Econometric Softwares: Stata

Jérémy Do Nascimento Miguel

Class 3 Spring 2024

Session 3 outline

- Overview of data and variables
- Frequency tables, counting and listing observations
- Generating and exporting descriptive statistics in Stata

General overview of data

Basic set of commands to get an overview of variables/data:

- describe [varlist]:overall descriptive of data, list of all variables, if varlist is specified then lists only those variables
- codebook [varlist] scribe content of all variables in more detail (but no statistics here), or varlist if specified
- list [varlist] [if]: list of all/selected (by if) observations from all variables or variables from varlist
  - → Check for irregular values: list if age<0
- count [if]: count nbr of observations in data

General overview of data

Basic set of commands to get an overview of variables/data:

- describe [varlist]:overall descriptive of data, list of all variables, if varlist is specified then lists only those variables
- codebook [varlist] scribe content of all variables in more detail (but no statistics here), or varlist if specified
- list [varlist] [if]: list of all/selected (by if) observations from all variables or variables from varlist
  - → Check for irregular values: list if age<0
- count [if]: count nbr of observations in data

Descriptive statistics: key commands

- summarize [varlist] [, detail]: descriptive statistics (mean, SD, max, min etc.) for all variables or variables from varlist , option detail gives more detailed descriptive statistics (percentile values, skewness etc.); short version sum
- tabulate varname1 [,summarize(varname2)]: one-way tabulation of all values (with frequency and cumulative distribution) from varname ; option summarize(varname2) provides descriptive statistics for variable varname2 across all values of varname1
- tabulate varname1 varname2 : two-way tabulation of all values from varname1 and varname2
- tabstat varlist [, stats(statistics) by(groupvar)]: compact descriptive statistics, one-way or two-way table (i.e. descriptive statistics by group)
- collapse: to generate data file with descriptive statistics (based on data in memory)

All commands allow: sample restriction using if or weights using [weight]

Descriptive statistics: *summarize*

```
. sum price mpg rep78

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906
         mpg |         74     21.2973    5.785503         12         41
       rep78 |         69    3.405797    .9899323          1          5
```

Descriptive statistics: *summarize*

```
. sum price, detail

                              Price
-------------------------------------------------------------
      Percentiles      Smallest
 1%         3291           3291
 5%         3748           3299
10%         3895           3667       Obs                  74
25%         4195           3748       Sum of wgt.          74

50%       5006.5                      Mean           6165.257
                        Largest       Std. dev.      2949.496
75%         6342          13466
90%        11385          13594       Variance        8699526
95%        13466          14500       Skewness       1.653434
99%        15906          15906       Kurtosis       4.819188
```

Descriptive statistics: *summarize*

```
. bysort foreign: sum price mpg rep78

-> foreign = Domestic

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         52    6072.423    3097.104       3291      15906
         mpg |         52    19.82692    4.743297         12         34
       rep78 |         48    3.020833     .837666          1          5

-> foreign = Foreign

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         22    6384.682    2621.915       3748      12990
         mpg |         22    24.77273    6.611187         14         41
       rep78 |         21    4.285714    .7171372          3          5
```

Descriptive statistics: one-way tabulate

```
tabulate headroom
```

Headroom (in.)	Freq.	Percent	Cum.
1.5	4	5.41	5.41
2.0	13	17.57	22.97
2.5	14	18.92	41.89
3.0	13	17.57	59.46
3.5	15	20.27	79.73
4.0	10	13.51	93.24
4.5	4	5.41	98.65
5.0	1	1.35	100.00
Total	74	100.00	

Descriptive statistics: two-way tabulate

```
. tabulate headroom foreign

Headroom |       Car origin
   (in.) | Domestic    Foreign |     Total
---------+------------------------------------
     1.5 |        3          1 |         4
     2.0 |       10          3 |        13
     2.5 |        4         10 |        14
     3.0 |        7          6 |        13
     3.5 |       13          2 |        15
     4.0 |       10          0 |        10
     4.5 |        4          0 |         4
     5.0 |        1          0 |         1
---------+------------------------------------
   Total |       52         22 |        74
```

Descriptive statistics: tabulate for group statistics

```
. tabulate headroom, summarize(price)
```

Headroom (in.)	Summary of Price Mean	Std. dev.	Freq.
1.5	5,509.5	1,009.613	4
2.0	4,822.846	854.11424	13
2.5	6,591.571	3,196.544	14
3.0	5,906.231	3,403.979	13
3.5	7,580.933	4,086.414	15
4.0	6,458.5	2,256.714	10
4.5	5,018	926.56606	4
5.0	4,060	0	1
Total	6,165.257	2,949.496	74

Descriptive statistics: tabstat

```
tabstat price mpg rep78
```

Stats	price	mpg	rep78
Mean	6165.257	21.2973	3.405797

Descriptive statistics: tabstat, statistics(*)

```
. tabstat price mpg rep78, statistics(mean sd min max count)

    Stats |     price       mpg     rep78
----------+------------------------------
     Mean |  6165.257   21.2973  3.405797
       SD |  2949.496  5.785503  .9899323
      Min |      3291        12         1
      Max |     15906        41         5
        N |        74        74        69
----------------------------------------
```

Descriptive statistics: tabstat, by groups

```
tabstat price mpg rep78, statistics(mean sd min max count) by(foreign)

Summary statistics: Mean, SD, Min, Max, N
Group variable: foreign (Car origin)

 foreign |     price       mpg     rep78
---------+----------------------------------
Domestic |  6072.423  19.82692  3.020833
         |  3097.104  4.743297   .837666
         |      3291        12         1
         |     15906        34         5
         |        52        52        48
---------+----------------------------------
 Foreign |  6384.682  24.77273  4.285714
         |  2621.915  6.611187  .7171372
         |      3748        14         3
         |     12990        41         5
         |        22        22        21
---------+----------------------------------
   Total |  6165.257   21.2973  3.405797
         |  2949.496  5.785503  .9899323
         |      3291        12         1
         |     15906        41         5
         |        74        74        69
```

Access to results

- Stata temporarily stores results after analysis (e.g. after summarize)
- Allows using them for subsequent calculations
- Results are usually stored in **r-class** or **e-class** commands
- r-class: results from commands that do not estimate parameters (e.g. summarize)
- e-class: results from commands that estimate parameters (e.g. regress)
- After command, stored results can be checked by typing return list or ereturn list
- Stored results can be accessed by typing r(*) or e(*)

Access to results

```
. sum price

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906

. return list

scalars:
                  r(N) =  74
              r(sum_w) =  74
               r(mean) =  6165.256756756757
                r(Var) =  8699525.974268788
                 r(sd) =  2949.495884768919
               r(min) =  3291
               r(max) =  15906
               r(sum) =  456229

. display r(mean)
6165.2568
```

Access to results

- tabulate or tabstat do not automatically store results in **r-class**
- For tabulate: option matcell(matrixname) used to store results in a matrix format
- For tabstat: option save used to store results in a matrix r(StatTotal)

Exercise 1 I

1. Open a new do-file for this session and set working directory (to the folder for this session).
2. Import data file ”hh.xls”.
3. Get an overview of data using commends like describe, codebook, or count.
4. Count number of observations with missing values of variable pccd.
5. Generate descriptive statistics, using summarize,for the following variables: pccd, hhsize, nchild, nmigrant, emplrat, hhh female, land, urban
6. Generate descriptive statistics for pccd, hhsize, nchild, nmigrant, emplrat, hhh female by residential location of household (i.e. rural/urban) using urban variable. Try using both summarize and tabstat command.

Exercise 1 II

7. One-way tabulations: generate frequency distribution of observations by education of HH head using variable hhh educ and tabulate (shortly tab) command.
8. Two-way tabulations: generate frequency distribution of observations by (i) gender of HH head, and (ii) education of HH head using variables hhh female and hhh educ.
9. Check 'detailed' descriptive statistics using summarize ..., detail for variable pccd (per capita monthly consumption).
10. Generate standardized version of pccd variable. To create standardized variable, you need to subtract the mean from actual values and divide by standard deviation.
11. Save the data as "hh.dta" (to be used later again).

Collapsing Data

collapse converts data in memory to data with its descriptive statistics

1. Short for generating means: collapse varlist
2. Specifying statistics type: collapse [(stat)] varlist1 [[(stat)] varlist2]
3. Specifying target variable name: collapse [(stat)] targetvar =varname
4. By groups: collapse varlist, by(groupvar)

Collapsing data: data before

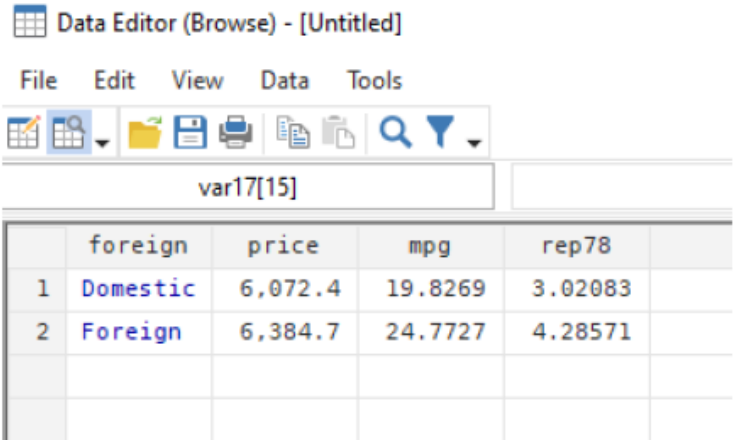
Data Editor (Browse) - [auto]

File Edit View Data Tools

var25[15]

	price	mpg	rep78	foreign
1	4,099	22	3	Domestic
2	4,749	17	3	Domestic
3	3,799	22	.	Domestic
4	4,816	20	3	Domestic
5	7,827	15	4	Domestic
6	5,788	18	3	Domestic
7	4,453	26	.	Domestic
8	5,189	20	3	Domestic
9	10,372	16	3	Domestic
10	4,082	19	3	Domestic

Collapsing data: data after



	foreign	price	mpg	rep78
1	Domestic	6,072.4	19.8269	3.02083
2	Foreign	6,384.7	24.7727	4.28571

Exporting descriptive statistics

Key commands:

- estout/esttab: formatted tables exported to MS Word, comma/tab delimited file (CSV), text file or Latex file
- outreg2: a faster/easier way to export (regression and summary) tables to MS Word, Excel or Latex files
- export excel: to use after collapse

Exporting descriptive statistics: estout/esttab

- estout and esttab exports estimation results to specified file format
- Requires results stored in e-class, hence after estimation commands
- Results from descriptive statistics can be stored in e-class using command estpost
- Syntax: estpost subcommand where *subcommand* can be summarize, tabstat, tabulate etc. (check help menu of estpost for full list of subcommands)
- Steps:
  1. Run estpost subcommand
  2. Results from are stored in e-class (check by typing ereturn list)
  3. Use estout or esttab to export results

Exporting descriptive statistics: estout/esttab

```
. estpost summarize price mpg rep78

             |  e(count)  e(sum_w)   e(mean)     e(Var)      e(sd)     e(min)     e(max)     e(sum)
-------------+---------------------------------------------------------------------------------------
       price |       74         74   6165.257   8699526   2949.496       3291      15906     456229
         mpg |       74         74    21.2973   33.47205   5.785503         12         41       1576
       rep78 |       69         69   3.405797   .9799659   .9899323          1          5        235

. esttab, cells("count mean sd min max") nomtitle nonumber noobs

------------------------------------------------------------------------
                      count       mean         sd        min        max
------------------------------------------------------------------------
price                    74   6165.257   2949.496       3291      15906
mpg                      74    21.2973   5.785503         12         41
rep78                    69   3.405797   .9899323          1          5
------------------------------------------------------------------------

. estout, cells("count mean sd min max") mlabels(,none)

------------------------------------------------------------------------
                      count       mean         sd        min        max
------------------------------------------------------------------------
price                    74   6165.257   2949.496       3291      15906
mpg                      74    21.2973   5.785503         12         41
rep78                    69   3.405797   .9899323          1          5
------------------------------------------------------------------------
```

Exporting descriptive statistics: estout/esttab

An example with statistics by groups

```
estpost tabstat price mpg rep78, by(foreign) statistics(mean sd) columns(statistics)

ummary statistics: mean sd
    for variables: price mpg rep78
 by categories of: foreign

    foreign |    e(mean)      e(sd)
------------+----------------------
omestic     |
      price |   6072.423   3097.104
        mpg |   19.82692   4.743297
      rep78 |   3.020833    .837666
------------+----------------------
oreign      |
      price |   6384.682   2621.915
        mpg |   24.77273   6.611187
      rep78 |   4.285714   .7171372
------------+----------------------
otal        |
      price |   6165.257   2949.496
        mpg |    21.2973   5.785503
      rep78 |   3.405797   .9899323

 esttab, main(mean) aux(sd) unstack noobs nonote nomtitle nonumber label

------------------------------------------------------------------------
                                 Domestic        Foreign          Total
------------------------------------------------------------------------
rice                               6072.4         6384.7         6165.3
                                  (3097.1)       (2621.9)       (2949.5)

Mileage (mpg)                       19.83          24.77          21.30
                                   (4.743)        (6.611)        (5.786)

epair record 1978                   3.021          4.286          3.406
                                   (0.838)        (0.727)        (0.990)
------------------------------------------------------------------------
```

Exporting descriptive statistics: estout/esttab

- Exporting as CSV file: esttab using filename.csv, ...
- Exporting as Word file: esttab using filename.rtf, ...
- Exporting as TEX file: esttab using filename.tex, ...

More examples of summary tables with estout/esttab at:http://repec.org/bocode/e/estout/estpost.html

Exporting descriptive statistics: outreg2

- esttab/estout is a powerful, but also a complicated tool to export tables
- Another drawback: sometimes exported CSV file may not open correctly in Excel
- An easy-to-use alternative is outreg2
- Exports regression/summary tables to Word, Excel, Latex format
- Options are straightforward and easier to use (compared esttab)

Exporting descriptive statistics: outreg2

outreg2 using sumtable, keep(price mpg rep78) sum(log) replace label excel

I12

	A	B	C	D	E	F	G
1							
2		(1)	(2)	(3)	(4)	(5)	
3	VARIABLES	N	mean	sd	min	max	
4							
5	Price	74	6,165	2,949	3,291	15,906	
6	Mileage (mpg)	74	21.30	5.786	12	41	
7	Repair record 1978	69	3.406	0.990	1	5	
8							
9							
10							

sum(log) gives basic table as in summarize. sum(detail) gives detailed descriptive statistics as in summarize,detail. In both cases, you can control which statistics you want to keep using optionseqdrop() or eqkeep(), default is all statistics from summarize or summarize, detail. **keep** is to select variables for which you need statistics, default is all variables

Exercise 2

1. Open the data "hh.dta" from Exercise 3.1.
2. Label variables pccd, hhsize, nmigrant and emplrat.
3. Export descriptive statistics for these variable using outreg2 to an Excel file. You should keep number of observations, mean, median, standard deviation, minimum and maximum for each variable. Make sure that variables are labelled in the exported table.
4. Collapsing data: now collapse your data by urban variable, only with variables listed above. The resulting data should have mean and standard deviations (as columns/variables) for these variables for urban and rural locations (as rows). Export the collapsed data to an Excel file.