Lecture 1: Stata Basics

Yi Chen

ShanghaiTech University

2021

Outline

Outline

Course Logistics

Goal: Equip students with basic knowledge about software, which can be useful throughout all stages in an economic research.

- Idea - Literature - Data - Analysis - Tables/Figures - Writing - Presenting

Requirements:

- Know how to use a software "correctly."
- Know how to use a software "efficiently."
  - Efficiency in programming time
  - Efficiency in running time
- Know how to use a software "nicely."
  - Nice to the readers
  - Nice to your advisor (if work as RA)
  - Nice to co-authors/researchers trying to replicate your research
  - Also to (future) yourself!

Even if you have no interest in academia ...

- Showing facts/writing/presenting are must-have skills for most technical jobs.
- Huge demand for skills in data analysis (data has been considered as the most valuable asset for many Internet companies).
- Jobs like data scientists/consulting welcome students with econ background.
  - e.g., Amazon is the second-largest employer of Econ PhD in the U.S. (behind Federal Reserve).

浏览器视频广告过滤功能是否构成不正当竞争？北京知识产权法院：构成

新华社新媒体
百家号 01-06 17:27

新华社北京1月6日电（记者熊琳）日前，腾讯公司与世界星辉公司不正当竞争纠纷一案尘埃落定。北京知识产权法院终审认定“世界之窗浏览器”过滤广告功能构成不正当竞争，判决世界星辉公司赔偿腾讯公司经济损失及合理支出189万余元。

腾讯公司一审诉称，“世界之窗浏览器”软件系世界星辉公司开发经营，该浏览器设置有广告过滤功能，用户可有效过滤“腾讯视频”网站在播放影片时的片头广告和暂停广告，使腾讯公司不能从该业务中获取直接收益。

一审法院认为，被诉行为不针对特定视频经营者，广告过滤功能属于行业惯例，网络用户对浏览器广告过滤功能的使用，虽造成广告被浏览次数的减少，但此种减少并不构成法律应予救济的“实际损害”。据此，一审法院驳回了腾讯公司全部诉讼请求。

腾讯公司提起上诉。二审中，腾讯公司提交了有关过滤广告功能对网络视频行业影响的经济学分析报告。

How to Learn this Course Well?

Did you have following experience?

I read several textbooks on a language. I know how each command works. But I just cannot finish a project on my own.

From bricks to castle

A separate command is like a brick. The entire project is like a castle. What's the missing link?

This course would put greater emphasizes on combining simple commands to achieve a certain goal.

My Expectations

- From passive learning to active learning
- From remembering to understanding
- From answering questions to asking questions

Outline

Features of Stata

Stata = statistics + data

Advantage:

- Language: intuitive and easy to learn.
- Do-file: for easy replication.
  - Imagine how you work with an Excel sheet.
- Designed for economists
  - Lots of powerful user-written commands.

Disadvantage:

- Need to put the data into the memory first. Not suitable for handling "huge" data.
- Cumbersome if you wish to build your own econometric model.

Q: how to determine which software to learn?

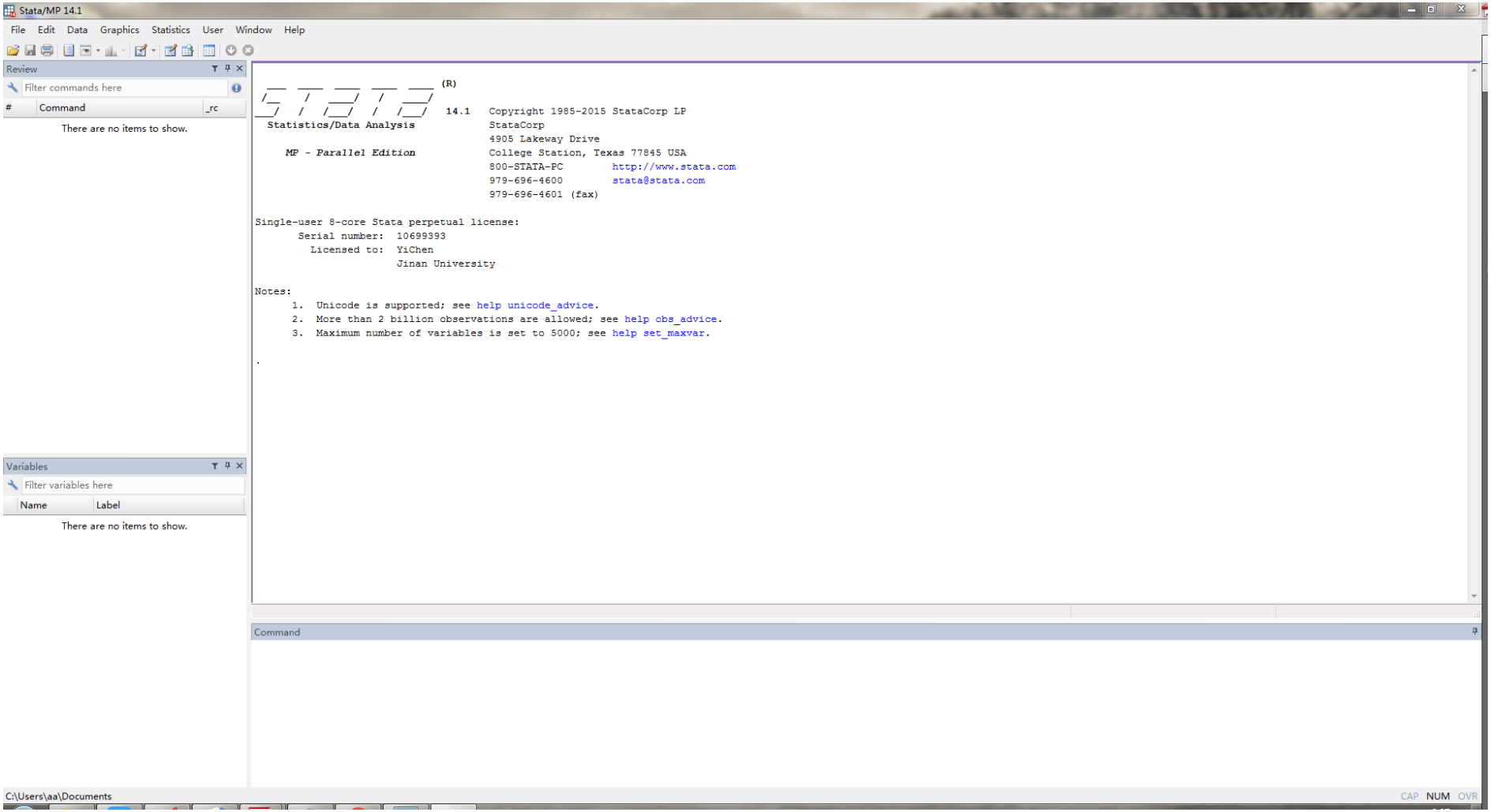
What You Think Is What You Get

Many students find it difficult to learn Stata at the beginning. Probably because of the following reason:

- Word/Excel/PowerPoint share a common feature—What You See Is What You Get
- Stata, along with many other softwares (Matlab, LaTeX, Python)—What You Think Is What You Get
- "Imagination" is very important! Once you get accustomed to this way of thinking, everything would be much easier!

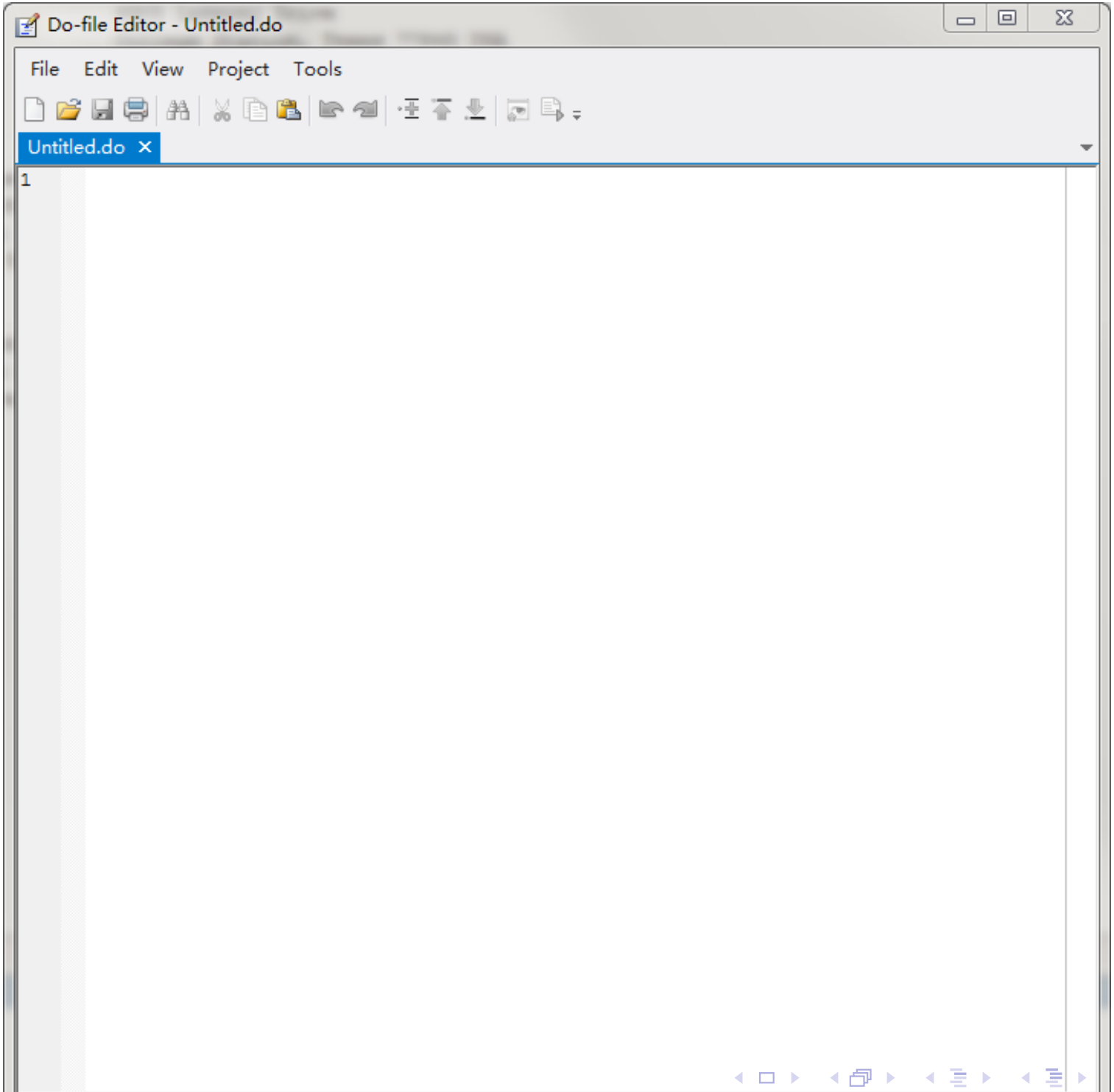
Goal of this Lecture

- After this lecture (Stata Basics), you should be able to—solve econometrics problem sets.
- After next lecture (Stata advanced), you should be able to—do REAL empirical analysis.


Stata/MP 14.1
File Edit Data Graphics Statistics User Window Help
Review
Filter commands here
Command _rc
There are no items to show.
Variables
Filter variables here
Name Label
There are no items to show.
(R)
Statistics/Data Analysis 14.1 Copyright 1985-2015 StataCorp LP
StataCorp
4905 Lakeway Drive
MP - Parallel Edition College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
Single-user 8-core Stata perpetual license:
Serial number: 10699393
Licensed to: YiChen
Jinan University
Notes:
1. Unicode is supported; see help unicode_advice.
2. More than 2 billion observations are allowed; see help obs_advice.
3. Maximum number of variables is set to 5000; see help set_maxvar.
.
Command
C:\Users\aa\Documents
CAP NUM OVR

Four Windows

- Result window—where you see the output
- Review window—where records the history of your command
- Variable window—where you can see the information of the variables after you load data
- Command window—where you type command
  - However, this is usually not the place where you type commands . . .


Do-file Editor - Untitled.do
File
Edit
View
Project
Tools
Untitled.do
1

First Lesson in Stata—ALWAYS Use DO-file

What's a do-file? Essentially, it's a TEXT file.

Why it's important? You need hundreds (even thousands) of operations to transit from raw data to final output.

- Recording
- Reminder
- Organization
- Replication

Comments in Do-file

A Do-file = a collection of commands + comments

- Comments are for annotation only and will not be executed by Stata.
  - But using comments properly is VERY important for a "nice" do-file!
- Three ways of commenting
  - begin the line with *
  - begin the comment with // (at the beginning or at the end)
    - if the // indicator is at the end of a line, it must be preceded by one or more blanks
  - place the comment between /* and */ delimiters
- Another use of comments: temporarily "save"

Stata General Syntax

All Stata commands can be expressed in follow syntax (or a subset)

[*prefix:*] command [*varlist*] [= *exp.*] [if *exp.*] [in *range*] [weight =] [using *filename*] [, *options*]

- `bysort` and `eststo` are the two mostly used prefix. For some reasons, `quietly` does not come with a ":"
- We will talk about weight in the Data Analysis part.

If Commands are Too Long . . .

change the end-of-line delimiter to ';' by using #delimit,

fragment of example.do

```
use mydata
#delimit ;
summarize weight price displ headroom rep78 length turn gear_ratio
          if substr(company,1,4)=="Ford" |
             substr(company,1,2)=="GM", detail ;
gen byte ford = substr(company,1,4)=="Ford" ;
#delimit cr
gen byte gm = substr(company,1,2)=="GM"
```

fragment of example.do

using /* */ comment brackets or to use the /// line-join indicator

fragment of example.do

```
use mydata
summarize weight price displ headroom rep78 length turn gear_ratio /*
      */  if substr(company,1,4)=="Ford" |    /*
      */     substr(company,1,2)=="GM", detail
gen byte ford = substr(company,1,4)=="Ford"
gen byte gm = substr(company,1,2)=="GM"
```

fragment of example.do

or

fragment of example.do

```
use mydata
summarize weight price displ headroom rep78 length turn gear_ratio ///
        if substr(company,1,4)=="Ford" |    ///
           substr(company,1,2)=="GM", detail
gen byte ford = substr(company,1,4)=="Ford"
gen byte gm = substr(company,1,2)=="GM"
```

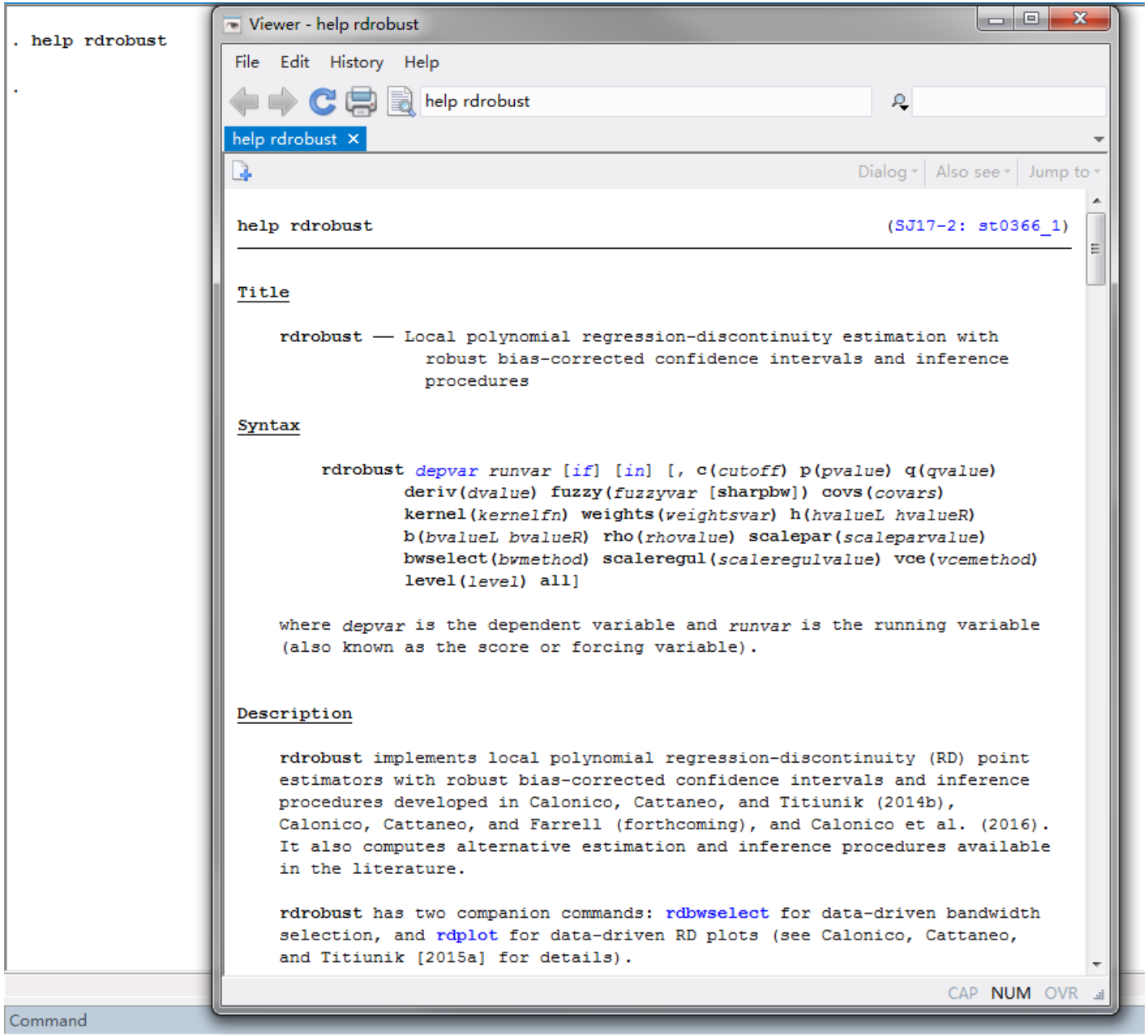
fragment of example.do

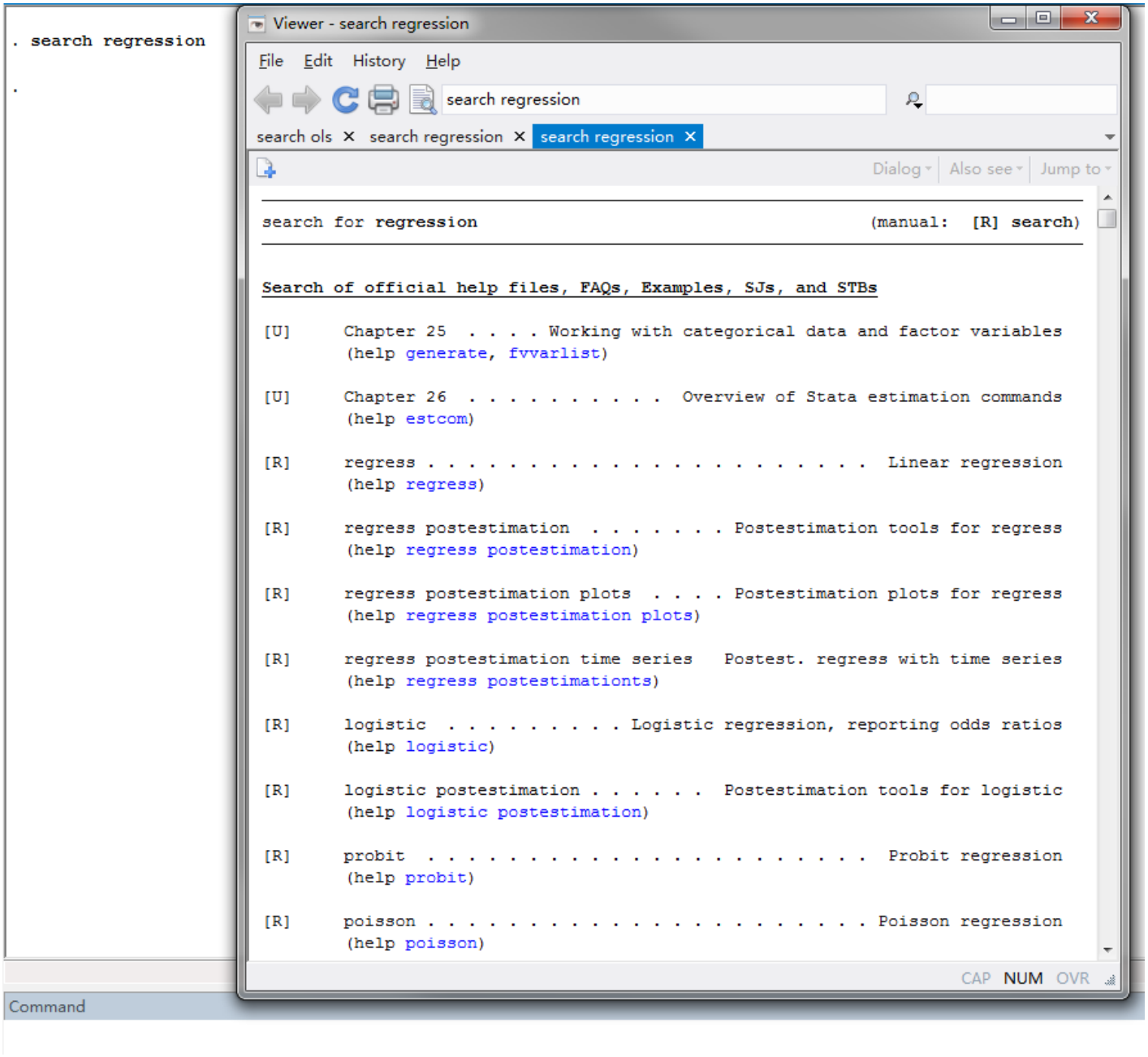
Several Useful Short-Cuts in the Do-file Editor

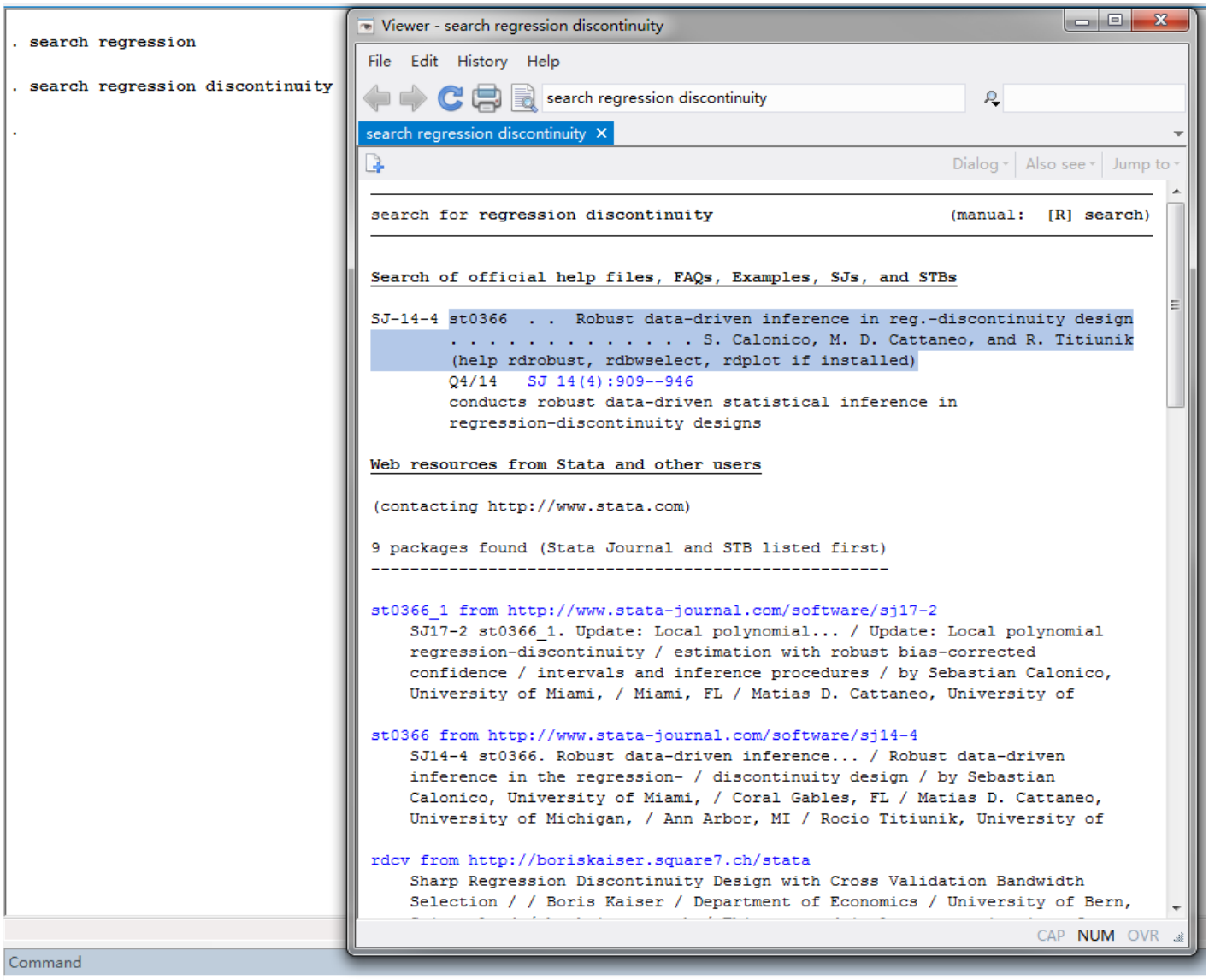
- Ctrl + D (w/o any highlight): execute the whole do-file
- Ctrl + D (w highlight): execute the selected commands (can involve multiple lines)
- Ctrl + Shift + D: execute all the remaining codes from the cursor
- Alt + Cursor: rectangle selection
  - Keep you do-file tidy is not just for good-looking!

Resources for Learning Stata

- Undoubtedly, the best resource for learning Stata is... Stata itself.
  - help: if you know the very specific command
  - search: if you only have a general idea
- Many tutorials outside
  - https://stats.idre.ucla.edu/stata/
  - http://wlm.userweb.mwn.de/Stata/wstatbas.htm
  - Fei Wang's short note (Reading Material 1.1)
- Stata Journal—very good for advanced econometric program, such as `rdrobust` (Reading Material 1.2)


. help rdrobust
.
Viewer - help rdrobust
File Edit History Help
help rdrobust
help rdrobust
Dialog
Also see
Jump to
help rdrobust (SJ17-2: st0366_1)
Title
rdrobust — Local polynomial regression-discontinuity estimation with robust bias-corrected confidence intervals and inference procedures
Syntax
rdrobust depvar runvar [if] [in] [, c(cutoff) p(pvalue) q(qvalue) deriv(dvalue) fuzzy(fuzzyvar [sharpbw]) covs(covars) kernel(kernelfn) weights(weightsvar) h(hvalueL hvalueR) b(bvalueL bvalueR) rho(rhovalue) scalepar(scaleparvalue) bwselect(bwmethod) scaleregul(scaleregulvalue) vce(vcemethod) level(level) all]
where depvar is the dependent variable and runvar is the running variable (also known as the score or forcing variable).
Description
rdrobust implements local polynomial regression-discontinuity (RD) point estimators with robust bias-corrected confidence intervals and inference procedures developed in Calonico, Cattaneo, and Titiunik (2014b), Calonico, Cattaneo, and Farrell (forthcoming), and Calonico et al. (2016). It also computes alternative estimation and inference procedures available in the literature.
rdrobust has two companion commands: rdbwselect for data-driven bandwidth selection, and rdplot for data-driven RD plots (see Calonico, Cattaneo, and Titiunik [2015a] for details).
CAP NUM OVR
Command


. search regression
.
Viewer - search regression
File Edit History Help
search regression
search ols
search regression
search regression
Dialog
Also see
Jump to
search for regression (manual: [R] search)
Search of official help files, FAQs, Examples, SJs, and STBs
[U] Chapter 25 . . . . Working with categorical data and factor variables
(help generate, fvvarlist)
[U] Chapter 26 . . . . . Overview of Stata estimation commands
(help estcom)
[R] regress . . . . . Linear regression
(help regress)
[R] regress postestimation . . . . . Postestimation tools for regress
(help regress postestimation)
[R] regress postestimation plots . . . . Postestimation plots for regress
(help regress postestimation plots)
[R] regress postestimation time series Postest. regress with time series
(help regress postestimationts)
[R] logistic . . . . . Logistic regression, reporting odds ratios
(help logistic)
[R] logistic postestimation . . . . . Postestimation tools for logistic
(help logistic postestimation)
[R] probit . . . . . Probit regression
(help probit)
[R] poisson . . . . . Poisson regression
(help poisson)
CAP NUM OVR
Command


. search regression
. search regression discontinuity
.
Command
Viewer - search regression discontinuity
File Edit History Help
search regression discontinuity
search regression discontinuity
Dialog
Also see
Jump to
search for regression discontinuity (manual: [R] search)
Search of official help files, FAQs, Examples, SJs, and STBs
SJ-14-4 st0366 . . Robust data-driven inference in reg.-discontinuity design
. . . . . S. Calonico, M. D. Cattaneo, and R. Titiunik
(help rdrobust, rdbwselect, rdplot if installed)
Q4/14 SJ 14(4):909--946
conducts robust data-driven statistical inference in
regression-discontinuity designs
Web resources from Stata and other users
(contacting http://www.stata.com)
9 packages found (Stata Journal and STB listed first)
st0366_1 from http://www.stata-journal.com/software/sj17-2
SJ17-2 st0366_1. Update: Local polynomial... / Update: Local polynomial
regression-discontinuity / estimation with robust bias-corrected
confidence / intervals and inference procedures / by Sebastian Calonico,
University of Miami, / Miami, FL / Matias D. Cattaneo, University of
st0366 from http://www.stata-journal.com/software/sj14-4
SJ14-4 st0366. Robust data-driven inference... / Robust data-driven
inference in the regression- / discontinuity design / by Sebastian
Calonico, University of Miami, / Coral Gables, FL / Matias D. Cattaneo,
University of Michigan, / Ann Arbor, MI / Rocio Titiunik, University of
rdcv from http://boriskaiser.square7.ch/stata
Sharp Regression Discontinuity Design with Cross Validation Bandwidth
Selection / / Boris Kaiser / Department of Economics / University of Bern,
CAP NUM OVR

Outline

Know the data sufficiently WELL before doing any econometric analysis!

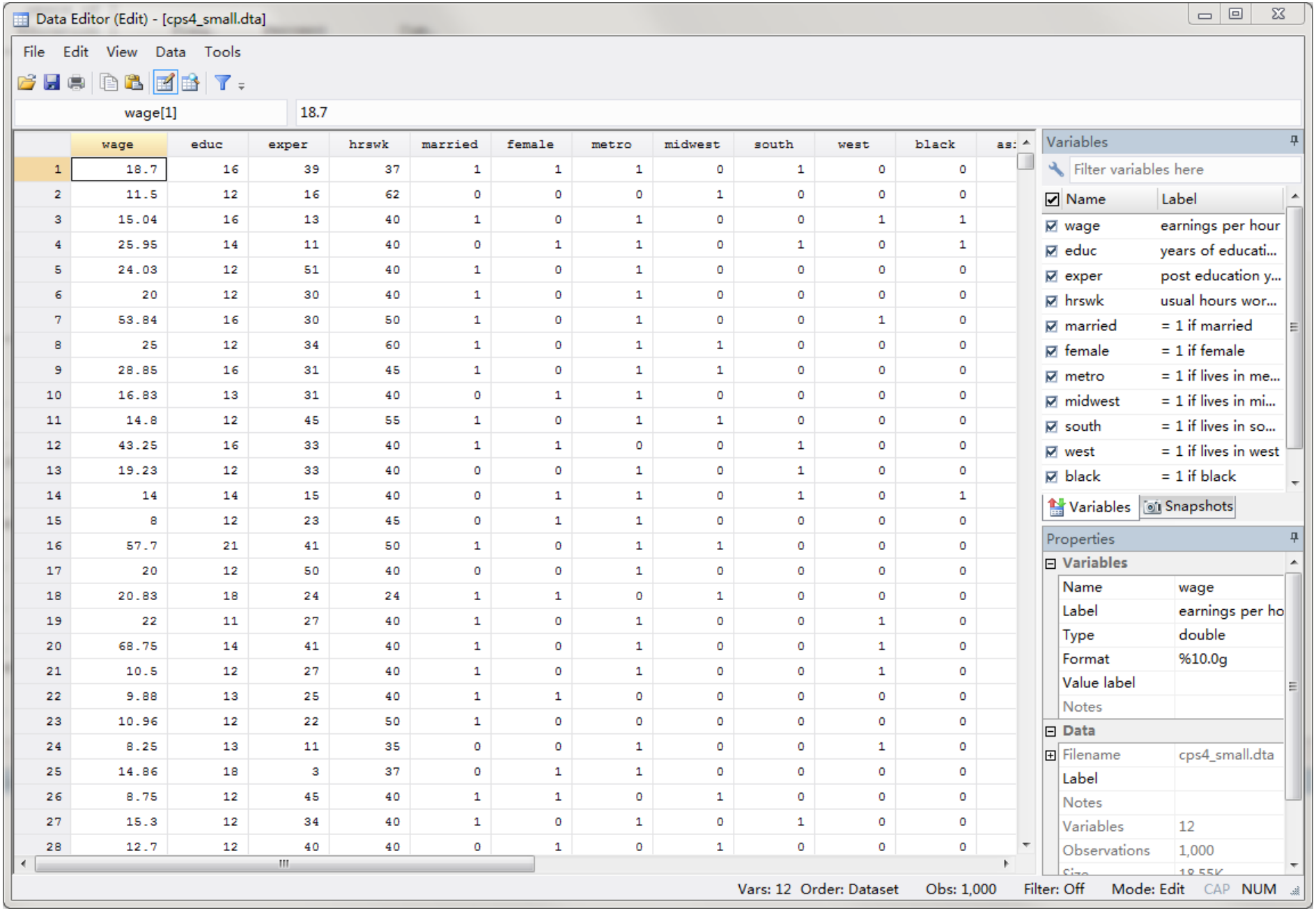
- Econometric models always need some assumptions.
  - Many of them are hard or even impossible to test.
- Data description does not require any assumption.

Load Data

- To load a data, first make sure the data is in the right place!
- Data loading: `use` command

```
use cps4_small.dta,clear
```

- Data in Stata looks similar to an Excel sheet:
  - Each row represents an "observation."
  - Each column represents a "variable."



	wage	educ	exper	hrswk	married	female	metro	midwest	south	west	black	as:
1	18.7	16	39	37	1	1	1	0	1	0	0	
2	11.5	12	16	62	0	0	0	1	0	0	0	
3	15.04	16	13	40	1	0	1	0	0	1	1	
4	25.95	14	11	40	0	1	1	0	1	0	1	
5	24.03	12	51	40	1	0	1	0	0	0	0	
6	20	12	30	40	1	0	1	0	0	0	0	
7	53.84	16	30	50	1	0	1	0	0	1	0	
8	25	12	34	60	1	0	1	1	0	0	0	
9	28.85	16	31	45	1	0	1	1	0	0	0	
10	16.83	13	31	40	0	1	1	0	0	0	0	
11	14.8	12	45	55	1	0	1	1	0	0	0	
12	43.25	16	33	40	1	1	0	0	1	0	0	
13	19.23	12	33	40	0	0	1	0	1	0	0	
14	14	14	15	40	0	1	1	0	1	0	1	
15	8	12	23	45	0	1	1	0	0	0	0	
16	57.7	21	41	50	1	0	1	1	0	0	0	
17	20	12	50	40	0	0	1	0	0	0	0	
18	20.83	18	24	24	1	1	0	1	0	0	0	
19	22	11	27	40	1	0	1	0	0	1	0	
20	68.75	14	41	40	1	0	1	0	0	1	0	
21	10.5	12	27	40	1	0	1	0	0	1	0	
22	9.88	13	25	40	1	1	0	0	0	0	0	
23	10.96	12	22	50	1	0	0	0	0	0	0	
24	8.25	13	11	35	0	0	1	0	0	1	0	
25	14.86	18	3	37	0	1	1	0	0	0	0	
26	8.75	12	45	40	1	1	0	1	0	0	0	
27	15.3	12	34	40	1	0	1	0	1	0	0	
28	12.7	12	40	40	0	1	0	1	0	0	0	

Data Description

- Data description—
  `describe`
  `summarize`
- Sometimes you can use abbreviations—e.g., you can type `sum` instead of `summarize`.
  - But at the beginning, I encourage you to type the full command.
  - Stata also allows abbreviations for variable names. But I recommend NEVER doing so.

How to describe data: tabulate command

- `tabulate` is a very useful command to describe the frequency *discrete variables* (DO NOT use the command for continuous variable!)
- You can tabulate one variable—
  `tabulate female`
- You can also tabulate two variables—
  `tabulate female married`
- `tabulate` can not only report frequencies, but also shares—
  `tabulate female married, column`
  `tabulate female married, row`

How to describe data: summarize command

- `summarize` not only can used to describe the whole data set, it can also be used on each variable separately
  `summarize wage`
- To see more details, you can add `detail` option,
  `summarize wage, detail`
- You can also summarize a list of variables:
  `summarize wage educ exper hrswk`

Using if condition

- Sometimes we do not wish to describe the full sample, instead, we may only want to learn the information of female.
- Logic conditions:
  - == equal to
  - != not equal to
  - <= less than or equal to; < less than
  - >= greater than or equal to; > greater than
- You can add `if` condition to most Stata commands:

```
summarize wage if female==1
tabulate married if female!=1
```

- IMPORTANT: "=" for expression; "==" for logic (Q: what does the following command mean?)

```
generate married_women = [married == 1] if female == 1
```

Combining Multiple Conditions

- You can also impose multiple `if` condition
  - | - or
  - & - and
- You can use bracket to specify the priority
- `count` command—count the observations that satisfy the condition(s)

```
count if female==1&married==1
count if (female==1 | married==1)&wage<=20
count if female==1 | (married==1 &wage<=20
```

Exercise 1.1—Old Age Support in China

Data support.dta contains the follow information from three waves of CHARLS (China Health and Retirement Longitudinal Study)

FN097_w2 Who do you think you can rely on financially for old-age support? 如果您将来老了干不动工作了，您认为生活来源主要将是什么？

1. Children 子女 → Skip to FN098_w2 请跳至 FN098_w2
2. Savings 储蓄结束本部分
3. Pension or retirement salary 养老金或退休金结束本部分
4. Commercial pension insurance 商业养老保险结束本部分
5. Other 其他结束本部分

Try to answer following questions by using only `tabulate`

1. Overall, what's the most important source of old-age support in China? Children or pension?
2. Is the answer different for urban and rural residents?
3. Chinese government has been improving the public pension system, especially in rural China. Can you find any sign of it from the data?

Sometimes, you can really know a lot simply by describing data!

How to describe data—plot graphs

Generally speaking, there are two types of graphs—oneway graph and twoway graph.

One-way graph is the graph that only requires one variable, a typical example is a histogram,

- histogram graph -
  `histogram wage`
- To save the graph, use `export graph` command (make sure the graph window is open!)
  `graph export hist_wage.wmf, replace`

Two-way graph is the graph that requires two variables

- scatter graph (when you have two or more variables, make sure to start with `twoway` command, *first y variable, then x variable*!)—
  `twoway scatter wage educ`
- You can also draw a fitted line—
  `twoway lfit wage educ`
- But sometimes you wish to draw above two graphs together, use "()" or "||"to separate different graphs
  `twoway (scatter wage educ) (lfit wage educ)`
  or `twoway scatter wage educ $||$ lfit wage educ`

Simple graphs can also be very informative—"One Graph is Worth a Thousand Words"

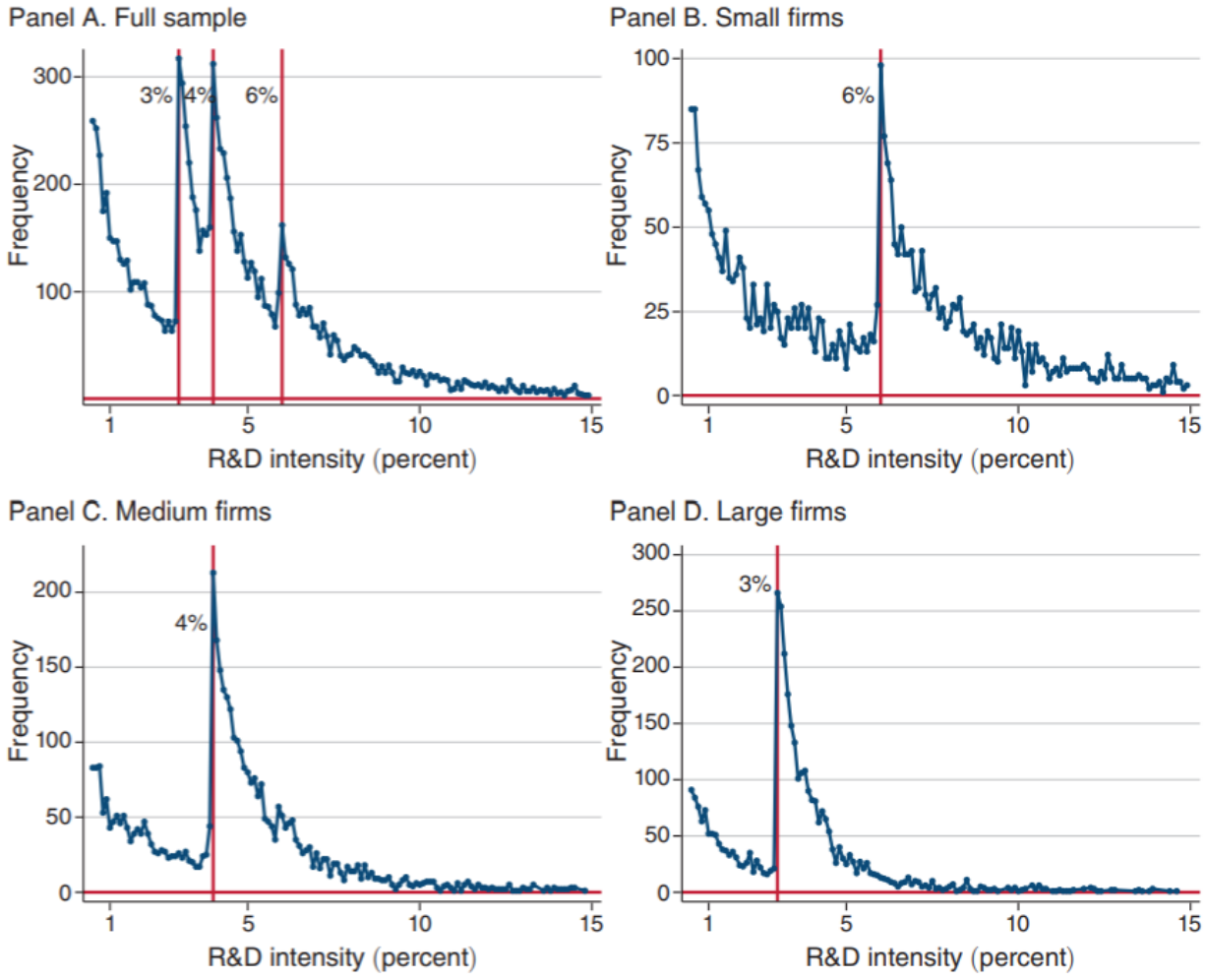


FIGURE 2. BUNCHING AT DIFFERENT THRESHOLDS OF R&D INTENSITY, 2011

Notes: This figure plots the empirical distribution of R&D intensity for all manufacturing firms with R&D intensity between 0.5 percent and 15 percent in the Administrative Tax Return Database. Panel A reports the pooled data distribution with all sizes of firms. Panels B, C, and D report the R&D intensity distribution of small, medium, and large firms, respectively. Note that large fractions of the firms bunch at the thresholds (6 percent for small, 4 percent for medium, and 3 percent for large) at which they qualify to apply for the InnoCom certification.

Source: Administrative Tax Return Database. See Section IIA for details.

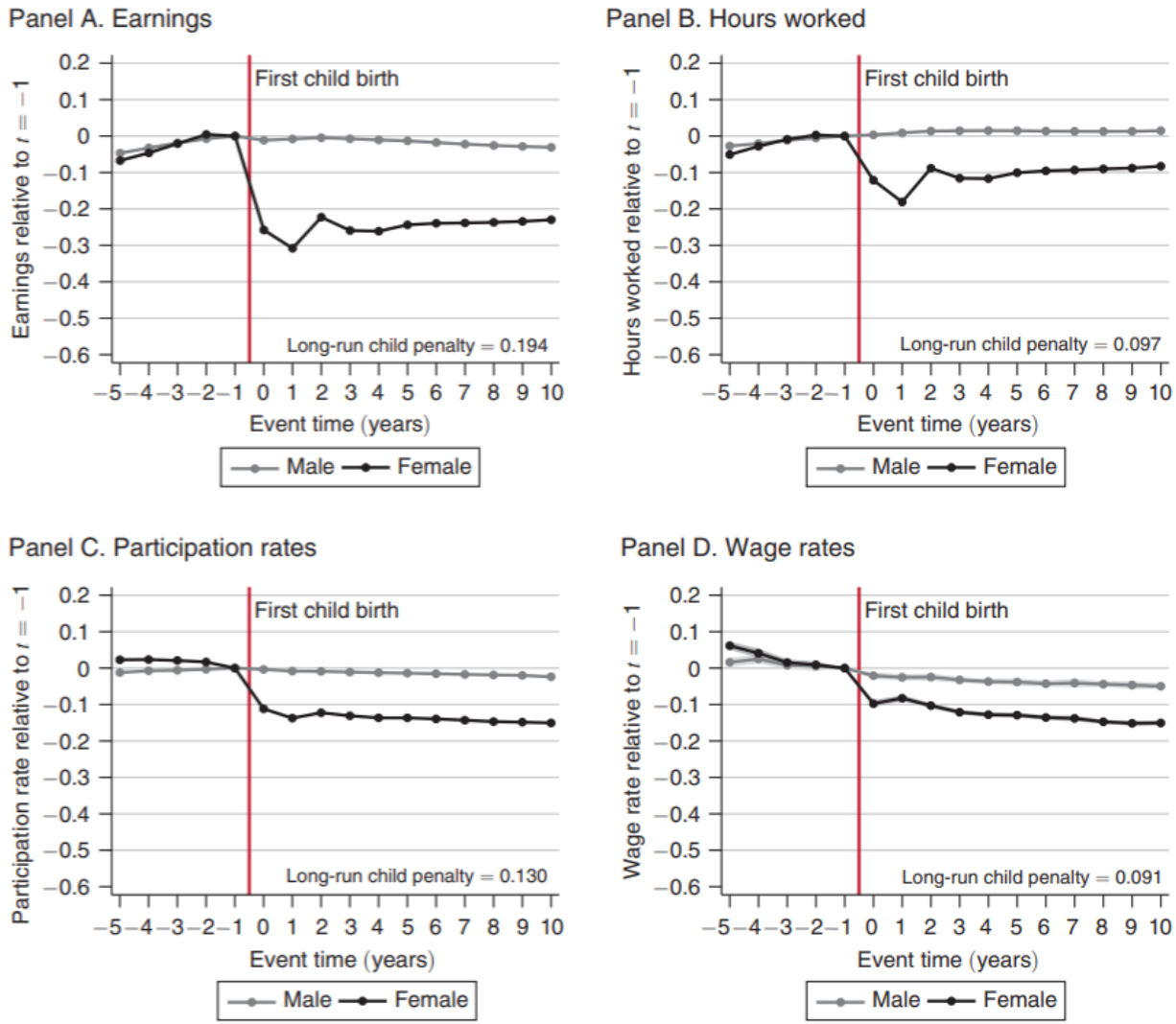


FIGURE 1. IMPACTS OF CHILDREN

From: Kleven, Henrik, Camille Landais and Jakob Egholt Søgaard. 2019. Children and Gender Inequality: Evidence from Denmark. American Economic Journal: Applied Economics, 11(4), 181–209.

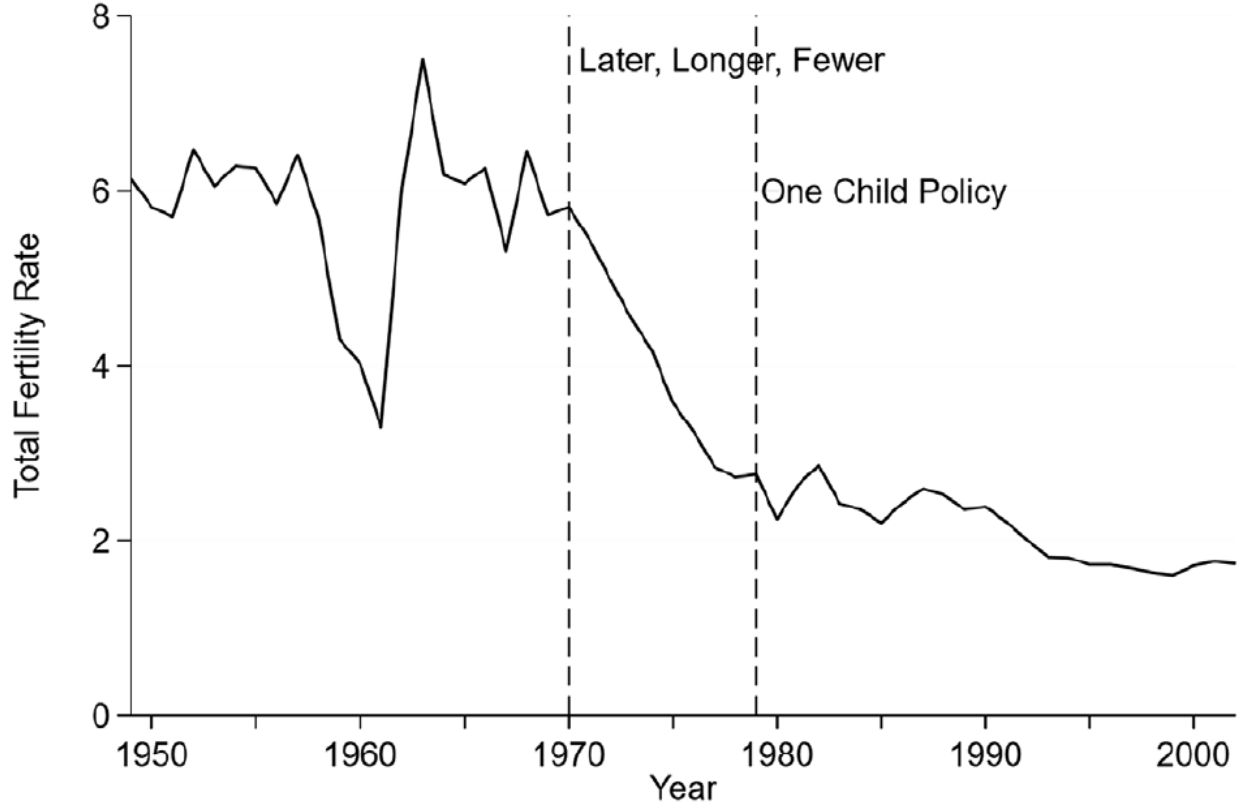


Fig. 1. National total fertility rate of China, 1949–2002.
Data source: Lu and Zhai (2009) "Sixty Years of New China Population."

From: Chen, Yi and Hanming Fang. 2021. The Long-Term Consequences of China's "Later, Longer, Fewer" Campaign in Old Age. Journal of Development Economics, 151, 102664.

Outline

Manipulate "Variables"

- After we know the data "sufficiently" well, we can start to analyze the data.
- But before running regressions, we may need to modify the data according to our needs.
- I assume you already know the basic commands, including: `generate`, `replace`, `drop`, `keep`.
  - `drop` and `keep` can be applied to BOTH variables and observations.
  - `replace` (or `recode`) are concrete examples of the conflict between "programming time" and "running time."
    - For large data sets, using generate→drop→rename maybe faster than using `replace`.

gen versus egen

- `egen`: extensions to `generate`
  - More complicated operations to variables/observations.
  - `help egen`
- Although the two commands look very similar and share some command features, they differ in some important aspects.
  - The same function may perform differently, e.g., `sum`
  - May treat missing value differently.
- It is important to check that what Stata DO is what you THINK.
  - "What You Think Is What You Get"

Stata Example—From CPI to Price Deflator

In economics, we often need to use "real" values. However, what we typically have is data on CPI.

You are given a time series of China's CPI "cpi.dta". You are asked to calculate a series of deflators that adjusts nominal values to the price level in 2000.

1. What are the deflators?
2. How to compute the deflators using Excel?
3. How to compute the deflators using Stata?

Stata Example—From CPI to Price Deflators (Cont.)

In the previous slides, we use the trick $\log(a \times b) = \log(a) + \log(b)$ to transform running multiplication to running sum.

Q: what if a (or b) is negative or equals to zero?

bysort: "generate" within each group

- Often, you wish to generate a value for each group. Take household for example,
  - Aggregate individual income to generate household income
  - Number of children/working adults/seniors in household
  - The best-educated in household
- When generating such kind of variables, you will not only use the information of the individual, but also use the information of other people in the same group.
- Stata example: count number of seniors (age$\geq$60) for each household

Exercise 1.2—Keep "Nuclear" Households

Multi-generational households are often very difficult to analysis empirically. To keep the analysis simple, researchers often focus on "nuclear" households—households that are at most composed of father, mother, and young children (age$<$18).

You are given a data (chip2002.dta) containing follow variables at individual level: age, relationship to household head. How to pick up those "nuclear" households?

Manipulate "Data"

- So far, we have been manipulating "variables."
- But sometimes, we may need to change the entire structure of the data. e.g., `reshape`.

long

i	*j*	*stub*
1	1	4.1
1	2	4.5
2	1	3.3
2	2	3.0

`reshape` <⟶>

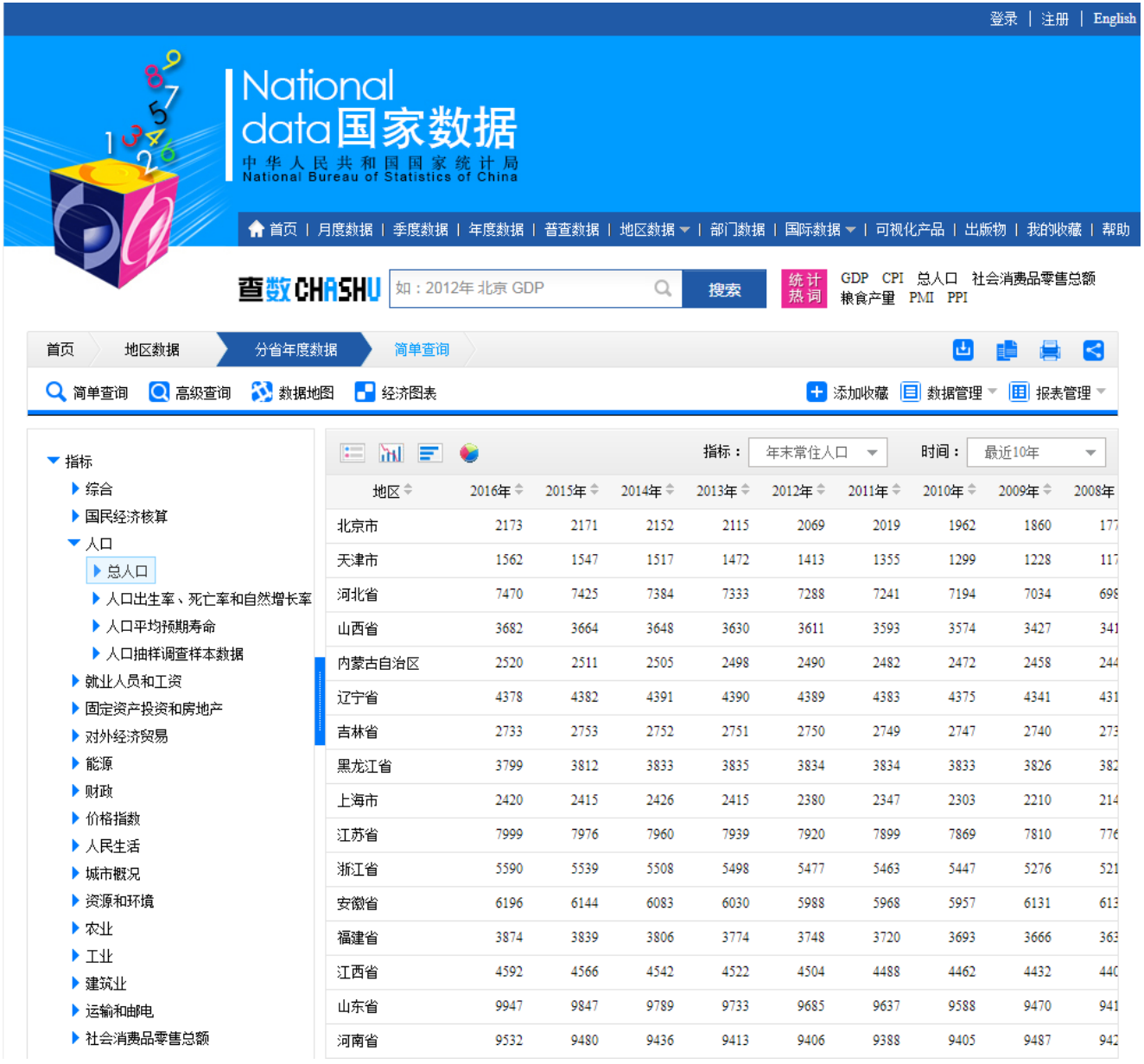
wide

i	*stub1*	*stub2*
1	4.1	4.5
2	3.3	3.0

- Other examples include: `xpose`, `merge`, `append`.
  - A "unique" identifier plays a central role in `merge` command. More details later.

Data in Excel format often looks like this:

GeoFips	GeoName	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957
00000	United States	1.93E+08	1.93E+08	2.14E+08	2.36E+08	2.49E+08	2.64E+08	2.71E+08	2.9E+08	3.1E+08	3.28E+08
01000	Alabama	2471947	2382967	2662225	3047380	3234463	3374775	3285389	3705568	3960676	4193495
02000	Alaska	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
04000	Arizona	856875	892996	978633	1198121	1338733	1440476	1499279	1641513	1840024	2011001
05000	Arkansas	1529473	1443263	1545153	1738594	1803288	1807509	1783027	1951398	2003164	2069298
06000	California	16328095	16930872	18884915	21407772	23483086	25383334	26392235	28848373	31709645	33987167
08000	Colorado	1706932	1742194	1907826	2229155	2369643	2395991	2487957	2717047	2968282	3295732
09000	Connecticut	3127814	3120784	3481317	3900173	4192305	4583649	4732791	5114012	5556616	5953327
10000	Delaware	435656	481420	553910	595260	634145	684774	716885	815169	960967	962796
11000	District of Columbia	1632778	1689982	1790171	1840925	1890377	1867970	1851862	1826556	1919532	1987674
12000	Florida	2942337	3095135	3520448	3899139	4377369	4893982	5193561	5933568	6824376	7549472
13000	Georgia	3048038	3080560	3562973	4078257	4349731	4515538	4521434	4977742	5317198	5481685
15000	Hawaii	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)	(NA)
16000	Idaho	689809	684967	741924	819019	890716	860570	876662	927298	1004086	1075804
17000	Illinois	13992219	13398699	14696564	15891440	16524568	17734981	18120771	19427214	20922155	21836490
18000	Indiana	5181118	4973942	5621879	6382212	6641141	7322346	7008221	7539316	8092990	8338189
19000	Iowa	3850529	3264637	3776405	3961572	4186452	3990428	4383449	4160718	4403187	4886884
20000	Kansas	2332838	2345851	2636654	2880927	3327186	3194350	3425624	3434360	3617804	3811267
21000	Kentucky	2639642	2530635	2748592	3186199	3334931	3504870	3480265	3637744	3830789	3976619
22000	Louisiana	2445645	2663637	2855397	3146798	3378092	3593308	3616680	3836780	4237082	4680189
23000	Maine	1028077	1018353	1043853	1136657	1222227	1234375	1275984	1410595	1473831	1530431
24000	Maryland	3156450	3245460	3624711	4048492	4406534	4750488	4874186	5233582	5672663	6056051
25000	Massachusetts	6472310	6506597	7215419	7696460	7882328	8417837	8694087	9160063	9756463	10295504
26000	Michigan	8790815	8878347	10050675	10950976	11585312	13078967	13007154	14484893	15013183	15465843
27000	Minnesota	3809163	3618102	3992061	4337904	4432944	4702507	4910583	5176944	5389100	5718265
28000	Mississippi	1581504	1413436	1621578	1770002	1864834	1896582	1845098	2089751	2110848	2137278
29000	Missouri	4859816	4760539	5246777	5677411	5908969	6266962	6427659	6864015	7234960	7438665
30000	Montana	814884	752157	924494	992718	1007430	1022775	1021182	1121511	1177350	1224754
31000	Nebraska	1812062	1653194	1967805	2004577	2156440	2038612	2213727	2135830	2192877	2575933



地区	2016年	2015年	2014年	2013年	2012年	2011年	2010年	2009年	2008年
北京市	2173	2171	2152	2115	2069	2019	1962	1860	177
天津市	1562	1547	1517	1472	1413	1355	1299	1228	117
河北省	7470	7425	7384	7333	7288	7241	7194	7034	698
山西省	3682	3664	3648	3630	3611	3593	3574	3427	341
内蒙古自治区	2520	2511	2505	2498	2490	2482	2472	2458	244
辽宁省	4378	4382	4391	4390	4389	4383	4375	4341	431
吉林省	2733	2753	2752	2751	2750	2749	2747	2740	273
黑龙江省	3799	3812	3833	3835	3834	3834	3833	3826	382
上海市	2420	2415	2426	2415	2380	2347	2303	2210	214
江苏省	7999	7976	7960	7939	7920	7899	7869	7810	776
浙江省	5590	5539	5508	5498	5477	5463	5447	5276	521
安徽省	6196	6144	6083	6030	5988	5968	5957	6131	613
福建省	3874	3839	3806	3774	3748	3720	3693	3666	363
江西省	4592	4566	4542	4522	4504	4488	4462	4432	440
山东省	9947	9847	9789	9733	9685	9637	9588	9470	941
河南省	9532	9480	9436	9413	9406	9388	9405	9487	942

Question: what should the data looks like in Stata?

Stata example: how to work with this type of data?

Outline

Return Values and Scalar

- Not all commands are used to generate the final output. Many of them are used to generate intermediate output, which will be used in further analysis.
  - e.g., deviation from the mean $z_i = (x_i - \overline{x})$
- If you have variable x_i in Stata, how would you generate z_i?
  - Use `summarize x` to find the mean of x_i, say 6.66
  - `generate z_i = x_i - 6.66`
  - It is very dangerous to do so!
- Principles of automation
  - Automate everything that can be automated.
  - Write a single script that executes all code from beginning to end.

The correct approach is to use the return values—a critical part for automation.

```
. summarize wage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wage	1,000	20.61566	12.83472	1.97	76.39

```
. return list

scalars:
                  r(N) =  1000
              r(sum_w) =  1000
               r(mean) =  20.61566
                r(Var) =  164.7301579223223
                 r(sd) =  12.8347246921125
                r(min) =  1.97
                r(max) =  76.39
                r(sum) =  20615.66

. display r(mean)
20.61566
```

Stata-defined scalars only keeps the return list from the last command. But you can define your own scalar to record the numbers of interest.

```
. summarize educ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
educ	1,000	13.799	2.711079	0	21

```
. scalar m_educ = r(mean)

.
. summarize exper
```

Variable	Obs	Mean	Std. Dev.	Min	Max
exper	1,000	26.508	12.85446	2	65

```
. display r(mean)
26.508

. display m_educ
13.799
```

Exercise 1.3—Computing Standard Deviations

In data `cps4_small.dta`, you are asked to calculate the standard deviation of wage using the following formula:

$$sd(\text{wage}) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\text{wage}_i - \overline{\text{wage}})^2}$$

You are not supposed to directly use the standard deviation from `summarize` command.

Hint: How many operations are included in the formula? How to realize them in Stata?

- Split one "big" problem to several "smaller" problems.

Break into several smaller problems:

1. Get necessary information such as $\overline{\text{wage}}$ and N
2. Generate a new variable $(\text{wage}_i - \overline{\text{wage}})^2$
3. Compute the summation of the above variable
4. Compute the standard deviation using the formula

Note that the whole process should be fully automated, which means you should not stop Stata until you get the final output $sd(\text{wage})$.

Using Local and Global

- The general idea of local/global is similar to scalar—they use "symbols" to represent something.
- The usage is also similar—first, define a local/global; then call it.
  - As its name suggests, locals are only effective "locally"—the defining command and calling command have to be executed in the same program.
  - Locals are usually the preferable method among scalar/local/global—it avoids some bugs that maybe difficult to find.
- The most severe bugs in programming are not those that software will report errors, but those that program can be executed but the outputs are wrong.
- Locals play a central role in our advanced course.

Loop—forvalues/foreach

- You often need to repeat an operation for several times. Instead of copying & pasting repeatedly, you should utilize the loop structure in Stata.
  - It makes your program look much nicer.
  - It makes things much easier when revising the code.
- Basic usage

Using Loops More "Automatically"

- The input of values seems trivial at the beginning—it is because either you have very limited values or the the values follow some simple rules. But it's not always the case!
- e.g., China's population census covers 31 provinces (excluding Taiwai, Hong Kong and Macau)
  - The provincial codes do not have a specific rule . . . and there are 31 codes!
  - The code may change by year! (Chongqing became a municipality directly under the central government in 1997. Such changes are much more common at prefecture level and county level.)
- `levelsof`: displays a sorted list of the distinct values of *varname*.

Provincial Code in China

北京 11 天津 12 河北 13 山西 14 内蒙古 15

辽宁 21 吉林 22 黑龙江 23

上海 31 江苏 32 浙江 33 安徽 34 福建 35 江西 36 山东 37

河南 41 湖北 42 湖南 43 广东 44 广西 45 海南 46

重庆 50 四川 51 贵州 52 云南 53 西藏 54

陕西 61 甘肃 62 青海 63 宁夏 64 新疆 65

香港 81 澳门 82

Using Loops More "Flexibly"

- Sometimes, you need to do a set of operations that are similar but not exactly the same.
- You don't wish to give up the loop structure merely because of several lines of different commands.
- Solution: use `if` loop within the `forvalues`/`foreach` loop.
  - Note `if` loop here is different from `if` conditions.

Stata example: China's aging process from 1990 to 2000 in different provinces.

Outline

Numbers in Regressions

- I assume that you already know how to run simple regressions using `regress` command.
- I also assume that you know how to interpret a standard regression output in Stata.
- Similar to `summarize` command, all numbers in `regress` output will be stored somewhere.
  - check `ereturn list`
  - The coefficient of *varname* will be stored in _b[*varname*] and the standard error will be stored in _se[*varname*].

- One often overlooked number is the number of "observations" used in regression.
  - Observations in data $\neq$ observations in regression.
  - `regress y x1` requires that the dependent variable (y) and the independent variables (x1) are not missing.
- Pay attention when the number of observations changes a lot across regressions.
- e.g., going from `regress y x1` to `regress y x1 x2` involves two effect:
  - A further restriction that x2 is not missing.
  - Adding a new variable x2.
- Look at following three regressions:
  1. regress y x1
  2. regress y x1 if missing(x2)==0
  3. regress y x1 x2

Hypothesis Test

- In Stata output, aside from Coef. and Std.Err., you can also see t-value and p-value.
  - In statistics and econometrics courses, you have to first define a "hypothesis" before computing t-value and p-value.
  - In Stata, those two values are for a specific hypothesis test: _b[*varname*]=0
- But we are not always interested in knowing whether the coefficient is zero or not. e.g., testing life-cycle hypothesis.
- Takeaway: report **standard errors** in tables! Not t-value, not p-value.
- Using the post-estimation command `test` to test linear hypothesis more flexibly.
  - Post-estimation commands are those that can only be used after an estimation. `help regress postestimation`

Nonlinear Models

- Nonlinearity in the variables

$$\begin{aligned} y &= \beta_1 + \beta_2 x^2 + e \\ y &= \beta_1 + \beta_2 \frac{1}{x} + e \\ \ln(y) &= \beta_1 + \beta_2 \ln(x) + e \end{aligned}$$

- Nonlinearity in parameters

$$\begin{aligned} y &= \beta_1 + \beta_2^2 x \\ y &= \beta_1 + \beta_1 \beta_2 x \end{aligned}$$

- When we are talking about nonlinear models, we refer to the latter case.

Probit Model

I will use Probit model as an example of nonlinear model.

- For binary dependent variable y:

$$\Pr(y = 1) = \Phi\left(\beta_1 + \beta_2 x_i\right),$$

  where Φ is the CDF of a standard normal distribution.

- The interpretation of β_2 is unclear here. Therefore, we usually does not directly report the output of `probit` command.

Marginal Effect:

$$ME = \frac{\partial \Pr(y_i = 1)}{\partial x_i} = \phi(\beta_1 + \beta_2 x_i)\beta_2$$

Notice that the marginal effect is a function of x_i. For reporting purpose, usually we prefer to report a single number.

- Average marginal effect (AME) is defined as

$$E\left(\frac{\partial \Pr(y_i = 1)}{\partial x_i}\right) = E\left[\phi(\beta_1 + \beta_2 x_i)\beta_2\right]$$

- Marginal effect at mean (MEM) is defined as

$$\left.\frac{\partial \Pr(y_i = 1)}{\partial x_i}\right|_{x_i = \overline{x}} = \phi(\beta_1 + \beta_2 \overline{x})\beta_2$$

- Economists usually prefer average marginal effect.
- Reading Material 1.3 for more details.