R/Stata Introduction

Alexandra Naumenko

1 R vs Stata Introduction

1.a Advantages of Stata

- user friendly
- convenient canned packages for statistical analysis
- quality control via StataCorp
- most popular statistical programming software used in academia (for the social sciences)

1.b Advantages of R

- free
- open source (code can easily be shared internationally)
- advanced graphics/data visualization options
- R is very marketable skill for data science oriented private sector jobs

1.c Objective

We will perform the same exercise in both R and Stata so you can familiarize yourself with the syntax and get a sense of which one you would enjoy using more for assignments.

1.d Resources

Stata

- Stata Manual http://www.stata.com/manuals/u.pdf
- the help command
  example: *help reg* will pull up more detailed information about executing a regression.

R

- *?'command'* will pull up more detailed information about executing a regression. *??command'* performs a search.
- Heiss' "Using R for Introductory Econometrics" (2016) available for a cheap price on Amazon or as a free ebook on their webpage.

2 Example Routine in Stata

Your version might not already have some packages installed so keep the following commands for installing useful packages in mind.

- ssc install estout, replace

2.a Importing and Viewing Data

Start by opening a do file editor. The editor is useful for staying organized and having a script to look back to at a later time. The command window is mainly useful for experimenting with commands and using the help command.

1. determine working directory
   cd "insert file pathway here"

2. I recommend always having *clear* at the top of your do file so you can rerun the script smoothly

3. import the data[1]
   i*mport delimited mroz*

4. view the data to assess how effectively Stata read the data file *browse*

5. assess summary stats
   sum

2.b Cleaning the Data

1. Often you want to rename variables to keep syntax more compact or to clarify what the variable is. In our case, let's shorten hourstotalin1975 to hours.
   rename hourstotalin1975 hours

2. Perhaps you want to drop missing values for wage.[2]
   drop if wage == .
   item A useful addition to the dataset might be a dummy variable for whether the women works fulltime (i.e., 40 hours a week)
   *gen fulltime = hours>40*52*

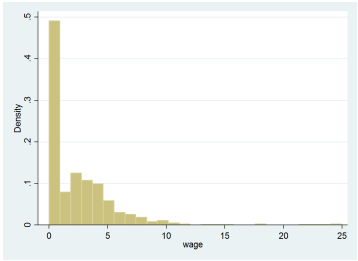
2.c Plotting data

You suspect that education and wage should be highly correlated. You want to investigate this question with visuals.
reg wage educ
predict fitted
scatter wage educ || line fitted educ

[1] syntax varies slightly depending on the type of file e.g., .dta, .xls etc
[2] you will learn in this class why this is often not a good idea

2.d Analysis

- We suspect that wage is not heavily distributed so we will look at a histogram
  hist wage



- Since the histogram validated our suspicion, we should log wage to correct for this. To simplify the analysis, we will restrict the sample to those who are employed (wage>0). All the log 0's would have resulting in missing values so the unemployed observations will drop out in the following regressions by default.
  gen lwage = log(wage)

- Education clearly doesn't explain wage entirely on its own so we would benefit from adding a control. Let's add in experience.
  reg lwage educ exper

2.e Presenting results

You are now happy with your analysis and you want to display your results in a table. Your audience will be interested in your significance levels so you want to make sure to include those!

reg lwage educ exper
eststo clear
eststo: quietly reg wage educ
eststo: quietly reg wage educ exper
esttab

Statistics/Data Analysis

```
      name:  <unnamed>
       log:  C:\Users\ANaumenko\Dropbox\590\log_day1.smcl
  log type:  smcl
 opened on:  22 Aug 2018, 11:51:56

1 .
2 . cd "C:\Users\ANaumenko\Dropbox\590"
  C:\Users\ANaumenko\Dropbox\590

3 .
4 . clear

5 . import delimited mroz
  (16 vars, 753 obs)

6 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
inlf	753	.5683931	.4956295	0	1
hoursto~1975	753	740.5764	871.3142	0	4950
kidslt6	753	.2377158	.523959	0	3
kidsge6	753	1.353254	1.319874	0	8
age	753	42.53785	8.072574	30	60
educ	753	12.28685	2.280246	5	17
wage	750	2.365276	3.241794	0	25
repwage	753	1.849734	2.419887	0	9.98
husage	753	45.12085	8.058793	30	60
huseduc	753	12.49137	3.020804	3	17
huswage	753	7.482179	4.230559	.4121	40.509
motheduc	753	9.250996	3.367468	0	17
fatheduc	753	8.808765	3.57229	0	17
unem	753	8.623506	3.114934	3	14
city	753	.6427623	.4795042	0	1
exper	753	10.63081	8.06913	0	45

```
7 .
8 . rename hourstotalin1975 hours

9 . drop if wage == .
  (3 observations deleted)

10. gen fulltime = hours>40*52

11.
12. regress wage educ
```

Source	SS	df	MS		
Model	802.963927	1	802.963927	Number of obs	= 750
Residual	7068.44834	748	9.44979724	F(1, 748)	= 84.97
				Prob > F	= 0.0000
Total	7871.41227	749	10.5092287	R-squared	= 0.1020
				Adj R-squared	= 0.1008
				Root MSE	= 3.0741

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.4531805	.0491626	9.22	0.000	.3566675	.5496935
_cons	-3.203405	.6144493	-5.21	0.000	-4.409656	-1.997155

```
13. predict fitted
  (option xb assumed; fitted values)

14. scatter wage educ || line fitted educ

15.
16. hist wage
  (bin=27, start=0, width=.92592593)

17.
18. gen lwage = log(wage)
  (325 missing values generated)

19.
20.
21. reg lwage educ exper
```

Source	SS	df	MS	Number of obs	=	425
				F(2, 422)	=	37.21
Model	33.3218942	2	16.6609471	Prob > F	=	0.0000
Residual	188.96186	422	.447776919	R-squared	=	0.1499
				Adj R-squared	=	0.1459
Total	222.283754	424	.524254136	Root MSE	=	.66916

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1098211	.0141757	7.75	0.000	.0819573	.137685
exper	.0157862	.0040258	3.92	0.000	.0078731	.0236993
_cons	-.4084307	.1906785	-2.14	0.033	-.7832286	-.0336327

```
22. eststo clear

23. eststo m1: quietly reg lwage educ

24. eststo m2: quietly reg lwage educ exper

25.
26. esttab
```

	(1) lwage	(2) lwage
educ	0.109***	0.110***
	(7.56)	(7.75)
exper		0.0158***
		(3.92)
_cons	-0.191	-0.408*
	(-1.03)	(-2.14)
N	425	425

```
  t statistics in parentheses
  * p<0.05, ** p<0.01, *** p<0.001

27. esttab m1 m2 using RegResults.rtf, label nonumber title("Education Effects on Wage")
  >  mtitle("Model 1" "Model 2") replace
  (output written to RegResults.rtf)
```

```
28.
29.
30. clear

31. input famid str4 name inc

        famid       name        inc
    1. 2 "Art" 22000
    2. 1 "Bill" 30000
    3. 3 "Paul" 25000
    4. end

32. save dads, replace
  file dads.dta saved

33. list
```

	famid	name	inc
1.	2	Art	22000
2.	1	Bill	30000
3.	3	Paul	25000

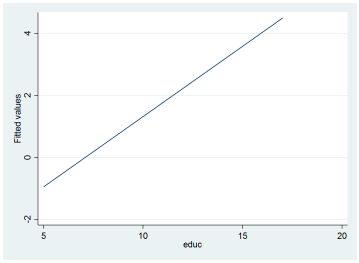
```
34.
35.
36.
37. clear

38. input famid str4 name inc

        famid       name        inc
    1. 1 "Bess" 15000
    2. 3 "Pat" 50000
    3. 2 "Amy" 18000
    4. end

39. save moms, replace
  file moms.dta saved

40. list
```

	famid	name	inc
1.	1	Bess	15000
2.	3	Pat	50000
3.	2	Amy	18000

```
41.
42. clear

43. use dads

44. append using moms

45. list
```

	famid	name	inc
1.	2	Art	22000
2.	1	Bill	30000
3.	3	Paul	25000
4.	1	Bess	15000
5.	3	Pat	50000
6.	2	Amy	18000

```
46.
47.
48. clear

49. use dads

50. rename name DadName

51. rename inc DadInc

52. merge 1:1 famid using moms

    Result                           # of obs.
    -----------------------------------------
    not matched                             0
    matched                                 3  (_merge==3)
    -----------------------------------------

53. list

     +-------------------------------------------------------------+
     | famid   DadName   DadInc   name     inc          _merge     |
     |-------------------------------------------------------------|
  1. |     1      Bill    30000   Bess   15000   matched (3)      |
  2. |     2       Art    22000    Amy   18000   matched (3)      |
  3. |     3      Paul    25000    Pat   50000   matched (3)      |
     +-------------------------------------------------------------+

54.
55.
56. log close
          name:  <unnamed>
           log:  C:\Users\ANaumenko\Dropbox\590\log_day1.smcl
      log type:  smcl
     closed on:  22 Aug 2018, 11:52:01
-------------------------------------------------------------------------------
```



	(1) wage	(2) wage
educ	0.453***	0.431***
	(9.22)	(9.02)
exper		0.0928***
		(6.87)
_cons	-3.203***	-3.924***
	(-5.21)	(-6.48)
N	750	750

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

3 Example Routine in R

Your version might not already have some packages used in this activity installed so keep the following commands for installing these packages in mind.

```
install.packages('data.table')
install.packages('stargazer')
```

This is a useful site for learning about the various packages:
https://cran.r-project.org/web/packages/available_packages_by_name.html

3.a Importing and Viewing Data

Start by opening an R script. The editor is useful for staying organized and having a script to look back to at a later time. The console window is mainly useful for experimenting with commands and using the help command.

1. determine working directory
   setwd(insert your path here)[3]

2. import the data[4]
   i*mroz <- read.csv("mroz.csv")*

3. view the data to assess how effectively R read the data file *head(mroz)*

4. assess summary stats
   summary(mroz)

[3] it's annoying but R requires forward slashes in between folders
[4] syntax varies slightly depending on the type of file e.g., .dta, .xls etc

3.b Cleaning the Data

1. Often you want to rename variables to keep syntax more compact or to clarify what the variable is. In our case, let's shorten hourstotalin1975 to hours.
   names(mroz)[names(mroz) == "hourstotalin1975"] <- "hours"

2. Perhaps you want to drop missing values for wage.[5]
   mroz2 <- na.omit(mroz)

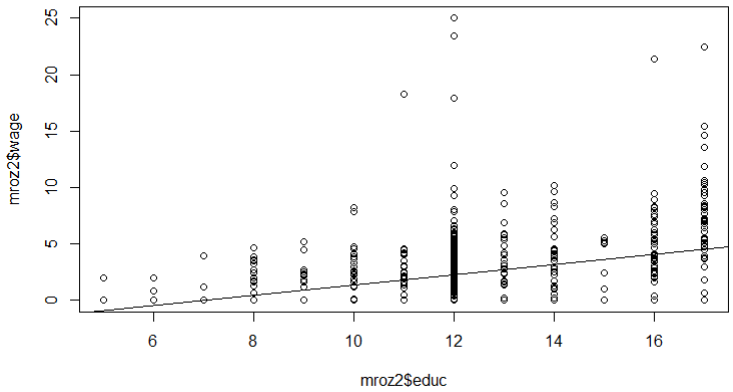
3. A useful addition to the dataset might be a dummy variable for whether the women works fulltime (i.e., 40 hours a week)

```
{mroz2$fulltime <- mroz2$hours > 40*52}
```

3.c Plotting data

You suspect that education and wage should be highly correlated. You want to investigate this question with visuals.

```
reg_1 <- lm(wage ~ educ, data = mroz2) \\
plot(mroz2$educ, mroz2$wage) \\
abline(lm(wage ~ educ, data = mroz2))
```

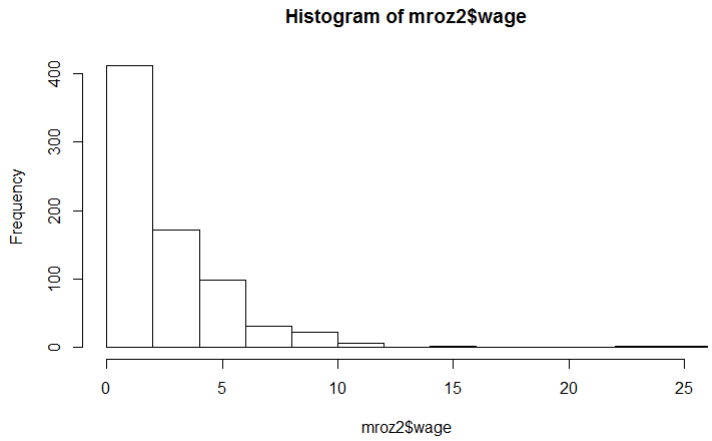


3.d Analysis

- We suspect that wage is not heavily distributed so we will look at a histogram

```
hist(mroz2$wage)
```

[5] you will learn in this class why this is often not a good idea



- Since the histogram validated our suspicion, we should log wage to correct for this. Unlike in Stata, we don't have to generate a new variable. We can just insert it into the regression command. To simplify the analysis, we will restrict the sample to those who are employed (wage>0).
- Education clearly doesn't explain wage entirely on its own so we would benefit from adding control. Let's add in experience.
- A benefit of R over Stata is that you can apply transformations to variables in the regression instead of generating them beforehand like you have to do in Stata[6]

```
install.packages('data.table')
library(data.table)
mroz3 <- as.data.table(mroz2)
mroz3 <- mroz3[which(wage >0),]


reg_2 <- lm(log(wage) ~ educ, data = mroz3)
reg_3 <- lm(log(wage) ~ educ + exper, data = mroz3)
```

3.e Presenting results

You are now happy with your analysis and you want to display your results in a table. Your audience will be interested in your significance levels so you want to make sure to include those!

```
install.packages('stargazer')
library(stargazer)
stargazer(reg_2,reg_3,type = "text")
```

[6]Stata will run using all the observations with missing values for lwage, but in R, you have to go out of your way to ensure your regression is only run on non-missing values to avoid an error.

Table 1:

	Dependent variable:	
	log(wage)	
	(1)	(2)
educ	0.109^{***}	0.110^{***}
	(0.014)	(0.014)
exper		0.016^{***}
		(0.004)
Constant	-0.191	-0.408^{**}
	(0.185)	(0.191)
Observations	425	425
R^2	0.119	0.150
Adjusted R^2	0.117	0.146
Residual Std. Error	0.680 (df = 423)	0.669 (df = 422)
F Statistic	57.099^{***} (df = 1; 423)	37.208^{***} (df = 2; 422)
Note:		*p<0.1; **p<0.05; ***p<0.01

.

```
setwd("C:/Users/ANaumenko/Dropbox/590")

mroz <- read.csv("mroz.csv")
head(mroz)
summary(mroz)
names(mroz)[names(mroz) == "hourstotalin1975"] <- "hours"

mroz2 <- na.omit(mroz)
mroz2$fulltime <- mroz2$hours > 40*52

reg_1 <- lm(wage ~ educ, data = mroz2)
plot(mroz2$educ, mroz2$wage)
abline(lm(wage ~ educ, data = mroz2))

hist(mroz2$wage)

library(data.table)
mroz3 <- as.data.table(mroz2)
mroz3 <- mroz3[which(wage >0),]


reg_2 <- lm(log(wage) ~ educ, data = mroz3)
reg_3 <- lm(log(wage) ~ educ + exper, data = mroz3)

library(stargazer)
stargazer(reg_2,reg_3,type = "latex")
```