

Handling Econometric data | R programming language

Preliminaries

Script Editor

```
488 # setup
489 rm(list = ls()) # clear the work space
490
491 # Install packages
492
493 install.packages("tidyverse")
494 install.packages("stargazer")
495 install.packages("magrittr")
496 install.packages("moments")
497
498 # Load packages
499
500 library(tidyverse)
501 library(magrittr)
502 library(stargazer)
503 library(moments)
```

Refer previous notes for assessing reflections in the console and environment panel

Cross-sectional data: Loading the dataset and summarizing data

- Cross-sectional data set on wages and other characteristics
- Observations are individuals

Script Editor

```
wage <- read.csv("C:/Users/ /OneDrive/Desktop/wage1.csv")
stargazer(wage, type = "text")
```

Environment

■ wage data.fra... 24 56.9 k 526 obs. of 24 var...

Console

```
> wage <- read.csv("C:/Users/ /OneDrive/Desktop/wage1.csv")
> stargazer(wage, type = "text")
```

Statistic	N	Mean	St. Dev.	Min	Max
wage	526	5.896	3.693	0.530	24.980
educ	526	12.563	2.769	0	18
exper	526	17.017	13.572	1	51
tenure	526	5.105	7.224	0	44
nonwhite	526	0.103	0.304	0	1
female	526	0.479	0.500	0	1
married	526	0.608	0.489	0	1
numdep	526	1.044	1.262	0	6
smsa	526	0.722	0.448	0	1
northcen	526	0.251	0.434	0	1
south	526	0.356	0.479	0	1
west	526	0.169	0.375	0	1
construc	526	0.046	0.209	0	1
ndurman	526	0.114	0.318	0	1
trcommpu	526	0.044	0.205	0	1
trade	526	0.287	0.453	0	1
services	526	0.101	0.301	0	1
profserv	526	0.259	0.438	0	1
profocc	526	0.367	0.482	0	1
clerocc	526	0.167	0.374	0	1
servocc	526	0.141	0.348	0	1
lwage	526	1.623	0.532	-0.635	3.218
expersq	526	473.435	616.045	1	2,601
tenursq	526	78.150	199.435	0	1,936

Handling Econometric data | R programming language

Cross-sectional data: Summarize the data for selected variables

Script Editor

```
wage %>%  
  select(wage, lwage, educ, exper, expersq, tenure, married, female) %>%  
  stargazer(type = "text")
```

This command is same as saying “take the dataset named wage and then select the variables wage, lwage, educ, exper, expersq, tenure, married, female and then arrange the same neatly in the form of a table using stargazer function.

The pipe operator helps in doing the abovementioned work using a single line of code instead of using multiple lines of codes as shown below:

```
a<- select(wage, lwage, educ, exper, expersq, tenure, married, female, wage=wage)  
stargazer(a, type = "text")
```

Console

```
> wage %>%  
+   select(wage, lwage, educ, exper, expersq, tenure, married, female) %>%  
+   stargazer(type = "text")
```

Statistic	N	Mean	St. Dev.	Min	Max
wage	526	5.896	3.693	0.530	24.980
lwage	526	1.623	0.532	-0.635	3.218
educ	526	12.563	2.769	0	18
exper	526	17.017	13.572	1	51
expersq	526	473.435	616.045	1	2,601
tenure	526	5.105	7.224	0	44
married	526	0.608	0.489	0	1
female	526	0.479	0.500	0	1

Cross-sectional data: Keep only selected variables in the dataset

Script Editor

'lhs %<>% rhs' is equivalent to 'lhs <- lhs %>% rhs'.

```
wage %<>% select(wage, lwage, educ, exper, expersq, tenure, married, female)
```

Environment

☐	wage	data.frame	24	56.9 KB	526 obs. of 24 vari...	
--------------------------	------	------------	----	---------	------------------------	--

Keeping only select variables in the dataset resulted in the reduction of total variables from 24 to 8 as reflected in the environment panel

☐	Name	Type	Length	Size	Value
☐	wage	data.frame	8	22.1 KB	526 obs. of 8 vari...

Console

```
> wage %<>% select(wage, lwage, educ, exper, expersq, tenure, married, female)
```

For more details on pipe operator (%>%) and pipe assignment operator (%<>%), refer previous notes

Handling Econometric data| R programming language

Cross-sectional data: Summarizing data

Script Editor

```
565 str(wage) <- this command gives structure of variables of the dataset named wage
566 stargazer(wage, type = "text") <- this command provides summary statistics
567 head(wage, 10) <- this command provides first 10 observations of the dataset wage
```

Console

```
> str(wage)
'data.frame': 526 obs. of 8 variables:
 $ wage : num 3.1 3.24 3 6 5.3 ...
 $ lwage : num 1.13 1.18 1.1 1.79 1.67 ...
 $ educ : int 11 12 11 8 12 16 18 12 12 17 ...
 $ exper : int 2 22 2 44 7 9 15 5 26 22 ...
 $ expersq: int 4 484 4 1936 49 81 225 25 676 484 ...
 $ tenure : int 0 2 0 28 2 8 7 3 4 21 ...
 $ married: int 0 1 0 1 1 1 0 0 0 1 ...
 $ female : int 1 1 0 0 0 0 0 1 1 0 ...
> stargazer(wage, type = "text")
```

```
=====
Statistic N Mean St. Dev. Min Max
-----
wage 526 5.896 3.693 0.530 24.980
lwage 526 1.623 0.532 -0.635 3.218
educ 526 12.563 2.769 0 18
exper 526 17.017 13.572 1 51
expersq 526 473.435 616.045 1 2,601
tenure 526 5.105 7.224 0 44
married 526 0.608 0.489 0 1
female 526 0.479 0.500 0 1
```

```
> head(wage, 10)
  wage lwage educ exper expersq tenure married female
1  3.10 1.131402  11     2      4      0      0      1
2  3.24 1.175573  12    22     484     2      1      1
3  3.00 1.098612  11     2      4      0      0      0
4  6.00 1.791759   8    44    1936    28      1      0
5  5.30 1.667707  12     7     49     2      1      0
6  8.75 2.169054  16     9     81     8      1      0
7 11.25 2.420368  18    15    225     7      0      0
8  5.00 1.609438  12     5     25     3      0      1
9  3.60 1.280934  12    26    676     4      0      1
10 18.18 2.900322  17    22    484    21      1      0
```

Cross-sectional data: Generating frequency tables

Script Editor

```
583 table(wage$female)
```

Console

```
> table(wage$female)
 0  1
274 252
```

There are 274 males (female=0) and 252 females(female=1) in the dataset

Note: In R, the \$ sign is used to access a specific column or element of a data frame or list. It allows you to reference a single variable within a data frame or list by name.

Handling Econometric data | R programming language

Time series data

Script Editor

```
595 prminwge <- read.csv("C:/Users/amalz/OneDrive/Desktop/prminwge.csv")
596 prminwge %<>% select(year, avgmin, avgcov, prunemp, prgnp)
597 str(prminwge)
598 stargazer(prminwge, type = "text")
599 table(prminwge$year)
```

Environment

R Global Environment					
Name	Type	Length	Size	Value	
prminwge	data.fr...	5	2.5 KB	38 obs. of 5 var...	
wage	data.fr...	8	22.1 ...	526 obs. of 8 va...	

Input and its interpretations are same as that of cross sectional data (except for the nature of the freq. table as output, since the variable year is an integer variable unlike the variable female of the previous dataset which was categorical).

Console

```
> prminwge %<>% select(year, avgmin, avgcov, prunemp, prgnp)
> str(prminwge)
'data.frame': 38 obs. of 5 variables:
 $ year : int 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 ...
 $ avgmin : num 0.198 0.209 0.225 0.311 0.313 ...
 $ avgcov : num 0.201 0.207 0.226 0.231 0.224 ...
 $ prunemp : num 15.4 16 14.8 14.5 15.3 ...
 $ prgnp : num 879 925 1016 1081 1104 ...
> stargazer(prminwge, type = "text")
```

Statistic	N	Mean	St. Dev.	Min	Max
year	38	1,968.500	11.113	1,950	1,987
avgmin	38	1.484	1.130	0.198	3.350
avgcov	38	0.416	0.153	0.201	0.596
prunemp	38	15.039	3.748	10.300	23.400
prgnp	38	2,604.411	1,210.735	878.700	4,496.700

```
> table(prminwge$year)
```

1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987						
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1						

Panel data

Panel data for wages of individuals across years

8-year panel data on wages of individuals and other variables

there are observations for the same individual for multiple years

person identifier is nr

time identifier is year

Input and its interpretations are same as that of cross sectional data (except for the nature of the freq. table as output, since the variable year is an integer variable unlike the variable female of the cross sectional dataset which was categorical. Moreover, the variable year differs between time series and panel data (hence the difference in the output of the same).

- Cross-sectional data: Each year appears only once with a frequency of 1.
- Panel data: Each year can appear multiple times, with the frequency equal to the number of entities in the panel dataset. (here 525)

Script Editor

```
648 wagepan <- read.csv("C:/Users/ /OneDrive/Desktop/wagepan.csv")
649 wagepan %<>% select(nr, year, lwage, exper, educ, hours)
650 str(wagepan)
651 stargazer(wagepan, type = "text")
652 head(wagepan, 10)
653 table(wagepan$year)
```

Handling Econometric data | R programming language

Console

```
> str(wagepan)
'data.frame':  4360 obs. of  6 variables:
 $ nr    : int  13 13 13 13 13 13 13 13 17 17 ...
 $ year  : int  1980 1981 1982 1983 1984 1985 1986 1987 1980 1981 ...
 $ lwage : num  1.2 1.85 1.34 1.43 1.57 ...
 $ exper : int  1 2 3 4 5 6 7 8 4 5 ...
 $ educ  : int  14 14 14 14 14 14 14 14 13 13 ...
 $ hours : int  2672 2320 2940 2960 3071 2864 2994 2640 2484 2804 ...
> stargazer(wagepan, type = "text")
```

```
=====
Statistic    N      Mean    St. Dev.    Min     Max
-----
nr           4,360  5,262.059  3,496.150    13   12,548
year         4,360  1,983.500    2.292   1,980   1,987
lwage        4,360    1.649    0.533   -3.579   4.052
exper        4,360    6.515    2.826     0     18
educ         4,360   11.767    1.746     3     16
hours        4,360  2,191.257   566.352   120   4,992
-----
```

```
> head(wagepan, 10)
  nr year   lwage exper educ hours
1  13 1980  1.1975402     1   14  2672
2  13 1981  1.8530600     2   14  2320
3  13 1982  1.3444617     3   14  2940
4  13 1983  1.4332134     4   14  2960
5  13 1984  1.5681251     5   14  3071
6  13 1985  1.6998910     6   14  2864
7  13 1986 -0.7202626     7   14  2994
8  13 1987  1.6691879     8   14  2640
9  17 1980  1.6759624     4   13  2484
10 17 1981  1.5183982     5   13  2804
```

```
> table(wagepan$year)

1980 1981 1982 1983 1984 1985 1986 1987
 545   545   545   545   545   545   545   545
```

Pooled cross sections

- # Pooled cross sections on housing prices
- # House prices for two years, 1978 and 1981
- # Houses are different in each year.
- # We do not have the price of the same house in both years.

Script Editor

```
hprice3 <- read.csv("C:/Users/amalz/OneDrive/Desktop/hprice3.csv")
hprice3 %<>% select(year, y81, price, lprice, rooms, baths)
str(hprice3)
stargazer(hprice3, type = "text")
head(hprice3, 10)
table(hprice3$year)
```

Handling Econometric data | R programming language

Console

```
> str(hprice3)
'data.frame': 321 obs. of 6 variables:
 $ year : int 1978 1978 1978 1978 1978 1978 1978 1978 1978 ...
 $ y81 : int 0 0 0 0 0 0 0 0 0 ...
 $ price : int 60000 40000 34000 63900 44000 46000 56000 38500 60500 55000 ...
 $ lprice: num 11 10.6 10.4 11.1 10.7 ...
 $ rooms : int 7 6 6 5 5 6 6 6 5 ...
 $ baths : int 1 2 1 1 1 3 2 2 2 2 ...
> stargazer(hprice3, type = "text")
```

```
=====
Statistic N      Mean      St. Dev.    Min      Max
-----
year      321 1,979.327      1.492    1,978    1,981
y81       321   0.442      0.497      0      1
price     321 96,100.660 43,223.730 26,000 300,000
lprice    321   11.378      0.438   10.166 12.612
rooms     321   6.586      0.901      4     10
baths     321   2.340      0.771      1      4
-----
```

```
> head(hprice3, 10)
  year y81 price  lprice rooms baths
1 1978  0 60000 11.00210     7     1
2 1978  0 40000 10.59663     6     2
3 1978  0 34000 10.43412     6     1
4 1978  0 63900 11.06507     5     1
5 1978  0 44000 10.69195     5     1
6 1978  0 46000 10.73640     6     3
7 1978  0 56000 10.93311     6     2
8 1978  0 38500 10.55841     6     2
9 1978  0 60500 11.01040     8     2
10 1978  0 55000 10.91509     5     2
```

```
> table(hprice3$year)
1978 1981
179 142
```

Pooled cross sections: Summary statistics for price and year

Script Editor

```
3 hprice3 %>%
4   select(price, year) %>%
5   stargazer(type = "text")
```

The command means:- Take the hprice3 dataset, then select only the price and year columns, and then use stargazer to create a nice summary table with the output in text format.

Console

```
=====
Statistic N      Mean      St. Dev.    Min      Max
-----
price     321 96,100.660 43,223.730 26,000 300,000
year      321 1,979.327      1.492    1,978    1,981
-----
```

Pooled cross sections: Summary statistics for selected samples

Script Editor

```
# Before is year 1978, see average price before
hprice3 %>%
  filter(year == 1978) %>%
  select(price) %>%
  stargazer(type = "text")

# After is year 1981, see average price after
hprice3 %>%
  filter(year == 1981) %>%
  select(price) %>%
  stargazer(type = "text")
```

Handling Econometric data | R programming language

Console

```
> # Before is year 1978, see average price before
> hprice3 %>%
+   filter(year == 1978) %>%
+   select(price) %>%
+   stargazer(type = "text")

=====
Statistic N      Mean      St. Dev.   Min      Max
-----
price    179 76,628.040 30,626.440 26,000 300,000
-----

> # After is year 1981, see average price after
> hprice3 %>%
+   filter(year == 1981) %>%
+   select(price) %>%
+   stargazer(type = "text")

=====
Statistic N      Mean      St. Dev.   Min      Max
-----
price    142 120,647.100 44,359.890 41,000 270,000
-----
```



ECONIMIUM
ECONOMICS FOR ALL