SOC 6304: Social Statistics

Introduction to Stata

Accessing Stata

- Lab
- For purchase
  - Stata/IC
  - For mid-sized datasets.
  - Perpetual: $198
  - Annual: $89
  - 6 months: $45

Stata File Types

- Data files
  - .dta
  - Includes rows/columns of your data
- Syntax files
  - Do files (.do)
  - Is the scripting file that tells Stata what to do
- Output files
  - Log files (.txt)
  - Includes the results of what you told Stata to do

Stata Syntax and Scripting

- Point and click is a waste of time
- Only reasonable way to conduct your stats is to script
- A syntax file is just a text file that stores your commands
- Syntax or script files are called “do files” in Stata (.do)

Stata/SE 14.0

File Edit Data Graphics Statistics User Window Help

Review

Filter commands here

#	Command	_rc

There are no items to show.

```
  ___  ____  ____  ____  ____  (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   14.0   Copyright 1985-2015 StataCorp LP
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  401406229773
         Licensed to:  Kate
                       UH

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
      3.  New update available; type -update all-

.
```

Command

Variables

Filter variables here

Name	Label

There are no items to show.

Variables Properties

Opening a Data Set

- There are several commands you should include at the top of every do file
- The “use” command opens the file in Stata

File Edit View Project Tools

syntax_lecture1 × Untitled.do ×

```
clear
clear matrix
capture log close
set more off
cd "C:\Users\kander20\Documents\Teaching\UH Fall 17 - 6304 Social Statistics\Lecture Examples"
log using log_lecture1.log, replace t

use GSS2016, clear

```

Opening Syntax for “do file”

- **clear –** closes any open datasets
- **capture log close –** closes any open log/output files
- **set more off –** tells Stata to produce all of the output without stopping
- **cd –** changes the working directory
- **log using –** creates a log file, which is a text file that stores your output
  - “**replace t**” is an option that overwrites the previous log file
- **use –** opens a dataset

Selecting a Subgroup of Variables

- If you know you are only going to use a small group of variables from a larger data, allows you to pare the data down
- **keep** – command which allows you to only “keep” the variables that you want
- Keep needs to be followed by a list of variables that you want to include
  - keep age race educ conlegis

Descriptive Statistics

- Can get the main descriptive statistics that we discussed today using one line of code in Stata
- **su, sep(0)**

- “su” means to summarize in Stata
- The “sep(0)” option tells Stata not to separate the variables with horizontal lines in the output
- If you don’t list variables, it will give you everything
- If you do, it will only give you the requested variables
  - su age, sep(0)

Frequencies

- For frequencies and cross-tabulations for variables with only a few categories, you can use the **tab** (tabulate) command
  - tab age
- More on cross-tabs later, but you can also use this option to cross-tabulate two variables
- You can also use the **tab1** command to get frequencies for several different variables
  - tab1 age race educ conlegis, m
- The **, m** option includes the missing values in the tabulation
- This is useful for data cleaning

Dropping Missing Cases

- Missing data is essentially missing information from the survey
- For example, if a respondent refused to answer a question
- We need to drop these values from our data because we cannot statistically analyze them as non-numeric values
- There are multiple ways to handle missing values, but for now we will just delete them
- Known as “list-wise case deletion”

Dropping Missing Cases

- 1) Get the descriptives and frequencies for each variable
- 2) Drop the missing cases
- 3) Look at the descriptives and frequencies to verify that the data are "clean"

- Drop the missing cases by using "drop if" or "keep if"

Do-file Editor - syntax_lecture1

File Edit View Project Tools

Assign1 × syntax_racehealth × syntax_lecture1 × syntax_lecture4 × Untitled.do ×

```
clear
clear matrix
capture log close
set more off
cd "C:\Users\kander20\Documents\Teaching\UH Fall 17 - 6304 Social Statistics\Lecture Examples"
log using log_lecture1.log, replace t

use GSS2016, clear

*Pulling out a subset of variables for analysis
keep age race educ conlegis

*Description of variables and basic descriptive statistics
describe
su, sep(0)

*Frequency Distributions
tab1 age race educ conlegis, m

*Cleaning the data set: listwise deletion of missing data
*The GSS uses a "." for all missing values
keep if age<.
keep if race<.
keep if educ<.
keep if conlegis<.

*Checking to see that the missing data points were dropped
tab1 age race educ conlegis, m
su, sep(0)

```