STATA - Examine Data

Prof. Tzu-Ting Yang

Institute of Economics, Academia Sinica

March 14, 2022

Examine Data

- Please see **C3_examine_data.do**

Examine Data

Consolas

```
1 use "$rawdata\cps_2014_16.dta",replace
```

- It is a good idea to examine your data when you first read it into Stata
- You should check that all the variables and observations are present and in the correct format

browse/edit: See the contents of a data file

STATA command: browse/edit

See the contents of a data file

Consolas

```
1 browse
2
3 edit
```

- You can see the contents of a data file using the **browse** or **edit** command
- You can also type **Ctrl+8** to call **Data Editor**

describe: Producing a summary of the dataset

STATA command: describe

Producing a summary of the dataset

Consolas

```
1 describe using "$rawdata\cps_2014_16.dta"
```

- **describe**: produces a summary of the dataset in memory or of the data stored in a Stata-format dataset

STATA command: describe

Producing a summary of the dataset

```
. describe using "$rawdata\cps_2014_16.dta"

Contains data
  obs:        60,000                          29 Mar 2020 16:29
 vars:            23
```

variable name	storage type	display format	value label	variable label
year	int	%8.0g		**survey year**
serial	long	%12.0g		**household serial number**
hwtsupp	double	%12.0g		**household weight, supplement**
cpsid	double	%12.0g		**cpsid, household record**
region	byte	%8.0g	REGION	**region and division**
statefip	byte	%8.0g	STATEFIP	**state (fips code)**
asecflag	byte	%8.0g	ASECFLAG	**flag for asec**
hflag	byte	%8.0g	HFLAG	**flag for the 3/8 file 2014**
county	long	%12.0g		**fips county code**
month	byte	%8.0g	MONTH	**month**
pernum	byte	%8.0g		**person number in sample unit**
cpsidp	double	%12.0g		**cpsid, person record**
wtsupp	double	%12.0g		**supplement weight**
age	byte	%8.0g	AGE	**age**
sex	byte	%8.0g	gendercode	**sex**
race	int	%8.0g	RACE	**race**
marst	byte	%8.0g	MARST	**marital status**

list: Show entire data within the results window

STATA command: list

Show entire data within the results window

Consolas

```
1 list serial year sex age
2 list serial year sex age in 55/60
```

- **list**: show entire data within the results window
- Although listing the entire dataset is only feasible if it is small
- Alternatively we could focus on all variables but list only a limited number of observations
  - Example: List the observation 55 to 60
  - The qualifier **in** ensures that commands apply only to a certain subset of the data

STATA command: list

Show entire data within the results window

```
. list serial year sex age in 55/60
```

	serial	year	sex	age
55.	47	2014	2	47
56.	47	2014	1	48
57.	48	2014	2	66
58.	48	2014	1	45
59.	48	2014	1	30
60.	48	2014	2	19

STATA command: list

Consolas

```
1 list serial year sex age if age==37
```

- A second way of subsetting the data is the **if** qualifier (more on this later on)
- The qualifier is followed by an expression that evaluates either to “true” or “false” (i.e. 1 or 0)

assert: Verifies whether a certain statement is true or false

STATA command: assert

Verifies whether a certain statement is true or false

Consolas

```
1 assert sex<3
2 assert sex>3
3 assert age>85
```

- With large datasets, it often is impossible to check every single observation using **list**
- **assert**: verifies whether a certain statement is true or false
- If the statement is true, **assert** does not yield any output on the screen.
- If it is false, **assert** gives an error message and the number of contradictions

STATA command: assert

Verifies whether a certain statement is true or false

```
. assert age>85
60,000 contradictions in 60,000 observations
assertion is false
```

codebook: Provides extra information on the variables

STATA command: codebook

Provides extra information on the variables

Consolas

```
1 codebook age sex incwage
2 codebook
```

- **codebook**: provides extra information on the variables, such as summary statistics of numerics, example data-points of strings, and so on
- **codebook** without a list of variables will give information on all variables in the dataset

STATA command: codebook

Provides extra information on the variables

```
. codebook age sex incwage
```

---

age

```
                  type:  numeric (byte)
                 label:  AGE, but 81 nonmissing values are not labeled

                 range:  [0,85]                       units:  1
         unique values:  82                       missing .:  0/60,000

              examples:  13
                         28
                         43
                         57
```

sum: Provides summary statistics

STATA command: sum

Provides summary statistics

Consolas

```
1 sum incwage
2 sum incwage,d
3 sum
```

- **summarize**: summary statistics, such as means, standard deviations, and so on
- Additional information about the distribution of the variable can be obtained using the **detail** option

STATA command: codebook

Provides summary statistics

```
. sum incwage,d

                    wage and salary income
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0            0
 5%            0            0
10%            0            0       Obs                  60,000
25%            0            0       Sum of Wgt.          60,000

50%        30000                    Mean                 2306039
                       Largest       Std. Dev.            4185713
75%       125000      9999999
90%      9999999      9999999       Variance             1.75e+13
95%      9999999      9999999       Skewness             1.293873
99%      9999999      9999999       Kurtosis             2.674723
```

tab: Produce a frequency table

STATA command: tab

Produce a frequency table

Consolas

```
1 tab age
2 tab age sex
```

- **tab**: Produce a frequency table of one variable or a cross-tab of two variables

STATA command: tab

Produce a frequency table

```
. tab age sex
```

age	sex 1	2	Total
under 1 year	**348**	**329**	**677**
1	**416**	**389**	**805**
2	**449**	**401**	**850**
3	**470**	**424**	**894**
4	**460**	**422**	**882**
5	**482**	**470**	**952**
6	**469**	**471**	**940**
7	**506**	**473**	**979**
8	**462**	**455**	**917**
9	**465**	**502**	**967**
10	**460**	**455**	**915**

STATA command: tab

Consolas

```
1 tab age, sum(incwage)
2 tab sex, sum(incwage)
```

- We can use the **tab** command combined with the **sum(varname)** option to gain a quick idea of the descriptive statistics of certain subgroups
- For example, we can know average wage by age (gender)

STATA command: tab

Produce a frequency table

```
. tab sex, sum(incwage)
```

sex	Summary of wage and salary income Mean	Std. Dev.	Freq.
1	**2421356.5**	**4251034**	**29,148**
2	**2197090.2**	**4120154.4**	**30,852**
Total	**2306038.8**	**4185713.2**	**60,000**

inspect: Show distribution of a variable

STATA command: inspect

Show distribution of a variable

Consolas

```
1 inspect age
2 inspect incwage
```

- **inspect**: This is a way to eyeball the distribution of a variable, including as it does a mini-histogram.
- It is also useful for **identifying outliers or unusual values**, or for spotting non-integers in a variable that should only contain integers

STATA command: inspect

Show distribution of a variable

```
. inspect age

age:  age                                       Number of Observations
-----------------------------------------------------------------------
                                                    Total      Integers   Nonintegers
  #                                 Negative            -             -             -
  #       #                         Zero              677           677             -
  #   #   #   #                     Positive        59,323        59,323             -
  #   #   #   #                     -----------------------------------------------
  #   #   #   #   #                 Total           60,000        60,000             -
  #   #   #   #   #                 Missing              -
-----------------------------------------------------------------------
0                     85                            60,000
  (82 unique values)

      age is labeled and all values are documented in the label.
```

isid: Check uniquely identify

STATA command: isid/duplicates

- Sometimes we investigate whether our dataset has exactly the same observation or the observations with the same value of variable
- We can use command **isid** and **duplicates** to explore this issue

STATA command: isid

Check uniquely identify

Consolas

```
1 isid serial
```

- **isid**: Check whether the specified variables (e.g. “serial”) uniquely identify the observations

STATA command: isid

Check uniquely identify

```
. isid serial
variable serial does not uniquely identify the observations
r(459);
```

duplicates: Detecting repeated observations

STATA command: duplicates

Detecting repeated observations

Consolas

```
1 duplicates report serial
2 duplicates report
```

- **duplicates**: Report, tag, or drop duplicate observations
- **Line 1:** Use option **report** to show the number of observation with the same "serial"
- **Line 2:** Use option **report** to show the number of observation with the same characteristics (value of all variables)

STATA command: duplicates

Detecting repeated observations

```
. duplicates report serial

Duplicates in terms of serial
```

copies	observations	surplus
1	**5555**	**0**
2	**13892**	**6946**
3	**11427**	**7618**
4	**14760**	**11070**
5	**8350**	**6680**
6	**3414**	**2845**
7	**1498**	**1284**
8	**536**	**469**
9	**324**	**288**
10	**80**	**72**
11	**66**	**60**
12	**72**	**66**
13	**26**	**24**

STATA command: duplicates

Detecting repeated observations

```
. duplicates report

Duplicates in terms of all variables
```

copies	observations	surplus
1	**60000**	**0**

STATA command: duplicates

Detecting repeated observations

Consolas

```
1 duplicates example serial age year in 1/100
2 duplicates list serial age year in 1/100
3 duplicates tag serial,gen(s_repeat)
4 tab s_repeat
5 duplicates drop serial,force
```

- **Line 1:** Use option **example** to list one example for each group of duplicates
  - in 1/100 means the first 100 observations
- **Line 2:** Use option **list** to show all repeated observations
- **Line 3:** Use option **tag** to generate a variable to indicate how many duplicates for this observation
- **Line 5:** Use option **drop** to drop observation with the same "serial"

STATA command: duplicates

Detecting repeated observations

```
. duplicates example serial age year in 1/100
```

Duplicates in terms of **serial age year**

group:	#	e.g. obs	serial	age	year
1	2	17	16	39	2014
2	2	25	22	85	2014
3	2	34	27	26	2014
4	2	62	50	32	2014
5	2	67	54	39	2014
6	2	77	55	4	2014
7	2	73	55	29	2014
8	2	97	61	53	2014

STATA command: duplicates

Detecting repeated observations

```
. duplicates list serial age year in 1/100
```

Duplicates in terms of **serial age year**

group:	obs:	serial	age	year
1	17	16	39	2014
1	19	16	39	2014
2	25	22	85	2014
2	26	22	85	2014
3	34	27	26	2014
3	35	27	26	2014
4	62	50	32	2014
4	63	50	32	2014
5	67	54	39	2014
5	70	54	39	2014
6	77	55	4	2014
6	78	55	4	2014
7	73	55	29	2014
7	74	55	29	2014
8	97	61	53	2014
8	98	61	53	2014

STATA command: duplicates

Detecting repeated observations

	serial	s_repeat	pernum
1	1	4	1
2	1	4	2
3	1	4	3
4	1	4	4
5	1	4	5
6	2	2	1
7	2	2	2
8	2	2	3