STM103 Spring 2008

INTRODUCTION TO STATA 8.0

This provides a brief introduction to using Stata 8.0 for the dataset analysis. Stata is available on all of the computers in the Kennedy School's computer lab. If you have a home computer you may want to purchase a copy of Stata from the CMO. Stata is available for Windows98, Windows 2000, Windows ME, Windows XP, Windows NT, Macintosh, and UNIX operating systems. The Stata User's Guide is also available from the CMO.

The commands outlined below assume that you are using Stata Release 8.0 for Windows. Throughout this text, anything appearing in **Bold** font is a Stata command, whereas anything in *red italics* is a variable name which you should change for your specific analysis. Menu commands are indicated as, e.g., **File | Open**, to indicate that you first go to the File menu and then choose the Open option. The `Blue Courier` is the type of output you should generate. As a shortcut, you can also just copy and paste any of the command lines here directly into your Stata Command window then run.

1. PREPARING BEFORE ENTERING THE LAB

GETTING THE SHARED DATASET

You will typically download data from the course web site, www.pippanorris.com under 'data'. Right click on the file that you want (which will end with ".dta") someplace convenient – say, in a folder on your M drive , or on a stick drive, or on your laptop C:/ hard drive. Call it something clear and simple (eg STM103) so you can easily find it again.

ASSIGNMENT 1 PREPARATION

The aim is to write a professional report assessing and comparing the problems of democratic governance reform in one world region. Pick your region:

- Latin America and the Caribbean,
- Africa,
- Asia,
- Central and Eastern Europe,
- Middle-East
- Western Europe

Think about the key problems of democratic governance in the region. From your experience and your reading, what are the priorities for agencies? Can you rank them? Focus on the most important 2-3 issues in the first instance. Then look carefully at the shared dataset codebook. Start by selecting 3-4 indicators which relate to the problems you have decided to focus upon. The shared class dataset provides the following indicators, along with many others:

1. Freedom House index of political rights and civil liberties
2. Polity IV Project Democracy and Autocracy scales
3. Cheibub and Gandhi Democracy-Autocracy classification
4. Vanhanen Democracy Index
5. World Values Survey/Global Barometers Attitudinal surveys
6. Kaufmann/Kray World Bank Institute Good governance indicators
7. Transparency International Corruption index

Here is an example we will start with

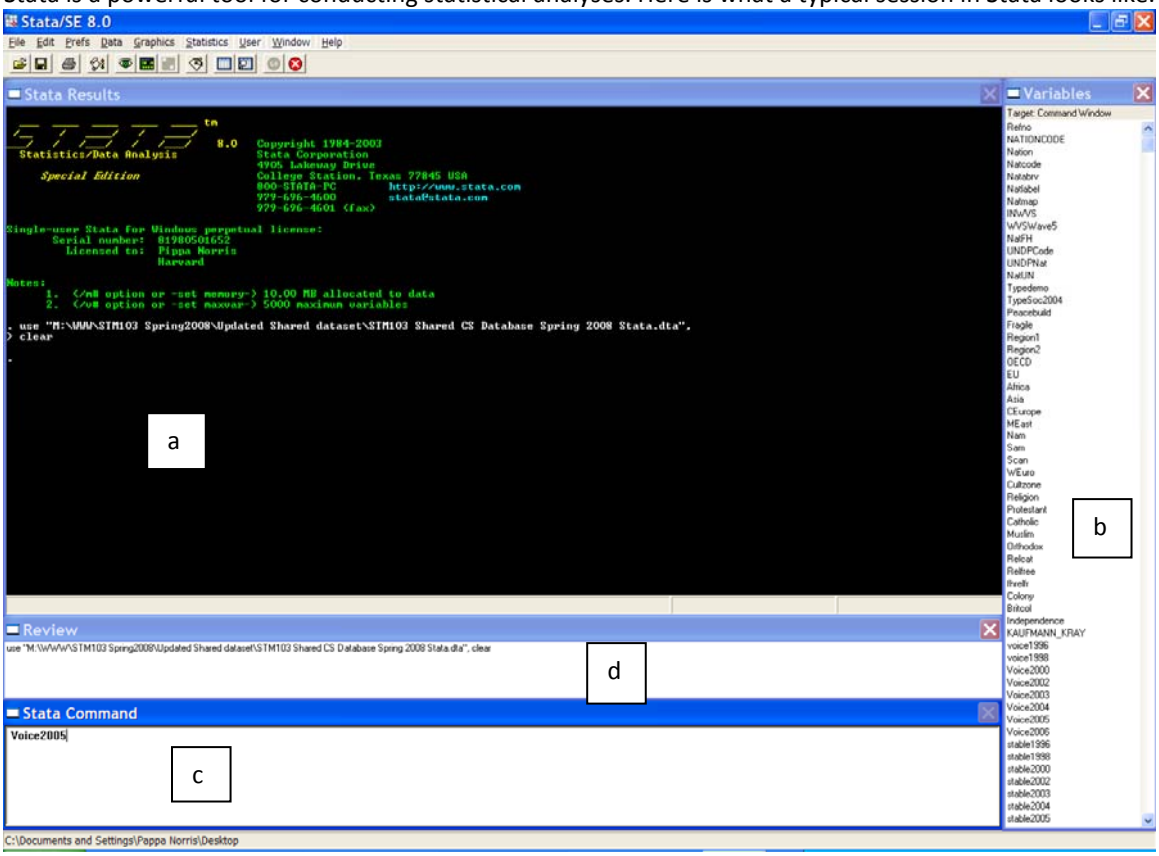
Name of your selected variables	**Notes**	**Type of data**
Typedemo2007	Freedom House classification of democracy in 2007	3 categories
Africa	Region Africa 1, rest of the world 0	2 categories
Typesoc2005	Type of society classified by Human Development Index	3 categories
Stable2006	Classification of political stability (Kaufmann)	100-point scale
GDP2006	Per capita GDP	100-point scale

Look at the codebook and jot down here the *exact* name of your 3-4 indicators to start the analysis:

Name of your selected variables	**Notes**	**Type of data**

2.STARTING A STATA SESSION

Start Stata by locating the link on the desktop and clicking.
Stata is a powerful tool for conducting statistical analyses. Here is what a typical session in Stata looks like.



The windows can be moved about and re-sized to suit your preferences. If you do not see any of these on your version, go to Window and add these until they look roughly like the above.

a. Here the *Results* window lists the outcome.

b. The *Variables* window, on the left, lists the names of all the variables included in the shared dataset.

c. You can enter commands in two ways. To start learning the program, you can use the drop down menus, similar to those common in Microsoft programs. This is useful for beginners. Once you become more familiar with the program, you will want to type in commands directly, to save time, using the *Command* window at the bottom center of the screen. A command tells Stata what to do – e.g, to open a file, to run a regression, to calculate a mean of a variable, etc.

d. The *Review* window shows a list of all the commands you have already run. (Here, it shows that I have opened a data file.) If you click on a previously-run command in the Review window, it will appear in the Command window and you can edit it or run it again.

OPENING, SAVING, AND CLOSING THE DATA FILE

You will then need to open the data file you have saved. You will need to boost the memory allocated.

Set memory 80000
File | Open
File| Save As
It is also always useful to have a backup copy of your data. This way, no matter what you do to change or recode the variables, you always have a copy of the older version. It is also useful practice to save your data file at the end of each session under a new sequential version (eg STM103_2, STM103_3, so that you have the old and newest file in case you need to revert back.

File | Exit
When ever you finish, to exit Stata.

KEEPING A LOG RECORD OF A STATA SESSION

File| log
To save a file ("log") of your results, you will need to create a log file. Stata gives you two choices of file formats for your log file, .log (text file) and .smcl (formatted log file). The .smcl files will look nicer when printed. You should ***never*** ever cut and paste your Stata output directly into your report; always simplify, clean and transfer in a professional and clean format.

To start a log file interactively, choose File | Log | Begin, select the directory you want to save the log file in, and give it a name (such as job1). Alternately, you can click on the fourth icon from the left on the icon bar, which looks like a scroll.

File |Log | Close

USING OPERATORS

Stata uses the following arithmetic operators:

+	add
-	subtract
*	multiply
/	divide
^	raise to the power

For relations, Stata uses:

= =	equal
~=	not equal
>	greater than
<	less than
>=	greater than or equal to
<=	less than or equal to

Note that a single equal sign (=) is used when assigning a value to a variable:
gen wage = salary/(hours*weeks)
but a double equal sign (= =) is used when asking Stata to make a comparison:
replace fulltime = 1 if hours == 40

For logical operations, Stata uses:

&	and
|	or (pipe sign; what you get when you hit Shift and the "\" key)
~	not

3.WORKING WITH DATA

Let's get some basic descriptive results. First we can look at some of the most common variables and once we have completed this exercise you should substitute the 3-4 variables have chosen to get a sense of what is available.

Let's start with some different types of variables. Nominal categories have no particular order, such as North, South, East, West. Ordinal categories have a sequential order but a limited number of categories, such as High, Medium and Low. Scale variables are ordered into a continuous series, for example level of GDP in dollars.

Note that when typing variable names the capitalization matters, follow the exact labels in the var list. Try the following commands.

TO SELECT VARIABLES AND SEE WHAT THEY CONTAIN:

. summarize *Typedemo2007 GDP2006*

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
Typedemo2007 |       190    1.757895    .7996797          1          3
     GDP2006 |       163    6184.601    10127.44         93      54779
```

This is very useful for looking at your selected variables to see what they are like, whether nominal, ordinal or scale (continuous). Typedemo2007 is an ordinal 3- category variable. Per capital GDP2006 is a continuous scale. Try this for a couple of your variables and add notes to your selected vars on page 2. **summarize** can be abbreviated to **sum**. You can also look at more detail. eg

sum GDP2006, detail

You can do the same just for your region

sum GDP2005 if Africa == 1, detail

TO LOOK AT THE % DISTRIBUTION IN CATEGORICAL VARIABLES

For categorical variables, try the following which generates some simple frequencies ie look at the number of countries (Freq) and the percent column.

. tab1 *Typedemo2007 Typesoc2005*

```
-> tabulation of Typedemo2007

    Type of |
  democracy |
 2007 (from |
    Freedom |
House 2008) |      Freq.     Percent        Cum.
------------+-----------------------------------
       Free |         89       46.84       46.84
Partly free |         58       30.53       77.37
   Not free |         43       22.63      100.00
------------+-----------------------------------
      Total |        190      100.00
```

```
-> tabulation of Typesoc2005

          Type of human |
    development, (3-cat, |
    classified from HDI) |
                UNDP2008 |      Freq.     Percent        Cum.
-------------------------+-----------------------------------
   Low human development |         25       14.12       14.12
Medium human development |         83       46.89       61.02
  High human development |         69       38.98      100.00
-------------------------+-----------------------------------
                   Total |        177      100.00
```

TO LOOK AT A TABLE COMBINING TWO CATEGORICAL VARIABLES

tab2 *Typedemo2007 Typesoc2005*, **col chi2**

```
-> tabulation of Typedemo2007 by Typesoc2005

    Type of |
  democracy |    Type of human development,
 2007 (from |   (3-cat, classified from HDI)
    Freedom |             UNDP2008
House 2008) | Low human  Medium hu  High huma |     Total
------------+---------------------------------+----------
       Free |         3         24         51 |        78
            |     12.00      28.92      73.91 |     44.07
------------+---------------------------------+----------
Partly free |        14         35          9 |        58
            |     56.00      42.17      13.04 |     32.77
------------+---------------------------------+----------
   Not free |         8         24          9 |        41
            |     32.00      28.92      13.04 |     23.16
------------+---------------------------------+----------
      Total |        25         83         69 |       177
            |    100.00     100.00     100.00 |    100.00

          Pearson chi2(4) =  43.7325   Pr = 0.000
```

TO LOOK AT A TABLE COMBINING TWO CATEGORICAL VARIABLES FOR YOUR SELECTED REGION

bysort *Africa*: **tab2** *Typedemo2007 Typesoc2005*, **col chi2**

```
-> Africa = 0
-> tabulation of Typedemo2007 by Typesoc2005

    Type of |
  democracy |    Type of human development,
 2007 (from |   (3-cat, classified from HDI)
    Freedom |             UNDP2008
House 2008) | Low human  Medium hu  High huma |     Total
------------+---------------------------------+----------
       Free |         0          17          50 |        67
            |      0.00       28.33       74.63 |     52.34
------------+---------------------------------+----------
Partly free |         1          26           8 |        35
            |    100.00       43.33       11.94 |     27.34
------------+---------------------------------+----------
   Not free |         0          17           9 |        26
            |      0.00       28.33       13.43 |     20.31
------------+---------------------------------+----------
      Total |         1          60          67 |       128
            |    100.00      100.00      100.00 |    100.00

          Pearson chi2(4) =  30.5063   Pr = 0.000
```

```
-> Africa = 1
-> tabulation of Typedemo2007 by Typesoc2005

    Type of |
  democracy |    Type of human development,
 2007 (from |   (3-cat, classified from HDI)
    Freedom |             UNDP2008
House 2008) | Low human  Medium hu  High huma |     Total
------------+---------------------------------+----------
       Free |         3           7           1 |        11
            |     12.50       30.43       50.00 |     22.45
------------+---------------------------------+----------
Partly free |        13           9           1 |        23
            |     54.17       39.13       50.00 |     46.94
------------+---------------------------------+----------
   Not free |         8           7           0 |        15
            |     33.33       30.43        0.00 |     30.61
------------+---------------------------------+----------
      Total |        24          23           2 |        49
            |    100.00      100.00      100.00 |    100.00

          Pearson chi2(4) =   3.6282   Pr = 0.459
```

This allows you to compare the rest of the world (Africa=0) in the first table with the situation in Africa (Africa=1) in the second.

Creating and Changing Values of Variables

recode

This helps to categorize the value of an existing scale variable. For example, suppose you have a variable called "GDP2006" that contains the per capita income for each nation in your study. But you want the income to be collapsed into two groups, say less than $5999 and over $6000. To do this, create a new variable called "GDPCAT2006" and recode it as follows:

gen GDPCAT2006= GDP2006
recode GDPCAT2006 min/5999=0 6000/max=1
sum GDPCAT2006

When recoding your data, be careful not to overwrite your original variable. You can check what you have done with the summary.

generate

Creates a new variable, in this case "GDP2006MULT", defined as "GDP2006" multiplied by 50:
generate GDP2006MULT =GDP2006 * 50

You can abbreviate this command with "gen". Note that Stata will tell you if any missing values were generated by attempting to perform a calculation with missing information. For example, if one of the observations was missing information on "hours," Stata would set "yrhrs" equal to missing for this observation. (See further notes about missing data later in this section.)

Once a variable with a particular name has been generated you can't generate another with the same name. Instead, you must replace the old one.

Labeling Variables and Values

Labeling variables and values helps you keep track of how you coded your variables and what they represent. It takes just a couple of seconds to add labels, and it can save you lots of time later when you can't remember what the a code of "4" means in your GDP category variable, for example, or how the variable "demo1" differs from "demo2".

label variable

To attach a label to a variable:

label variable GDPCAT2006 "Per Capita GDP 2006"

label value

label define GDPCAT2006 0 "Poor" 1 "Rich"

USING FUNCTIONS

Functions are special calculations used with other commands, such as generate or replace. Stata has the capability to calculate many functions. Here are some examples of the most commonly used ones.

ln(x)

Calculates the natural log of x, where x may be a constant or a variable such as "GDP2006", or an equation such as (GDP2006 + EDUC2006)

ln(1.5) or ln(GDP2006) or ln(GDP2006 + EDUC2006)

In a command, you might use the ln function like this:

gen logGDP2006 = ln(GDP2006)

DELETING VARIABLES AND OBSERVATIONS

drop

The drop command can delete either variables or observations. Deleting a variable removes an entire variable (column) from the data set, whereas deleting an observation removes an entire observation (row) from the data set. Be careful when doing this – the variables and observations are permanently deleted once you save the data file! It is far better to retain the whole dataset but to filter for the selected region.

To eliminate a variable, in this case "GDP2005":
drop GDP2005

To eliminate observations, in this case those affluent nations for which "Typesoc2005" is one:
drop if Typesoc2005 == 1

LIST

Prints all variables and observations to the screen. You'll probably never want to do this since your data sets will be too large.

list

You can print a limited set of variables:

list Nation Typedemo2007 Typesoc2005

You could also print a limited set of observations according to another criteria, in this case "Africa" being equal to 1:

list Nation Typedemo2007 Typesoc2005 if Africa == 1

codebook

Provides even more information (mean, standard deviation, range, percentiles, labels, number of missing values, etc.) about a variable:

codebook GDP2006

ANALYSIS OF CONTINUOUS (SCALE) VARIABLES

EXAMINING MEANS BY CATEGORY

In this case the category is Typedemo2006 and the mean is calculated for Voice2006.

bysort Typedemo2006**: tabstat** Voice2006**, columns(variables)**

CORRELATIONS

corr Stable2006 GDP2006

With signicance (P) printed below in stars for all coefficients significant at .05 or above

pwcorr Stable2006 Voice2006 GDP2006, **star(5)**

TESTING HYPOTHESES ABOUT MEANS

ttest

ANOVA is used to test a hypothesis that the means of two groups are significantly different, in this case where the two groups are defined by the regional variable "Africa":

ttest GDP2006, by(Africa)

ESTIMATING LINEAR MODELS (OLS AND 2-STAGE LEAST SQUARES)

regress

Calculates an ordinary least squares (OLS) regression, in this case for a regression of the dependent "Stable2006" on the independents "GDP2006" and "Africa". Note that the dependent variable is the first variable listed.

regress Stable2006 GDP2006 Africa

If you wish to only include observations with "Africa" equal to 1 in the regression:

regress Stable2006 GDP2006 fhrate07 if Africa == 1

To run a regression with robust standard errors:

regress Stable2006 GDP2006 Africa, **robust**

To run two-stage least squares where "GDP2006" is endogenous and "z1" is an exogenous instrumental variable:

regress Stable2006 GDP2006 Africa (GDP2006 z1)

Note: If you run a regression containing more than 40 variables, Stata will return an error code saying: matsize too small

To overcome this problem, reset the maximum number of variables Stata will estimate using the matsize command; the number should be greater than or equal to the total number of variables in the regression.

set matsize 150

predict

Calculates the predicted value for each observation using the coefficients from the last regression estimated and saves these as a variable called "yhat":

predict yhat

To calculate the residual for each observation using the most recently estimated regression model and save these as a variable called "ehat":
predict ehat, residual

test

Calculates an F-test of a joint hypothesis concerning the coefficients in the most recently estimated linear regression model, in this case with the null hypothesis H0: βage = βsex = 0:

test Stable2006 GDP2006

ESTIMATING NON-LINEAR MODELS (LOGIT AND PROBIT)

LOGIT

Estimates a model suitable for a dichotomous dependent variable. In this case, the variable "CheibubType2" equals 1 for democracy and 0 for autocracy.

logit CheibubType2 Stable2006 GDP2006

If you wish to find a predicted probability for each observation based on the most recent model run and save these as a variable called "phat":

predict phat

PROBIT

Estimates a model suitable for a dichotomous dependent variable. In this case, the variable "CheibubType2" equals 1 for democracy and 0 for autocracy. If you wish to estimate the probability of "CheibubType2" conditional upon Stable2006 and GDP2006:

probit CheibubType2 Stable2006 GDP2006

If you wish to find a predicted probability for each observation based on the most recent model run and save these as a variable called "phat":

predict phat

4. MAKING GRAPHS

Stata 8 has a Graphics menu that lets you create graphs from a windows menu, as an alternative to using command language. The Graphics menu is a particularly user-friendly way of creating graphs, since graphs contain so many options for labels, axes, etc. The Graphics menu is fairly intuitive to use—simply pull down the menu and choose the type of graph you want. The options are self-explanatory. For those interested in using command language to create graphs, some of the basics are covered below, and you can reply on the graphics manual for more complicated creations. Also SPSS has better and far more flexible graphics. You may want to consider this program for this function alone. You can also cut and paste the results of tables into Excel for flexible formats and control over elements.

HISTOGRAM

This is the default when only one variable is specified:
histogram fhrate07

You can also draw a normal density over the histogram:
histogram fhrate07, normal

To have STATA graph only certain observations, in this case those for which "Africa" is 1:

histogram fhrate07 if Africa == 1, bin(30)

To add a title:

histogram fhrate07, title("Freedom House Rating of Liberal Democracy, 2007, in Africa")

SCATTERPLOT

This is the default if two variables are specified:

scatter Stable2006 GDP2006

Conditions, axes, titles, labeling and reference lines can be specified as above. For example:

scatter Stable2006 GDP2006, t1(Stability by income)

After performing a regression, you may want to graph predicted and actual values of the dependent variable against the independent variable:

scatter yhat1 Stable2006 GDP2006, xlabel ylabel symbol(o.)

BAR GRAPHS

This is produced with a graph command followed by one variable. A second variable is used to define groups. To produce a graph with bar heights representing the mean for each group:

sort Stable2006
graph bar (mean) Stable2006, over(Africa)

Conditions, y-axis options, most titles, and horizontal reference lines can be specified as described above with regard to histogram:

```
sort Stable2006
graph bar (mean) Stable2006, over(Africa) t1(Political Stability in Africa) t2(Title 2) l1(Mean Stability 2006)
        l2(Another Title) yline(33000)
```

PRINTING YOUR GRAPH

Stata allows you to print (File | Print Graph) and save (File | Save Graph) your graphs. The easiest way to incorporate your graph into a Word document is to copy the graph to the clipboard using Edit | Copy Graph and then paste it into your document. Remember that all graphs should have a clear headline, to illustrate your report, with a full note below specifying the source of the data and any notes explaining variables. All graphs should be self-contained without looking further in your report.

5. UTILITIES

VIEWING THE DATA

Once you have opened a data set, you may wish to look at the variables and observations in spreadsheet format. Stata provides two ways to do this, "browse" and "edit". The browse command lets you see the data but not make changes, whereas the edit command allows you both to browse and to make changes. It is probably best to use browse unless you actually intend to make changes to your data manually; otherwise you may accidentally change something and ruin your data.

To browse, enter browse into the Command window or select the Browse icon (third from the right, a spreadsheet with a magnifying glass on it). To edit, enter edit into the Command window or select the Edit icon (fourth from the right, a spreadsheet with no magnifying glass).

CREATING AND SUBMITTING A DO FILE

Although Stata can be run interactively by just typing one command at a time, Stata commands can also be submitted in batches by using a "do file." A do file is simply a text file which contains a series of Stata commands. You enter the Stata commands in the same order as you would enter them interactively, and Stata then runs these commands automatically instead of your having to type them in line by line.

For your problem sets, it is strongly recommended that you use do files. Some of the problem sets will require many Stata commands, and it is inevitable that you will need to make changes and run these series of commands a number of times. When you have all of your commands in a single file, it is much easier to go back to that file and make the necessary changes than to have to retype every command.

Creating a Do file

To start creating a do file, click on the Do file editor button (fifth from the right, looks like an envelope with a pencil on it), choose the Do file editor option under the Window menu, or type doedit in the Command window. Note that since a do file is a written list of commands as entered in the Command window, you cannot use the Stata menus within a do file. Instead you need to use the typed (Command window) commands.

CONVERTING DATA FILES (EXCEL TO STATA)

The easiest way to convert data files is to use the software program StatTransfer. This program is on the lab computers and allows you to convert your data to or from a variety of different file formats (Stata, SAS Transport, Excel, SPSS, QuatroPro, FoxPro, etc.).

To convert a file from Excel to Stata:

a) Click on the application StatTransfer in the "Data Analysis" folder.

b) Select "Excel Worksheet" for "Input File Type."

c) Use "Browse" to identify the Excel file you want to convert from. (If the first row of the worksheet contains the variable names, the program will use these as the variable names.)

d) Select "Stata Version 8" as the "Output File Type." (Since Stata 8.0 is a recent release, it is possible that the version of StatTransfer you're using will not have Stata Version 8 as an option. If this is the case, save it as a version 7 file; you should still be able to open the file in version 8.)

e) Type in the path and name of the file you wish to create.

f) Begin the conversion by clicking on "Begin Transfer."

Stata also allows you to read in binary and ASCII files directly. However, in most cases it is easier to first convert your data to a spreadsheet and then convert it to Stata using StatTransfer.