計量經濟學電腦實習課

L3 Panel Data Review, and Binary Dependent Variable Model

陳柏瑜 Chen Boyie

May 4, 2020

Graduate Institute of Economics
National Taiwan University
r08323004@ntu.edu.tw

Brief

Panel Data Review

Binary Dependent Variable

Commands in Panel Data

Stata Homework 4 Announcement

Panel Data Review

EE10.1

我們直接從課本的 Empirical Exercise 10.1 來複習如何用 Stata 處理 Panel Data 的資料，並得到需要的統計推論。

Gun Data

1. Observations: 50 states, 1 district
2. Time Period : 1977-1999

Variable Definitions

Variable	Definition
vio	violent crime rate (incidents per 100,000 members of the population)
rob	robbery rate (incidents per 100,000)
mur	murder rate (incidents per 100,000)
shall	= 1 if the state has a shall-carry law in effect in that year = 0 otherwise
incarc_rate	incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year)
density	population per square mile of land area, divided by 1000
avginc	real per capita personal income in the state, in thousands of dollars
pop	state population, in millions of people
pm1029	percent of state population that is male, ages 10 to 29
pw1064	percent of state population that is white, ages 10 to 64
pb1064	percent of state population that is black, ages 10 to 64
stateid	ID number of states (Alabama = 1, Alaska = 2, etc.)
year	Year (1977-1999)

We will discuss the State and Time Fixed Effect respectively.

EE10.1 Models

Before we consider any State/Time Fixed Effect, the two models are:

$$log(vio_{it}) = \beta_0 + \beta_1 shall_{it} + u_{it}$$

and

$$log(vio_{it}) = \beta_0 + \beta_1 shall_{it} + controls + u_{it}$$

Notice that the above models can be estimated by OLS, or say Pooled OLS.

EE10.1 OLS

```
. reg lvio shall, r //想想看直接用reg跑OLS是什麼意思？ 當作有NT個obs的cross sectional data

Linear regression                               Number of obs     =      1,173
                                                F(1, 1171)        =      86.86
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0866
                                                Root MSE          =     .61735

------------------------------------------------------------------------------
             |               Robust
        lvio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       shall |  -.4429646   .0475283    -9.32   0.000    -.5362148   -.3497144
       _cons |   6.134919   .0193039   317.81   0.000     6.097045    6.172793
------------------------------------------------------------------------------
```

EE10.1 Pooled OLS

```
. reg lvio shall, vce(cluster stateid) //還是只有修正s.e.，允許加入serial correlations

Linear regression                               Number of obs     =      1,173
                                                F(1, 50)          =       7.96
                                                Prob > F          =     0.0068
                                                R-squared         =     0.0866
                                                Root MSE          =     .61735

                                   (Std. Err. adjusted for 51 clusters in stateid)
------------------------------------------------------------------------------
             |               Robust
        lvio |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       shall |  -.4429646   .1570184    -2.82   0.007    -.7583452   -.1275839
       _cons |   6.134919   .0790269    77.63   0.000     5.976189    6.293649
------------------------------------------------------------------------------
```

EE10.1 Models

After we consider State and Time Fixed Effect, the two models are:

$$log(vio_{it}) = \alpha_i + \beta_1 shall_{it} + \lambda_t + u_{it}$$

and

$$log(vio_{it}) = \alpha_i + \beta_1 shall_{it} + \lambda_t + controls + u_{it}$$

EE10.1 (a.) Table

VARIABLES	(1) POLS 1	(2) POLS 2	(3) S.FE 1	(4) S.FE 2	(5) T.FE 1	(6) T.FE 2
shall	-0.443***	-0.368***	0.114***	-0.0461	0.00188	-0.0280
	(0.157)	(0.114)	(0.0360)	(0.0418)	(0.0403)	(0.0407)
incarc_rate		0.00161***		-7.10e-05		7.60e-05
		(0.000600)		(0.000250)		(0.000208)
density		0.0267		-0.172		-0.0916
		(0.0415)		(0.138)		(0.124)
avginc		0.00121		-0.00920		0.000959
		(0.0241)		(0.0130)		(0.0165)
pop		0.0427***		0.0115		-0.00475
		(0.0117)		(0.0142)		(0.0152)
pb1064		0.0809		0.104***		0.0292
		(0.0714)		(0.0327)		(0.0495)
pw1064		0.0312		0.0409***		0.00925
		(0.0341)		(0.0135)		(0.0238)
pm1029		0.00887		-0.0503**		0.0733
		(0.0341)		(0.0207)		(0.0525)
Constant	6.135***	2.982	6.000***	3.866***	5.820***	3.766***
	(0.0790)	(2.167)	(0.00876)	(0.770)	(0.0260)	(1.152)
Observations	1,173	1,173	1,173	1,173	1,173	1,173
R-squared	0.087	0.564	0.039	0.218	0.383	0.418
Number of stateid			51	51	51	51
State FE			YES	YES	YES	YES
Year FE					YES	YES

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

EE10.1 a (i.)Interpretations

Is the estimate large or small in a real-world sense?

Notice that we use $log(vio)$ as dependent varaible, and the unit of `vio` is "percentage", that is, we have to transform the coeficient we estimated before interpretations.

Suppose $log(vio_{it}) = \beta_1 shall_{it}$, then $vio_{it} = e^{\beta_1 shall}$.

Since $\frac{\partial vio}{\partial shall} = \beta_1 e^{\beta_1 shall}$, the interpretation of β_1 would be:

If `shall` increase by 1 unit, then `vio` would increase by $\beta_1 e^{\beta_1}$ units.

EE10.1 a (i.)Interpretations

But notice that `shall` is a binary variable here, so the coeficient estimated should be interpreted as:

If `shall` changes from 0 to 1, then the violent crime rate would increase by $\beta_1 e^{\beta_1} - \beta_q e^0 = \beta_1(e^{\beta_1} - 1)$. (單位是：起 / 每十萬人)

1. POLS 1 : (-0.443)*(exp(-0.443)-1) = 0.159
2. POLS 2 : (-0.368)*(exp(-0.368)-1) = 0.113
3. State FE 1: (0.114)*(exp(0.114)-1) = 0.014
4. State FE 2: (-0.0461)*(exp(-0.0461)-1) = 0.002
5. Time FE 1 : (0.00188)*(exp(0.00188)-1) $\approx$ 0
6. Time FE 2 : (-0.028)*(exp(-0.028)-1) = 0.0008

The crime rate increased (起/每十萬人) is not relatively large in every model.

EE10.1 a (ii.)

If we do not consider any fixed effect, we're basically using an OLS approach with clustered standard errors. And we should look at column (1) and column (2). The coefficient estimated are both statistically significant but not large in real-world scale.

(Haven't asked yet) After adding control variables in state fixed effect model (from column (3) to column (4)), we observe that the coefficient estimated β_1 is not statistically significant.

EE10.1 a (iii.)

讓 Z_i 作為第 i 個州的 idiosyncratic variable，且 Z_i 不隨時間改變。

則可能的 Z_i 可以是：

1. 人口結構：除了主流族裔，有亞裔、非裔、拉丁裔等少數族裔的地區可能呈現不同的文化特色。
2. 地緣位置：東西岸或南北方州可能基於工業化速度不同，而對於擁槍與否持不同意見。

EE10.1 b

After adding state fixed effects, we now focus on column (3) and (4).

The two new columns show that the OLS approach overestimate the effect of `shall` on `vio`.

And column (4) controls more factors that are statistically significant, that is, we tend to believe that column (4) helps us achieve "ceteris paribus."

EE10.1 c

After adding time fixed effects, we now focus on column (5) and (6).

The argument is the same as the previous one in (b.)

And column (6) controls more factors, that is, we tend to believe that column (6) helps us achieve "ceteris paribus."

```
. test $yeardum

 ( 1)  y78 = 0
 ( 2)  y79 = 0
 ( 3)  y80 = 0
 ( 4)  y81 = 0
 ( 5)  y82 = 0
 ( 6)  y83 = 0
 ( 7)  y84 = 0
 ( 8)  y85 = 0
 ( 9)  y86 = 0
 (10)  y87 = 0
 (11)  y88 = 0
 (12)  y89 = 0
 (13)  y90 = 0
 (14)  y91 = 0
 (15)  y92 = 0
 (16)  y93 = 0
 (17)  y94 = 0
 (18)  y95 = 0
 (19)  y96 = 0
 (20)  y97 = 0
 (21)  y98 = 0
 (22)  y99 = 0

       F( 22,    50) =   21.62
            Prob > F =    0.0000
```

Binary Dependent Variable

Review Binary Dependent Model

We say we have a binary model if

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

where y_i is either 0 or 1.

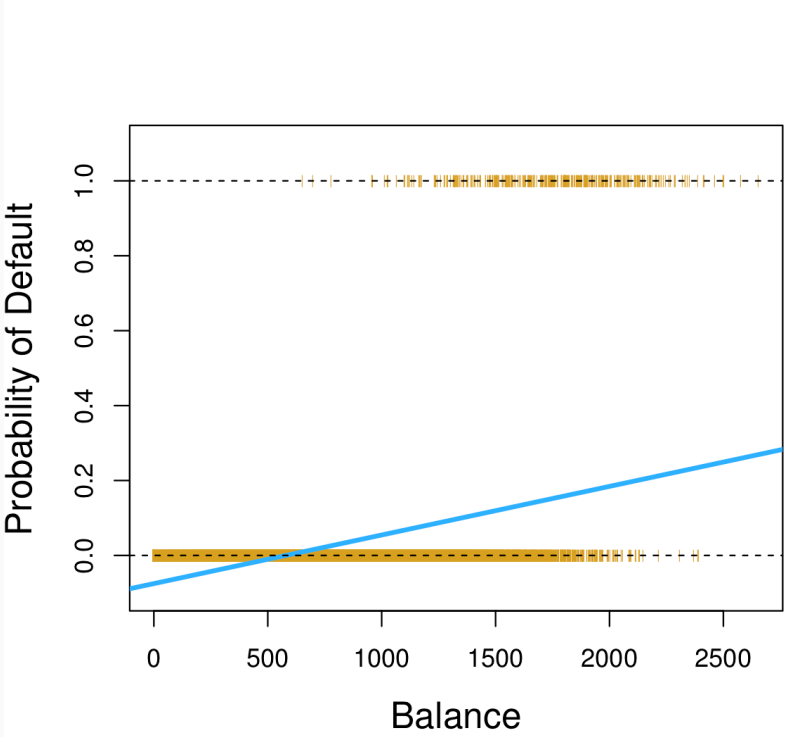
If we do somthing like:

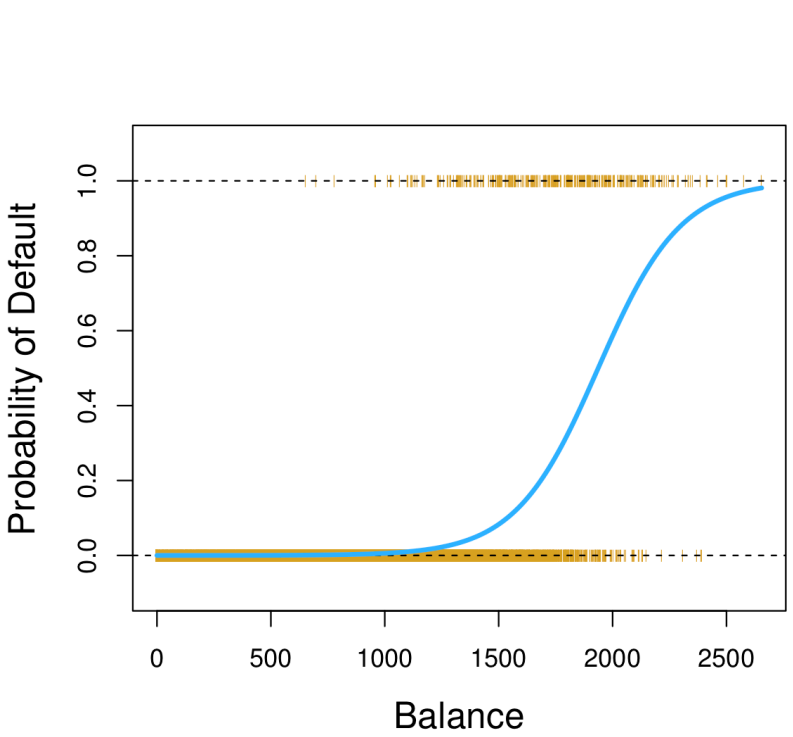
$$Pr(y_i = 1|X_i) = \xi(\beta_0 + \beta_1 X_i)$$

We model a expected mean (or conditional prob.) function.

Note that if $\xi(\cdot)$ is linear in β, then $\xi(\cdot)$ is LPM. Check Prof. Luoh's lecture note p.32.

LPM, Logit/Probit





LPM

Simply use OLS estimator.

Command:

```
#in Stata
reg y x,r

#in R
lm(y~x)
```

Probit

Instead of being an Ordinary least square estimator, $\hat{\beta}$ obtained from Probit models are called "Generalized Least Square" (GLS) estimators. Still check Prof. Luoh's lecture note p.32.

Command:

```
#in Stata
probit y x,r

#in R
glm(y ~ x, family = binomial(link = "probit"))
```

Logit

Command:

```
#in Stata
logit y x,r

#in R
glm(y ~ x, family = binomial(link = "logit"))
```

EE11.2

我們直接從課本的 Empirical Exercise 11.2 來複習如何用 Stata 處理 Binary Dependent Variable 的資料，並得到需要的統計推論。

EE11.2 Questions

E11.2 Believe it or not, workers used to be able to smoke inside office buildings. Smoking bans were introduced in several areas during the 1990s. Supporters of these bans argued that in addition to eliminating the externality of secondhand smoke, they would encourage smokers to quit by reducing their opportunities to smoke. In this assignment, you will estimate the effect of workplace smoking bans on smoking, using data on a sample of 10,000 U.S. indoor workers from 1991 to 1993, available on the text website, **http://www.pearsonglobaleditions.com**, in the file **Smoking**. The data set contains information on whether individuals were or were not subject to a workplace smoking ban, whether the individuals smoked, and other individual characteristics.[7] A detailed description is given in **Smoking_Description**, available on the website.

EE11.2 Questions

a. Estimate the probability of smoking for (i) all workers, (ii) workers affected by workplace smoking bans, and (iii) workers not affected by workplace smoking bans.

b. What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.

c. Estimate a linear probability model with *smoker* as the dependent variable and the following regressors: *smkban*, *female*, *age*, age^2, *hsdrop*, *hsgrad*, *colsome*, *colgrad*, *black*, and *hispanic*. Compare the estimated effect of a smoking ban from this regression with your answer from (b). Suggest an explanation, based on the substance of this regression, for the change in the estimated effect of a smoking ban between (b) and (c).

d. Test the hypothesis that the coefficient on *smkban* is 0 in the population version of the regression in (c) against the alternative that it is nonzero, at the 5% significance level.

e. Test the hypothesis that the probability of smoking does not depend on the level of education in the regression in (c). Does the probability of smoking increase or decrease with the level of education?

EE11.2 Questions

f. Repeat (c)–(e) using a probit model.

g. Repeat (c)–(e) using a logit model.

h. i. Mr. A is white, non-Hispanic, 20 years old, and a high school dropout. Using the probit regression and assuming that Mr. A is not subject to a workplace smoking ban, calculate the probability that Mr. A smokes. Carry out the calculation again, assuming that he is subject to a workplace smoking ban. What is the effect of the smoking ban on the probability of smoking?

ii. Repeat (i) for Ms. B, a female, black, 40-year-old college graduate.

iii. Repeat (i)–(ii) using the linear probability model.

Smoking Data

"Supporter of the workplace smoking bans argued that the bans would encourage smokers to quit."

1. Observations: 10000
2. Time Period : 1991, 1993 (different respondents)

Variable Definitions

Variable	Definition
smoker	=1 if current smoker, =0 otherwise
smkban	=1 if there is a work area smoking ban, =0 otherwise
age	age in years
hsdrop	=1 if high school dropout, =0 otherwise
hsgrad	=1 if high school graduate, =0 otherwise
colsome	=1 if some college, =0 otherwise
colgrad	=1 if college graduate, =0 otherwise
black	=1 if black, =0 otherwise
hispanic	=1 if Hispanic =0 otherwise
female	=1 if female, =0 otherwise

EE11.2 Table

VARIABLES	(1) LPM	(2) Probit	(3) Logit
smkban	-0.0472***	-0.159***	-0.262***
	(0.00897)	(0.0291)	(0.0495)
female	-0.0333***	-0.112***	-0.191***
	(0.00857)	(0.0288)	(0.0492)
age	0.00967***	0.0345***	0.0599***
	(0.00190)	(0.00688)	(0.0118)
age_sq	-0.000132***	-0.000468***	-0.000818***
	(2.19e-05)	(8.26e-05)	(0.000143)
hsdrop	0.323***	1.142***	2.017***
	(0.0195)	(0.0730)	(0.134)
hsgrad	0.233***	0.883***	1.579***
	(0.0126)	(0.0604)	(0.116)
colsome	0.164***	0.677***	1.230***
	(0.0126)	(0.0614)	(0.118)
colgrad	0.0448***	0.235***	0.447***
	(0.0120)	(0.0654)	(0.126)
black	-0.0276*	-0.0843	-0.156*
	(0.0161)	(0.0535)	(0.0913)
hispanic	-0.105***	-0.338***	-0.597***
	(0.0140)	(0.0494)	(0.0862)
Constant	-0.0141	-1.735***	-2.999***
	(0.0414)	(0.152)	(0.265)
Observations	10,000	10,000	10,000
R-squared	0.057		

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

MLE (Optional)

現在我們來試著解釋為什麼要用 MLE 來估 Probit 或 Logit Model 的 $\hat{\beta}$，以及在 Stata 輸入 `probit` 或 `logit` 指令後出現的 iteration 是什麼意思。

MLE (Optional)

Recall:

$$Pr(y_i|X) = \Phi(\beta_0 + \beta_1 X_i)$$

The OLS estimator $\hat{\beta}_{OLS}$ solves the following maximizing problem:

$$\max_{\hat{\beta}_0, \hat{\beta}_1} - \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

The NLS estimator $\hat{\beta}_{NLS}$ solves the following maximizing problem:

$$\max_{\hat{\beta}_0, \hat{\beta}_1} - \sum_{i=1}^{n} (y_i - \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

MLE (Optional)

$$\max_{\hat{\beta}_0, \hat{\beta}_1} -\sum_{i=1}^{n} (y_i - \Phi(\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Write the maximizing problem into matrix form with parameter β:

$$\hat{\beta}_{NLS} = arg\max_{\beta} -\sum_{i=1}^{n} (y_i - \Phi(X_i'\beta))'(y_i - \Phi(X_i'\beta))$$

This is the NLS estimator for parameter β. Check Prof. Luoh's lecture note in p.32.

We will see it yields the same estimate $\hat{\beta}$ as MLE.

MLE (Optional)

Since we impose the assumption for the conditional probability function for y_i, which is:

$$Prob(y_i = 1|X) = \Phi(\beta_0 + \beta_1 X_i)$$

and

$$Prob(y_i = 0|X) = 1 - \Phi(\beta_0 + \beta_1 X_i)$$

We first write the above equations in matrix form.

$$Prob(y_i = 1|X) = \Phi(X_i'\beta)$$

MLE (Optional)

$$Prob(y_i = 1|X) = \Phi(X_i'\beta)$$

Note that after we condition on X, y_i follows Bernoulli distribution.

$$y_i|X \sim Bernoulli(p)$$

where $p = \Phi(X_i'\beta)$ is a function of X

MLE (Optional)

Write the density function of y_i as $f_{Y|X}(y_i)$:

$$f_{Y|X}(y_i) = (\Phi(X_i'\beta))^{y_i}(1 - \Phi(X_i'\beta))^{1-y_i}$$

And the likelihood function $\mathcal{L}$ is:

$$\mathcal{L} = \prod_{i=1}^{n}(\Phi(X_i'\beta))^{y_i}(1 - \Phi(X_i'\beta))^{1-y_i}$$

MLE (Optional)

Take log,

$$\ell = log\mathcal{L} = \sum_{i=1}^{n}(y_i log(\Phi(X_i'\beta)) + (1 - y_i) log(1 - \Phi(X_i'\beta)))$$

Then this is our objective function to maximize, i.e.

$$\hat{\beta}_{MLE} = arg\max_{\beta} \ell$$

MLE (Optional)

$$\ell = log\mathcal{L} = \sum_{i=1}^{n}(y_i log(\Phi(X_i'\beta)) + (1 - y_i) log(1 - \Phi(X_i'\beta)))$$

We may define our log-likelihood function in R:

```
mll = function(beta){
  logLikelihood = sum(Y*log(pnorm(X%*%beta))
                +(1-Y)*log(1-pnorm(X%*%beta)))
  return(-logLikelihood)
}
```

MLE (Optional)

Note that for a given β, which is a $K \times 1$ vector, the `mll(beta)` returns a scalar.

Thus we use a numerical maximization function `nlm()` in R to apply the Newton's Method.

```
est_probit = nlm(mll, rnorm(2),
                 print.level = 2, hessian = TRUE)
```

```
iteration = 9
Parameter:
[1] -0.5545698 -0.2448055
Function Value
[1] 5498.667
Gradient:
[1] -0.004574758 -0.001799890

Relative gradient close to zero.
Current iterate is probably solution.
```

MLE (Optional)

The parameter estimated is close to the result given by `glm()` function and Stata's command.

```
data = read.dta("~/Smoking.dta")
m.probit = glm(smoker ~ smkban,
               family = binomial(link = "probit"),
               data = data)
```

```
Call:
glm(formula = smoker ~ smkban, family = binomial(link = "probit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8269  -0.8269  -0.6904  -0.6904   1.7612

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.55457    0.02123 -26.126   <2e-16 ***
smkban      -0.24481    0.02787  -8.784   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11074  on 9999  degrees of freedom
Residual deviance: 10997  on 9998  degrees of freedom
AIC: 11001

Number of Fisher Scoring iterations: 4
```

MLE (Optional)

For the NLS estimator, we have silimar objective function:

$$\hat{\beta}_{NLS} = arg\max_{\beta} -\sum_{i=1}^{n}(y_i - \Phi(X_i'\beta))'(y_i - \Phi(X_i'\beta))$$

We may define our objective function to minimize in R:

```
lsq = function(beta){
  least_square = sum(t(Y-pnorm(X%*%beta))%*%(Y-pnorm(X%*%beta)))
  return(least_square)
}
```

MLE (Optional)

Note that for a given β, which is a $K \times 1$ vector, the `mll(beta)` returns a scalar.

Thus we use a numerical maximization function `nlm()` in R to apply the Newton's Method.

```
est_probit_lsq = nlm(lsq, rnorm(2),
                     print.level = 2, hessian = TRUE)
```

```
iteration = 8
Parameter:
[1] -0.5545686 -0.2448063
Function Value
[1] 1821.594
Gradient:
[1] -0.0007196377 -0.0004545200

Relative gradient close to zero.
Current iterate is probably solution.
```

MLE (Optional)

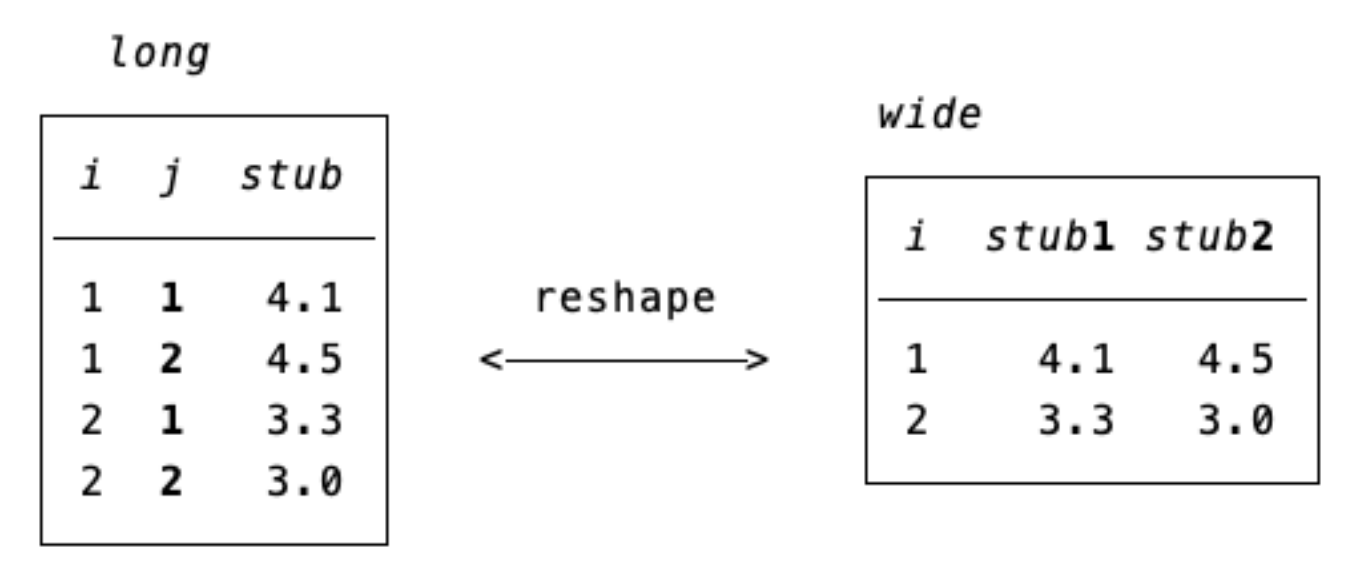
For now, we do not need to know why $\hat{\beta}_{NLS}$ and $\hat{\beta}_{MLE}$ give us the same result of estimate. We just need to know that they give us an insight about estimators, that is:

In most of the time, estimators are the solutions to some maximizing problems. And those maximized objective functions are some kind of distances. That's all.

Commands in Panel Data

Reshape

reshape wide [vlb_not_varies_by_time], i(individual) j(time) reshape long [vlb_with_time_subscript], i(individual) j(create_time_index)



Collapse

collapse (stat) [vlb] [fw = vlb]

mean	means (default)
median	medians
p1	1st percentile
p2	2nd percentile
...	3rd–49th percentiles
p50	50th percentile (same as **median**)
...	51st–97th percentiles
p98	98th percentile
p99	99th percentile
sd	standard deviations
semean	standard error of the mean (**sd/sqrt(n)**)
sebinomial	standard error of the mean, binomial (**sqrt(p(1-p)/n)**)
sepoisson	standard error of the mean, Poisson (**sqrt(mean)**)
sum	sums
rawsum	sums, ignoring optionally specified weight except observations with a weight of zero are excluded
count	number of nonmissing observations
percent	percentage of nonmissing observations
max	maximums
min	minimums
iqr	interquartile range
first	first value
last	last value
firstnm	first nonmissing value
lastnm	last nonmissing value

Outreg2

Please check the do file of EE10.1.

Stata Homework 4 Announcement

HW

1. Textbook Empirical Exercise 11.1 (a.)-(e.)
2. Notice that (f.)-(h.) are not included.

HW 格式要求

1. Upload only one pdf file.
2. All formulas should be expressed in LaTeX format.
3. Deadline is 5/19 Tue. 14:10

Example Answers

Check Stata handout.

EE11.1 Questions

Empirical Exercises

E11.1 In April 2008, the unemployment rate in the United States stood at 5.0%. By April 2009, it had increased to 9.0%, and it had increased further, to 10.0%, by October 2009. Were some groups of workers more likely to lose their jobs than others during the Great Recession? For example, were young workers more likely to lose their jobs than middle-aged workers? What about workers with a college degree versus those without a degree or women versus men? On the text website, **http://www.pearsonglobaleditions.com**, you will find the data file **Employment_08_09**, which contains a random sample of 5440 workers who were surveyed in April 2008 and reported that they were employed full-time. A detailed description is given in **Employment_08_09_Description**, available on the website. These workers were surveyed one year later, in April 2009, and asked about their employment status (employed, unemployed, or out of the labor force). The data set also includes various demographic measures for each individual. Use these data to answer the following questions.

EE11.1 Questions

a. What fraction of workers in the sample were employed in April 2009? Use your answer to compute a 95% confidence interval for the probability that a worker was employed in April 2009, conditional on being employed in April 2008.

b. Regress *Employed* on *Age* and Age^2, using a linear probability model.

i. Based on this regression, was age a statistically significant determinant of employment in April 2009?

ii. Is there evidence of a nonlinear effect of age on the probability of being employed?

iii. Compute the predicted probability of employment for a 20-year-old worker, a 40-year-old worker, and a 60-year-old worker.

c. Repeat (b) using a probit regression.

d. Repeat (b) using a logit regression.

e. Are there important differences in your answers to (b)–(d)? Explain.