Tutorial: Introduction to Stata

Exercise 1: Do the following tasks.
(The command sequences and some further explanation is attached)

1. Open a session in Stata
2. Open the editor window
3. Clear and set the memory to 1 megabite
4. Download the dataset named auto.dta
5. Describe the dataset
6. List all the variables
7. Sort the observations by miles per gallon
8. List make and mpg for the first 5 observations only
9. List make and mpg for the last 5 observations only

10. Get the summary statistics for all the variables in the dataset
11. Get the summary statistics for the variable price
12. Get the summary statistics for the variable price if mpg is less than 21.3
13. Get the complete summary statistics (including various percentile and the median) for mpg
14. Get summary statistics on mpg separately by foreign status

15. Get the distribution for foreign (also try the no label option)
16. Define a label for rep78 (reparation in 1978) THIS IS THE FREQUENCY OF REPAIR RECORD ON A 1-5 SCALE 1=POOR...5=EXCELLENT
17. Then get the distribution of rep78
18. Generate a cross tabulation of repair and foreign status
19. Use a Chi square test to check whether the distribution of repair differs by foreign status
20. Generate a cross tab of repair and foreign status but with the cell frequency

21. Correlate mpg and weight
22. Correlate mpg and weight separately by foreign status and test significance of correlation
23. Generate a correlation matrix fro mpg, weight, price, length and displacement

24. Plot the relation between mpg and weight by foreign status and for the full sample

25. Generate a variable for the square of weight
26. Regress mpg on the quadratic in weight and foreign status. How do you interpret the coefficient on foreign
27. Generate the predicted value for mpg
28. Sort the observations by weight
29. Plot the relationship between mpg and weight, as well as your fitted regression. Do it separately by foreign status

Stata output with comments -showing commands required

Before you start copy the file auto.dta to your filespace and make a note of where it is

```
**********************************************************************
***********DIGRESSION
* This asterix means that the following text is just a comment which
STATA ignores but where the researcher can write notes
*It's very useful as when someone else looks at your do-file they
will understand what is going on.
*Equally if you haven't looked at a do file yourself in a long time
it will clear things up
/* you can also write comments like this */
/*stata will ignore any text between the "forward-slash-asterix" and
the "asterix-forward-slash".... */
/* .....even if
......
......it scrolls onto another line....
*/
*Everything I have written so far has been just a comment
**********************************************************************
*********************

*STARTING THE ANALYSIS
clear                                        /*THIS CLEARS THE MEMORY.
*/
set mem 1m              /*THIS SETS THE MEMORY SIZE - TOO SMALL FOR A
GIVEN DATA SET - STATA WON'T WORK SETTING IT TOO BIG WILL SLOW THE
OPERATION OF THE COMPUTER*/

use "h:\auto.dta"             /*SPECIF THE CORRECT PATH!!! THIS
COMMAND LOADS IN THE DATA - SO YOU MUST SPECIFY THE CORRECT PATH
OTHERWISE STATA WON'T KNOW WHERE TO LOOK YOU CAN USE EXPLORER TO
CHECK THE CORRECT FILE PATH NOTICE THAT STATA DATA SETS HAVE THE
".DTA" FILE ENDING*/

                                     /*WE HAVE NOW LOADED IN A SMALL
DATASET WHICH COMES WITH STATA AND IS USED FOR TUTORING*/

**********************************************************************
*****************************
describe

list                                 /*LETS YOU VIEW THE DATASET IN
THE MAIN STATA RESULTS WINDOW THIS IS OK IN THIS INSTANCE AS THE
DATASET IS QUITE SMALL NOT SUCH A GOOD IDEA IF THE DATASET IS LARGE
INSTEAD USE THE DATA EDITOR BUTTON (FOURTH FROM RIGHT ON THE TOOLBAR)
AN EXCEL-TYPE SPREADSHEET POPS UP.YOU CANNOT PROCEED UNLESS THIS
WINDOW IS CLOSED*/
                                     /*THIS DATASET CONTAINS VARIABLES
ON A SAMPLE OF CARS*/

*HIT THE SPACE BAR WHEN THE BLUE "MORE" MESSAGE IS DISPLAYED


sort mpg                             /*THIS SORTS THE DATA BY THE
VARIABLE MPG*/
```

```
list make mpg in 1/5                        /*THIS LISTS THE MAKE AND
MPG IN THE FIRST FIVE OBSERVATIONS*/

list make mpg in -5/-1                      /*THIS LISTS THE MAKE AND
MPG IN THE LAST FIVE OBSERVATIONS*/


/*AT THIS POINT WE SHOULD LOOK AT THE SYNTAX OF STATA COMMANDS

[prefix_cmd:] command [varlist] [if] [in] [, options]

anything inside square brackets [ ] is optional

so every line must have a command
 [varlist]  obviously tell STATA which variable(s) to use.
                  not specifying a variable or a list of variable
means that all variables are used
[if]              the inclusion of an "if" statement means the
command will only use observations satisfying the if statement
                  ==, <, >, <=, >=, != are the operators.
                  Note the double equals sign for equality. Brackets
are allowed.
                  “|” is "or".
                  “&” is "and".
[in]        "in" statements are less commonly used but when they are
used only a specified subset of observations will be used
 [, options]      some STATA commands have in-built options pre-
programmed. The comma is important
[prefix_cmd:]     usually involves the command "by VARNAME:". This
must be preceded with by sorting the data by VARNAME. It will then
perform the command for each group
            (we will return to "by" later)
*/
**********************************************************************
**************


*SOME DESCRIPTIVES
summarize price
summarize
summarize price if mpg<21.3
summarize mpg, detail                       /*THIS TELLS YOU A LITTLE
BIT MORE*/

summarize price mpg if foreign==0      /*IF STATEMENT USE A DOUBLE
EQUALS SIGN */
summarize price mpg if foreign==1




ttest mpg, by (foreign)       /*WE WANT TO TEST IF THE MPG OF OF
FOREIGN AND DOMESTIC CARS ARE EQUAL - THE COMMAND RUNS TWO TAILED AND
ONE TAILED TESTS*/
                        /*WE SEE THAT 1978 DOMESTIC CARS HAVE POORER
MILEAGE THAN FOREIGN ONES*/
**********************************************************************
**********************
*SIMPLE TABLES & LABELS

tabulate foreign
```

```
tabulate foreign, nolabel           /*THE DATA IS ACTUALLY SAVED IN A
BINARY WAY BUT IT HAS BEEN LABELLED "FOREIGN" & "DOMESTIC"*/

tabulate rep78                /*THIS IS THE FREQUENCY OF REPAIR
RECORD ON A 1-5 SCALE 1=POOR...5=EXCELLENT BUT IT ISN'T LABELLED IT'S
SUPERFICIAL BUT IMPORTANT IF SHARING DATA SO EVERYONE KNOWS THE
CODING*/
label define rating 1 "poor" 2 "fair" 3 "good" 4 "very good" 5
"excellent"
label values rep78 rating
tabulate rep78                /*LOOKS NICER BUT WE HAVE A PROBLEM 5
CARS DO NOT HAVE A VALUE FOR REP78*/


tabulate rep78 foreign        /*CROSS TABS - THERE SEEM TO BE POORER
FREQ OF REPAIR RECORDS FOR DOMESTIC CARS IS IT SIGNIFICANT?*/

tabulate rep78 foreign, chi2  /*FREQ OF REPAIR RECORDS DIFFER
STATISTICALLY*/

tabulate rep78 foreign,  row col          /*GIVES CELL PERCENTAGES*/


*************************************************************************
*****************************
*Correlation
pwcorr mpg weight       /*PRODUCES A MATRIX - TOP RIGHT HAND CORNER
IS OMITTED AS MATRIX IS SYMMETRIC*/


sort foreign
by foreign: pwcorr mpg weight,sig         /*AGAIN SAME RESULTS TWO
DIFFERENT METHODS. NOTE THE USE OF THE "SORT" AND "BY" COMMANDS*/

pwcorr mpg weight price length displacement           /*A LARGER
MATRIX*/
*************************************************************************
*********************************
*GRAPH
scatter mpg weight, by(foreign, total)    /*AN INTERESTING
BREAKDOWN*/
/*
In STATA 8 the power to create graphs using the menu/dialog boxes
greatly increased to allow for the varying needs of users
IN the main stata window select
Graphics>Twoway graph (scatterplot, line, etc.)
choose Scatter for the plot type
Enter weight in the X variable field and mpg in the Y variable field
Click on the by tab
Enter foreign in the variables field
check the Graph Total checkbox
Click ok
Gives you the exact same graph as before
*/


*THE RELATIONSHIP SEEMS TO DIFFER DEPENDING ON WHERE THE CAR IS FROM
*************************************************************************
************************************
*REGRESSION
*LETS MODEL THE RELATIONSHIP BETWEEN MPG AND WEIGHT
```

```
generate wtsq = weight^2
regress mpg weight wtsq foreign           /*KEEP IN MIND THAT FOREIGN
IS A DUMMY*/
predict mpghat                /*THIS IS A POST ESTIMATION COMMAND -
WE GENERATE THE FITTED VALUES*/

sort weight

twoway (scatter mpg weight) (line mpghat weight), by(foreign)
    /* OVERLAY PLOT ADDING THE FITTED VALUES TO THE SCATTER. */
*ALTERNATIVELY YOU COULD USE
scatter mpg weight || line mpghat weight, by(foreign)
```

***THIS EXERCISE MAY BE USEFUL: IT IS LESS ABOUT DATA ANALYSIS BUT MORE ABOUT THE PRACTICAL MANAGEMENT OF DATA**

```
**********************************************************************
************************************
*THE VIEWER
*the Viewer button in the main STATA window is important - it's the
picture of an eye
*here you can make use of STATA's help facility
*click on it and type "help summarize" - you will be shown the
correct syntax of the command and a few examples. Related commands
might appear
*the search button lets you type in a key word and displays existing
net resources provided by STATA Corp or other users
*the commands that STATA corp provide are called .ado files
*people often write programs (.do files) which are easily
downloadable
**********************************************************************
*******************************
*SAVING DATA
*after changing the data you may want to save it
*click on the save icon in the main STATA window. Be sure not to
overwrite existing data - use another name
**********************************************************************
*********************************
*CLEANING & EDITING DATA

*If using data which has been professionally collected  you are
unlikely to carry out any major changes to the data  in terms of
cleaning the data.
*If you have collected the data yourself cleaning can be done on an
observation by observation basis
*not exactly STATA's strong point but it got better in STATA VERSION
9
* Cleaning or raw data editing is never fun
*alernatively you can use commands like "drop, keep &  replace"
*combine these with if statements and they become very powerful

use "h:\qmss\tutorial 1\auto.dta", clear
keep if mpg<=14         /*we lose 66 observations*/


**********************************************************************
************************************
*READING RAW DATA FROM DIFFERENT FORMATS
```

```
*Many organistions will have the data in STATA format already
*A program like stattransfer will be useful but you can do this in
other ways


*Assume the dataset was created using a spreadsheet like Excel
*copying and pasting is an option if there isn't a whole lot of data
although problems can arise

*STATA HAS THREE MAIN METHODS
*first you have to tell stata where to look
clear             /*clears the memory*/
pwd               /*this tell you which is currently the working
drive*/
cd  "h:\qmss\tutorial 1"            /*change the directory to where
the file sample.txt is saved*/


*METHOD 1
*So the method STATA Corp recommends is as follows:
*take the spreadsheet sample.xls for example.  open it and look at
it.
*it has also been saved a txt file sample.txt - open this file up
using a text editor - the data is tab delimited
*this can be read into stata
insheet using sample.txt
list              /*ths variable names aren't very descriptives*/
clear             /*let's clear the data and try again*/
insheet make price mpg weight gear_ratio using sample.txt

list              /*much nicer*/


*METHOD 2
*A different example is sample2.txt which is  formatted in a
different way.
*open it up using a text editor
*Note that text is enclosed in "" and there are .  and *** for
missing values. Generally speaking the formatting is messier
*This is typical of what you might receive from a US Agency like BJS
clear
infile str18 make price mpg weight gear_ratio using sample2.txt
      /*we use an infile command*/
list

*METHOD 3
*How about this formatted data
*open up sample3.txt in a text editor - it looks fine but see what
happens when you try to read it in
*the data looks tab delimited but it isn't there are "hard" spaces
between the characters
clear             /*let's clear the memory and try again*/
infile make price mpg weight gear_ratio using sample3.txt

list              /*not good it's all messed up*/

*now you have to open up a text editor and write a dictionary
*A dictionary is a program that tells STATA how to delimit the data
*I've done this for you already
clear
```

```
infile using sample3dict.dct
list                    /*much nicer*/
*****************************************************************
```