Introduction to Stata

Syllabus

Chris Ledford
Department of Political Science
University of Kentucky
E-mail: chris.ledford@uky.edu

Course description

Stata is a general-purpose statistical software package and is one of the main statistical packages used in the social sciences. At UK Stata is taught at the graduate level in political science, sociology, economics, public policy, and a number of other departments.

Introduction to Stata is a short, two-hour course intended for students with limited or no previous experience with Stata. The course focuses on simple data analysis such as reading files, combining and modifying data, and statistical analysis. After taking this course, students will be able to do basic data analysis using Stata for their classes and research projects.

Topics to be covered

- Background, advantages, and disadvantages of Stata
- Using help
- Data files
- Programs
- Importing and exporting
- Looking at data
- Modifying data sets
- Graphing
- Macros
- Statistical analysis
- Making a table

References

- http://www.ats.ucla.edu/stat/stata/
- http://data.princeton.edu/stata/
- Short YouTube tutorials on
    - linear regression,
    - Fixed and random effects
    - logistic regression
    - count models
- Hamilton, Lawrence. "Statistics with STATA: version 12" 8th Edition. Stata Press, 2012
- Acock, Allan C. "A Gentle Introduction to Stata." 3rd ed., Stata Press, 2010.
- Long, Scott. "The Workflow of Data Analysis Using Stata" Stata Press 2008

```
version 13
clear all
set more off
cd "C:\Users\cwled\Dropbox\Stata Workshop
capture log close
log using 1-25-2016.smcl, replace
***************************************************************************
*File Name:          Stata Workshop.do
*Date:               January 25, 2016 - last modified Jan. 24, 2016
*Author:             Chris Ledford
*Purpose:            Introduction to Stata Workshop
*Input Files:
*Output File:
*To do:
***************************************************************************

**************************************
*The help command is your friend *
**************************************
        help

*****************************
*Loading Data into Stata *
*****************************

        ******************************************
        *Importing dataset from the Internet *
        ******************************************
                use http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi, replace
                clear all
        ***********************
        *Importing csv file *
        ***********************
                insheet using workshop.csv, clear
                clear all
        ****************************
        *Importing an excel file*
        ****************************
                import excel workshop.xlsx, sheet("workshop")
                clear all
        *****************************
        *Importing Stata dataset *
        *****************************
                use workshop_a.dta, clear
```

```
***********************
*Merging Two Datasets *
***********************

    merge 1:1 id using workshop_b.dta


    ***************************************************************************
        *This command performs a one-to-one merge; it merges the dataset already *
        *opened in Stata with another dataset.*

    ***************************************************************************

******************
*Data Management *
******************

    *******************
    *Basic Operations *
    *******************

        browse
            *************************
            *Browse the variable(s) *
            *************************

        sort id
            ***********************
            *Sort the variable(s) *
            ***********************

        edit
            ***********************
            *Edit the variable(s) *
            ***********************


    ***************************************
    *Examining Dataset for Missing Values *
    ***************************************

        misstable sum, all
            **************************************
            *creates a missing observation table *
            **************************************
    *****************
    *Modifying data *
    *****************

        lookfor schtyp
            **********************************
            *the lookfor command is also your friend*
            **********************************

        label variable schtyp "type of school"
            *********************************************
            *Label the variable schtyp "type of school" *
            *********************************************



        rename gender female
            *************************************************
            *Change the variable name from gender to female *
            *************************************************

        generate lnread=ln(read)
            *****************************************************
            *Generates a new variable, which is the log of read *
            *****************************************************

        drop schtyp
            *********************************************************
            *Drops and deletes the variable schtyp from the dataset *
            *********************************************************

        tabulate race, generate (r)


************************************************************************
         *From the categorical variable race, generates a set of dummy variables*

************************************************************************

        replace string = subinstr(string, "$", "",.)

    ***********************************************************************
            *cleans the variable by removing symbols that Stata doesn't understand*

    ***********************************************************************

        destring string, generate(string2)

    **********************************************************************
            *converts the variable from a block of text into a numerical variable*
    **********************************************************************



********************************************************
*Summary Statistics and Examining a Variable in Detail *
********************************************************

    describe
        ************************************
        *Describe the variables in dataset *
        ************************************

    tab female race
        ****************************************************
        *Performs a crosstab for variables gender and race *
        ****************************************************



    list female-read in 1/10
        *******************************************************************
        *List the first 10 observations for variables from gender to read *
        *******************************************************************

    summarize
        ********************************************
        *Summary statistics table of all variables *
        ********************************************

    summarize read math science write
        ******************************************************************
        *Summary statistics for variables read, math, science, and write *
        ******************************************************************

    summarize if read>=60
        *************************************************************
        *Summary statistics for all variables if variable read > 60 *
        *************************************************************

    summarize if prgtype=="academic"
        ************************************************************************
        *Summary statistics for all variables if variable prgtype = "academic" *
        ************************************************************************

    summarize read, detail
        **************************************
        *Summary the variable read in detail *
        **************************************
        centile
                ***************************************
                *Looking at variable values by percentile*
                ***************************************

        codebook
                ***********************************
                *Shows more details about a variable *
                ***********************************


        ******************************
        *Breaking down data by a group *
        ******************************


                tabulate prgtype

                tabstat read write math, by(race) stat(n mean sd)


        **************************************************************
                        *By race, show the N, mean, and standard deviation for variables *
                        *read, write, and math.*

        **************************************************************



****************************
*Looking at data graphically *
****************************

        histogram read, normal
                ***************************************************************
                *Histogram of variable read, and add a normal density to the graph *
                ***************************************************************

        ladder read
                **************************************************************
                *list of ladder of powers transformations. Pick the lowest value. *
                **************************************************************

        gladder read
                ****************************
                *Ladder of powers histograms *
                ****************************

        qladder read
                **************************************
                *Ladder of powers quantile-normal plots *
```

```
          *********************************************




*********************************************************************
*Graphing to show relationships between two or more variables *
*********************************************************************

     graph matrix read math science write, half
          ****************************************************************
          *Scatterplots of variables read, math, science, and write *
          ****************************************************************

     graph twoway (scatter write read)
          **********************************************
          *Scatterplot of variables write and read *
          **********************************************

     graph twoway (lpolyci write read)
          ***************************************************************
          *Local polynomial smooth plots with confidence intervals *
          ***************************************************************

     graph twoway (scatter write read) ///
          (lpoly write read)
          *********************************
          *Combine two previous graphs *
          *********************************

     graph twoway (scatter write read) ///
          (lpoly write read if race==1) ///
          (lpoly write read if race==2) ///
          (lpoly write read if race==3)
          **********************************************************
          *Previous graph by race; graph looks a little messy *
          **********************************************************

     graph twoway (scatter write read) ///
          (lpoly write read), by(race)
          ***********************************************
          *Better graph to show differences by race *
          ***********************************************
```

```
    graph twoway (scatter write read, mcolor(forest_green)) ///
        (lfit write read, lcolor(blue) lpattern(solid) cmissing(y)), ///
        by(, title(Reading Scores by Writing Scores) ///
        subtitle(Broken down by Socioeconomic Status) ///
        note(The blue line is a fitted value)) by(ses) ///
        graphregion(color(white)) bgcolor(white)
        *******************************************
        *More complicated version *
        *******************************************


        help scheme
        set scheme s2mono
        set scheme s1mono
        set scheme s2color

************
*Using e() *
************

    help regress
        *******************************************
        *Ask Stata to explain the regress command *
        *******************************************

    regress write read string2
        **************************
        *Just a simple OLS model *
        **************************

    histogram read
        ********************
        *Regular histogram *
        ********************

    graph save read.gph, replace
        *****************
        *Saving a graph *
        *****************

    histogram read if e(sample)
        *********************************************************************
        *Histogram of the variable using only the observations that made it *
        *into the model                                   *
        *********************************************************************
        graph save read2.gph, replace
                ****************************************************
                *Save previous graph and call the file "read2.gph" *
                ****************************************************

        graph combine read.gph read2.gph
                ********************************************
                *Combine graphs "read.gph" and "read2.gph" *
                ********************************************


***************************************
*Using factor (categorical) variables *
***************************************

        regress write read female race

        regress write read female r1 r2 r3 r5

        regress write read female ib4.race
                *********************************
                *Same results as previous model *
                *********************************

********************************
*Using global macros/variables *
********************************

        global y write
                *****************************************************
                *Assign the variable write to the global variable y *
                *****************************************************

        global x read female ib4.race

        *************************************************************************
                *Assign ariables read, female, and factors of variable race to global x *

        *************************************************************************

        reg $y $x
                ****************
                *Run the model *
                ****************

        regress write read female ib4.race
                *********************************
                *Same results as previous model *
                *********************************
```

```
********************************
*Outputting Results to a Table *
********************************

        regress $y $x
                ****************
                *Run the model *
                ****************
    *Install outreg2 if needed*
                ssc install outreg2

        outreg2 using workshop, ///
                excel replace label alpha(.01, .05) symbol(**, *) title(OLS)
                ***************************************************************
                *Generates an Excel file with the results from previous model *
                ***************************************************************
```