robust

outreg2 using "C:\Stata Workshop\outreg.xls", excel dec(3) append