Stata Workshop: DAY 2

Taisei NODA

Graduate School of Economics, Osaka University

June 18, 2019

Overview

Routine Work

- Let's begin by the "routine work"
- Change the working directory to your personal folder (e.g. "L:")
- Do not forget to open your log file
- Today we are going to work on "wb.csv".
  - This data comes from **Education Statistics in 2015** by World bank.
  - I downloaded the original data from DataBank and modify it. This data is public data.
  - Make sure you have downloaded the csv file into the working directory.
  - Use **dir**

See If You Have the Dataset: dir

```
. dir
 902.3k  10/20/14 13:30  LBftf@fCf‹f fbfN2_v209.exe
  <dir>  12/25/17  8:17  LBftf@fCf‹f fbfN2_Readme
  <dir>   3/20/19 18:27  System Volume Information
1008.1k   3/26/19 19:32  rufus-3.4p.exe
1802.7M   3/26/19 19:34  ubuntu-ja-18.04.1-desktop-amd64.iso
   0.5k   5/08/19 18:16  LB FileLock2.dat
  14.3M   5/20/19 17:35  fj.lock
  <dir>   5/08/19 14:55  fj
   4.8k   6/17/19 10:18  wb.csv
  <dir>   6/17/19 12:32  io
  <dir>   6/17/19 12:33  temp
   1.1k   6/17/19 13:08  stata2019_day2.log
```

Import Data

- To import csv file, we use **import delimited**
- **varnames(1)** option specifies the row of variable names
- For dta file, we use "**use....,clear**" command
- For excel file(.xlsx), **import excel** command is available
  - see **help import excel**

Replace Parts of Data

- Today, we will use the variables of "countryname, countrycode, var1, var3, var4, var5, var6, var7".
- We can choose variables of interest by **drop** and **keep**

Why No Observation ?

```
. su
```

Variable	Obs	Mean	Std. Dev.	Min	Max
countryname	0				
countrycode	0				
var1	0				
var3	0				
var4	71	460.0912	56.75294	324.0882	570.706
var5	71	462.7892	53.374	338.6303	542.0488
var6	0				
var9	71	.1971831	.4007036	0	1

Why Colored Red?

	countryname	countrycode	var1	var3	var4	var5	var6	var9
1	Albania	ALB	.74197	90.54527	412.8957	406.6567	11800	0
2	Algeria	DZA	..	..	356.8391	348.7972	14260	0
3	Argentina	ARG	2.38742	89.85171	407.4132	427.6997	20030	0
4	Australia	AUS	1.62449	92.71159	495.3505	508.7095	45230	0
5	Austria	AUT	2.17453	84.95183	500.7409	490.9135	49390	0
6	Belgium	BEL	..	85.47176	512.7295	507.1101	45330	0
7	Brazil	BRA	2.58198	83.31311	371.3515	404.8044	15360	0
8	Bulgaria	BGR	..	95.81945	440.8897	436.6862	17820	0
9	Canada	CAN	..	99.94666	517.6515	531.3152	43720	0
10	Chile	CHL	1.43118	77.97418	421.6818	460.5004	22290	0
11	China	CHN	..	..	537.6656	501.1963	14440	1
12	Colombia	COL	1.59344	77.33186	386.3858	425.2064	13530	0
13	Costa Rica	CRI	2.38546	77.60461	397.616	426.7143	15050	0
14	Croatia	HRV	..	97.60586	462.2638	488.5583	22860	0
15	Cyprus	CYP	2.63435	95.32878	437.5244	446.8593	31980	0
16	Czech Republic	CZE	1.72534	84.94709	494.1965	492.3631	31420	0
17	Denmark	DNK	..	91.7673	513.4732	505.4148	50360	0
18	Dominican Republic	DOM	1.2897	61.07617	324.0882	353.7894	13700	0
19	Estonia	EST	1.35091	91.43698	520.6267	522.5063	28570	0
20	Finland	FIN	2.65718	96.59672	513.8009	534.3968	42530	0

Change String to Numeric: destring

- We can transform string to numeric by **destring**
- Before do that, you need to replace all the text with numeric values or missing(".")
- Now var1, var3 and var 6 have ".." instead of ". Stata recognizes ".." as string.
- **replace** changes row values
- To avoid repetitive task, we use "loop syntax"

Loop: foreach

```
*Loop: same operation for each variable in the variable list
foreach var of varlist var1 var3 var6 {
replace `var'="." if `var'==".."
}
destring _all,replace
```

- This command executes replacement of ".." with "." for each variable in the "varlist". i.e. var1, var3, and var6.
- Then, run **destring**

Label for a Variable: label variable

- You can put labels for variables by **label variable**

Variables

Variable	Label
countryname	
countrycode	
expend	Government expenditure on education as % of GDP (%)
enroll	Net enrollment rate(%), secondary
math	PISA math, median
read	PISA reading, median
gni_pc	GNI per capita
asia	Asia dummy

Label for Values

- Also, you can put labels for each value
- e.g. "yes" if the value = 1, "no" if the value = 0

ll	math	read	gni_pc	asia
4527	412.8957	406.6567	11800	no
.	356.8391	348.7972	14260	no
5171	407.4132	427.6997	20030	no
L159	495.3505	508.7095	45230	no
5183	500.7409	490.9135	49390	no
7176	512.7295	507.1101	45330	no
L311	371.3515	404.8044	15360	no
L945	440.8897	436.6862	17820	no
4666	517.6515	531.3152	43720	no
7418	421.6818	460.5004	22290	no
.	537.6656	501.1963	14440	yes
3186	386.3858	425.2064	13530	no
0461	397.616	426.7143	15050	no
0586	462.2638	488.5583	22860	no
2878	437.5244	446.8593	31980	no
4709	494.1965	492.3631	31420	no

Output Summary Tables: outreg2

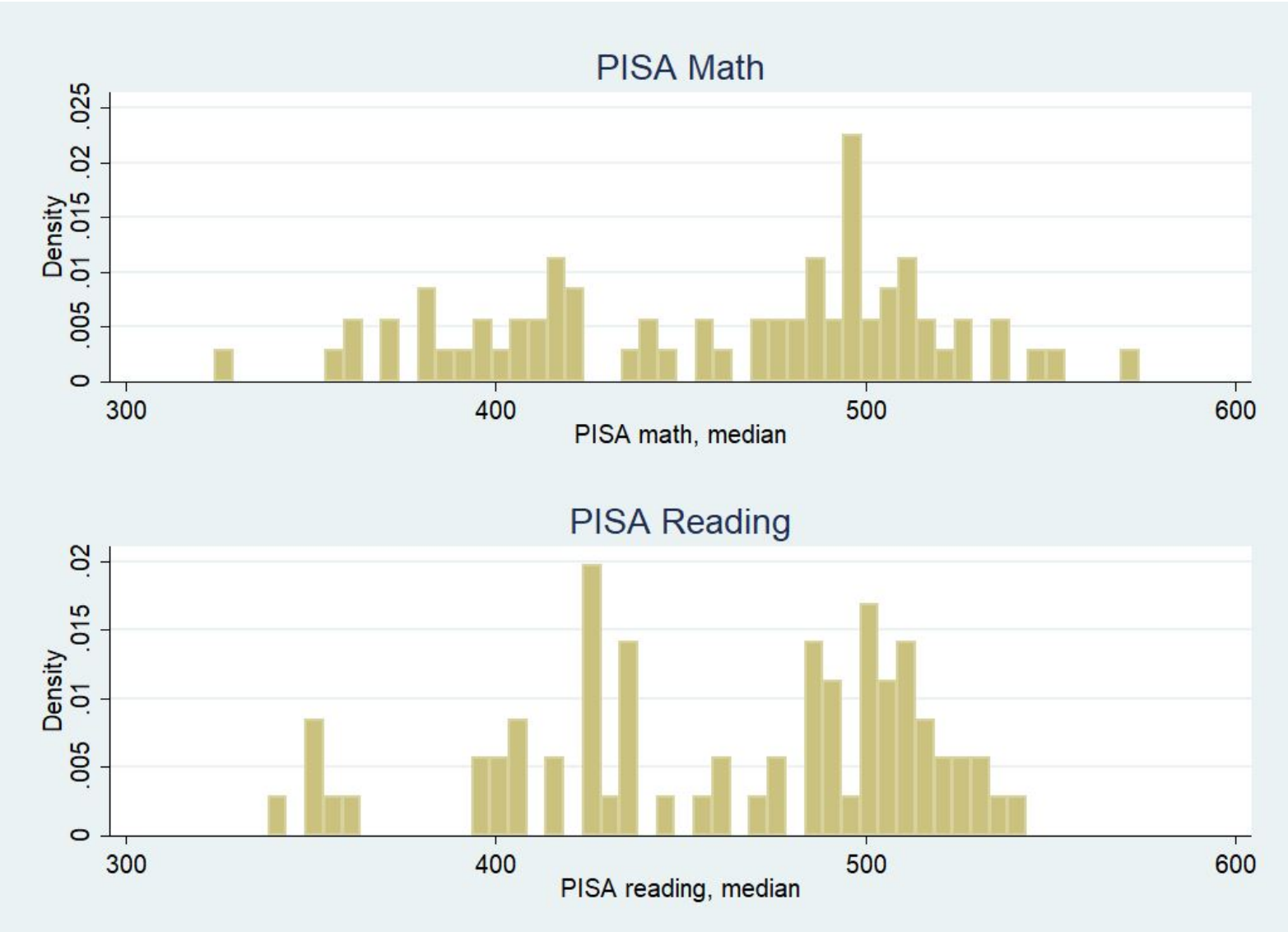
- **outreg2** provides a fast and easy way to produce an illustrative table of outputs.
- This is user-written file (we call "ado file"), and then not pre-installed. You need to install by yourself.
- **ssc install outreg2**

Summary Table by outreg2

- Your tabel should be saved in your working directory

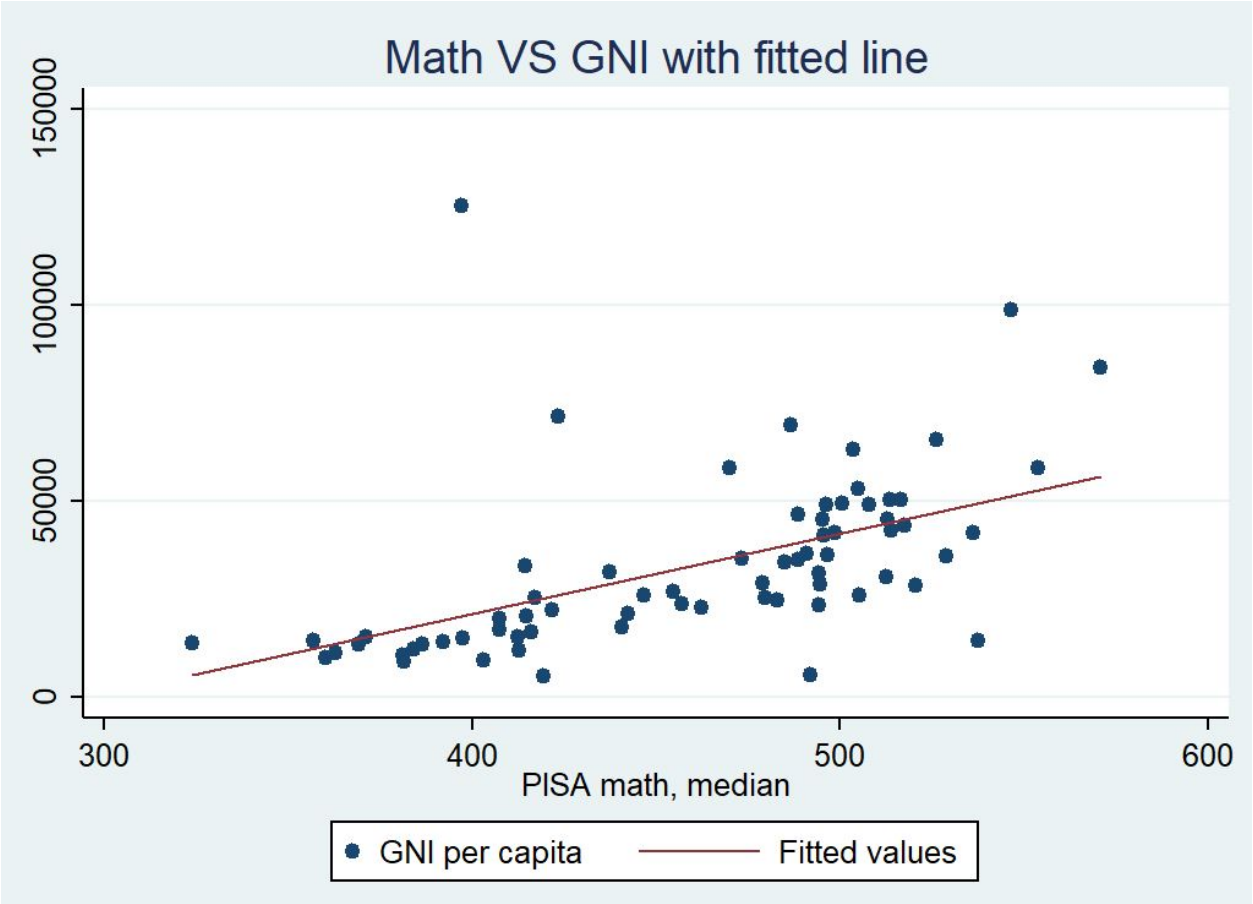
VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
expend	48	1.875	0.615	0.742	4.724
enroll	58	90.52	7.939	61.08	99.95
math	71	460.1	56.75	324.1	570.7
read	71	462.8	53.37	338.6	542.0
gni_pc	70	33,486	22,307	5,430	125,200
asia	71	0.197	0.401	0	1

Histogram


PISA Math
Density
0
.005
.01
.015
.02
.025
300
400
500
600
PISA math, median
PISA Reading
Density
0
.005
.01
.015
.02
300
400
500
600
PISA reading, median

Scatter Plot: twoway scatter

- You can draw scatter plot by **twoway scatter**
- Also, you can add fitted line on the figure
  - "||" combines two types of graphs into one figure
  - "lfit" draws a fitted line



Correlation:corr

```
. corr gni_pc math read enroll expend
(obs=44)
```

	gni_pc	math	read	enroll	expend
gni_pc	1.0000				
math	0.7481	1.0000			
read	0.6969	0.9454	1.0000		
enroll	0.3607	0.6691	0.6445	1.0000	
expend	0.1190	0.1299	0.1829	0.2355	1.0000

T Test:ttest

- **ttest** performs group mean comparison, so called t-test

```
. ttest math,by(asia)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
no	57	454.835	7.112264	53.69641	440.5874	469.0826
yes	14	481.4914	17.54294	65.63965	443.5922	519.3906
combined	71	460.0912	6.735335	56.75294	446.658	473.5244
diff		-26.6564	16.74593		-60.06361	6.750819

```
    diff = mean(no) - mean(yes)                               t =  -1.5918
Ho: diff = 0                     degrees of freedom =       69

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0580         Pr(|T| > |t|) = 0.1160          Pr(T > t) = 0.9420
```

Linear Regression (OLS):reg

```
reg gni_pc math
```

Source	SS	df	MS
Model	9.5355e+09	1	9.5355e+09
Residual	2.4798e+10	68	364671472
Total	3.4333e+10	69	497581844

Number of obs	= 70
F(1, 68)	= 26.15
Prob > F	= 0.0000
R-squared	= 0.2777
Adj R-squared	= 0.2671
Root MSE	= 19096

gni_pc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	205.8102	40.2482	5.11	0.000	125.4962	286.1243
_cons	-61150.79	18647.44	-3.28	0.002	-98361.19	-23940.4

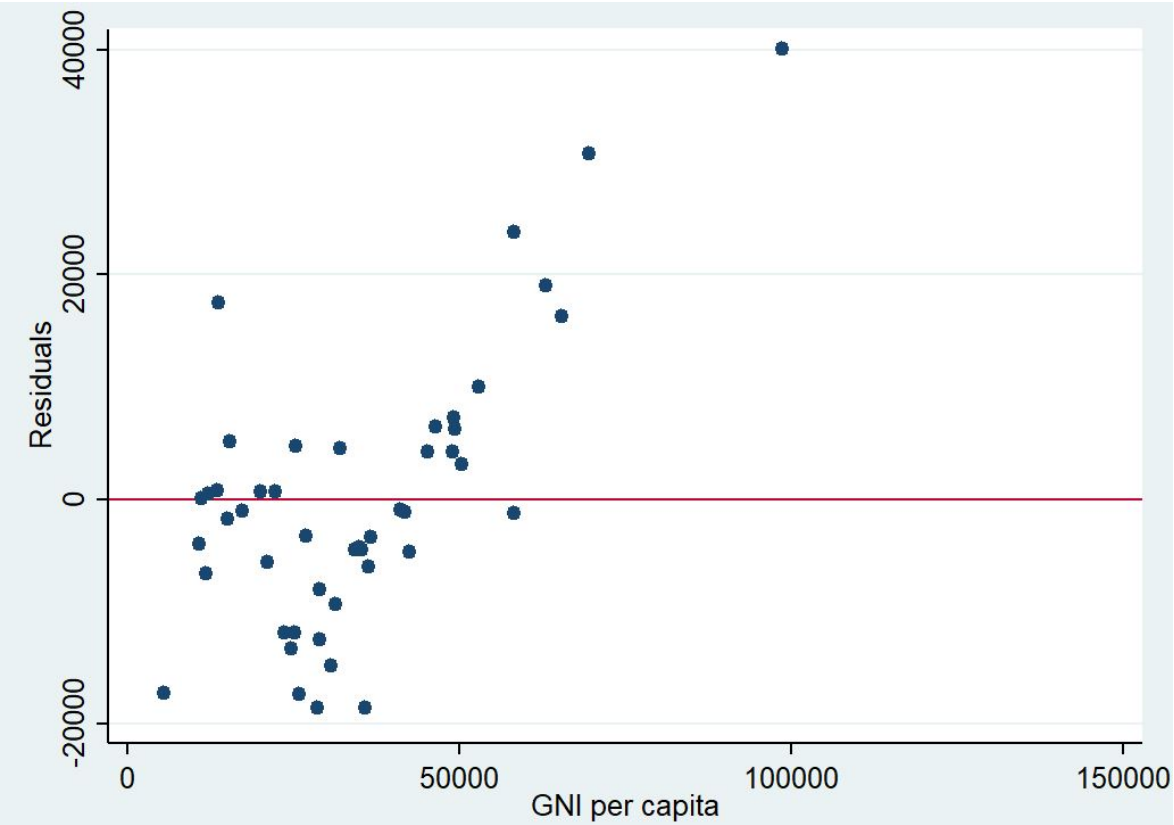
- **reg** command has various **post estimation**
  - **outreg2**
  - **predict double resid,residual**
  - **coefplot**
  - etc

Illustrative Table of Regression by outreg2

	(1)	(2)
VARIABLES	gni_pc	gni_pc
math	205.8***	258.5***
	(40.25)	(35.68)
asia		4,157
		(5,712)
expend		1,398
		(3,081)
Constant	-61,151***	-89,337***
	(18,647)	(18,065)
Observations	70	47
R-squared	0.278	0.567
Standard errors in parentheses		
*** p<0.01, ** p<0.05, * p<0.1		

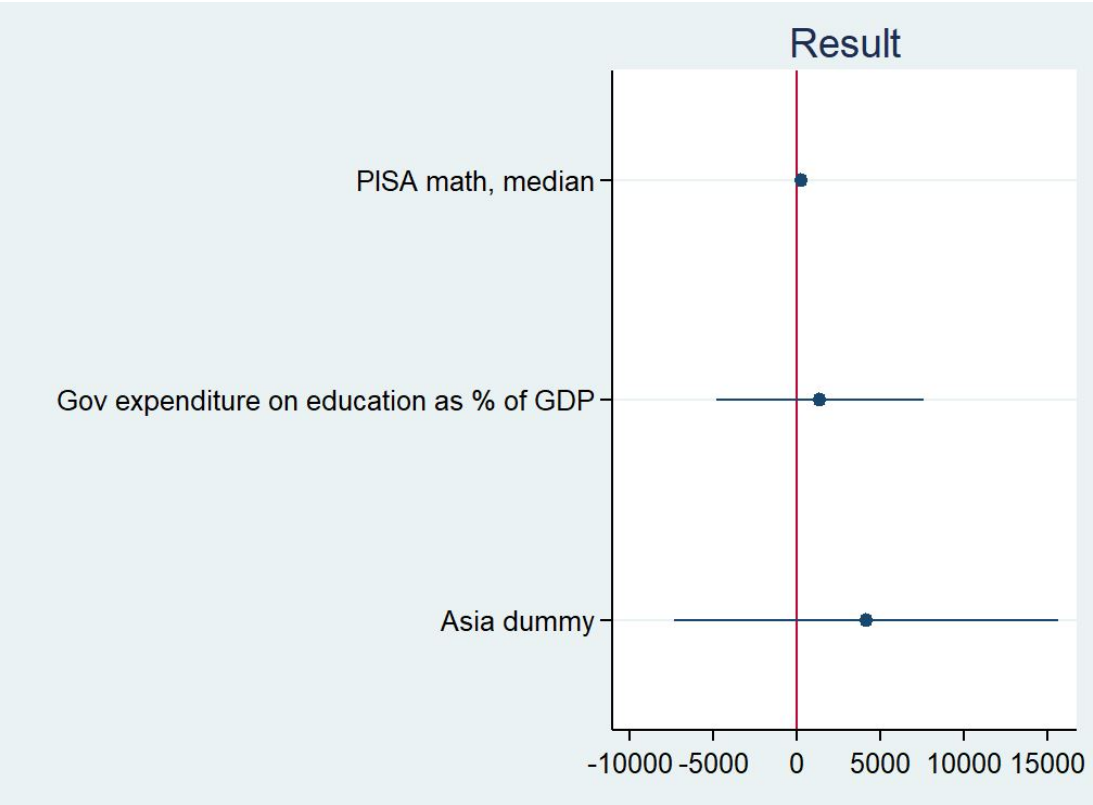
Calculating Residuals: predict double resid,residuals

- **predict double resid,residuals**
  - **resid** names the residuals. Any name is fine.
  - Note: You can also calculate predicted values by this command. See the help file.
- draw scattering plot to see the distribution of the residuals



Visualization of Regression Results: coefplot

- **coefplot** plots point estimated coefficients and the confidence intervals.



Save Data:save

- **save**
- **saveold** command saves your dta file in older version (e.g. stata 12)
- Note that Stata has often compatibility problem. Older version sometimes does not work for dta file generated by the newest version (Stata 13 cannot open dta file generated by Stata 15).