計量經濟學電腦實習課

L4 IV, Qusi-Experiments

陳柏瑜 Chen Boyie

June 2, 2020

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

Brief

Instrument Variable EE12.2

Quasi Experiment EE13.1

Stata Homework 5 Announcement

Instrument Variable EE12.2

EE12.2

我們直接從課本的 Empirical Exercise 12.2 來複習如何用 Stata 處理內生性問題，並實作 2SLS。

EE12.2 Questions

E12.2 Does viewing a violent movie lead to violent behavior? If so, the incidence of violent crimes, such as assaults, should rise following the release of a violent movie that attracts many viewers. Alternatively, movie viewing may substitute for other activities (such as alcohol consumption) that lead to violent behavior, so that assaults should fall when more viewers are attracted to the cinema. On the text website, **http://www.pearsonglobaleditions.com**, you will find the data file **Movies**, which contains data on the number of assaults and movie attendance for 516 weekends from 1995 through 2004.[7] A detailed description is given in **Movies_Description**, available on the website. The data set includes weekend U.S. attendance for strongly violent movies (such as *Hannibal*), mildly violent movies (such as *Spider-Man*), and nonviolent movies (such as *Finding Nemo*). The data set also includes a count of the number of assaults for the same weekend in a subset of counties in the United States. Finally, the data set includes indicators for year, month, whether the weekend is a holiday, and various measures of the weather.

a. i. Regress the logarithm of the number of assaults [$ln\_assaults = \ln(assaults)$] on the year and month indicators. Is there evidence of seasonality in assaults? That is, do there tend to be more assaults in some months than others? Explain.

ii. Regress total movie attendance ($attend = attend\_v + attend\_m + attend\_n$) on the year and month indicators. Is there evidence of seasonality in movie attendance? Explain.

b. Regress *ln_assaults* on *attend_v*, *attend_m*, *attend_n*, the year and month indicators, and the weather and holiday control variables available in the data set.

i. Based on the regression, does viewing a strongly violent movie increase or decrease assaults? By how much? Is the estimated effect statistically significant?

ii. Does attendance at strongly violent movies affect assaults differently than attendance at moderately violent movies? Differently than attendance at nonviolent movies?

[7]These are aggregated versions of data provided by Gordon Dahl of University of California–San Diego and Stefano DellaVigna of University of California–Berkeley and were used in their paper "Does Movie

EE12.2 Questions

iii. A strongly violent blockbuster movie is released, and the weekend's attendance at strongly violent movies increases by 6 million; meanwhile, attendance falls by 2 million for moderately violent movies and by 1 million for nonviolent movies. What is the predicted effect on assaults? Construct a 95% confidence interval for the change in assaults. [*Hint:* Review Section 7.3 and material surrounding Equations (8.7) and (8.8).]

c. It is difficult to control for all the variables that affect assaults and that might be correlated with movie attendance. For example, the effect of the weather on assaults and movie attendance is only crudely approximated by the weather variables in the data set. However, the data set does include a set of instruments—*pr_attend_v*, *pr_attend_m*, and *pr_attend_n*—that are correlated with attendance but are (arguably) uncorrelated with weekend-specific factors (such as the weather) that affect both assaults and movie attendance. These instruments use historical attendance patterns, not information on a particular weekend, to predict a film's attendance in a given weekend. For example, if a film's attendance is high in the second week of its release, then this can be used to predict that its attendance was also high in the first week of its release. (The details of the construction of these instruments are available in the Dahl and DellaVigna paper referenced in footnote 5.) Run the regression from (b) (including year, month, holiday, and weather controls) but now using *pr_attend_v*, *pr_attend_m*, and *pr_attend_n* as instruments for *attend_v*, *attend_m*, and *attend_n*. Use this IV regression to answer (b)(i)–(b)(iii).

EE12.2 Questions

d. The intuition underlying the instruments in (c) is that attendance in a given week is correlated with attendance in surrounding weeks. For each movie category, the data set includes attendance in surrounding weeks. Run the regression using the instruments *attend_v_f*, *attend_m_f*, *attend_n_f*, *attend_v_b*, *attend_m_b*, and *attend_n_b* instead of the instruments used in (c). Use this IV regression to answer (b)(i)–(b)(iii).

e. There are nine instruments listed in (c) and (d), but only three are needed for identification. Carry out the test for overidentification summarized in Key Concept 12.6. What do you conclude about the validity of the instruments?

f. Based on your analysis, what do you conclude about the effect of violent movies on (short-run) violent behavior?

EE12.2 Data Descrptions

Movie Data

1. Observations: 516 weekends
2. Time Period : 1995-2004

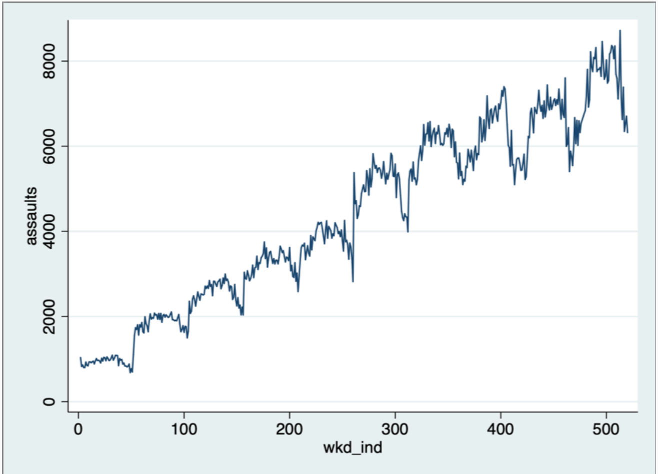
Variable Name	Description
Assaults and Movie Attendance	
assaults	number of assaults and intimidation in a subset of U.S. counties
attend_v	attendance stongly violent movies (in millions)
attend_m	attendance mildly violent movies (in millions)
attend_n	attendance nonviolent movies (in millions)
Weather, Holiday and Calendar Variables	
year1 to year10	indicator variable for year of the sample (1995-2004)
month1 to month12	indicator variables for month of the year (January-December)
h_chris	indicator variable for Christmas weekend
h_newyr	indicator variable for New Years weekend
h_easter	indicator variable for Easter weekend
h_july4	indicator variable for July 4 (U.S. Independence Day) weekend
h_mem	indicator variable for Memorial Day weekend
h_labor	indicator variable for Labor Day weekend
w_rain	fraction of locations with rain
w_snow	fraction of locations with snow
w_maxa	fraction of locations with maximum daily temperature between 80ºF and 90ºF
w_maxb	fraction of locations with maximum daily temperature between 90ºF and 100ºF
w_maxc	fraction of locations with maximum daily temperature greater than 100ºF
w_mina	fraction of locations with minimum daily temperature less than 10ºF
w_minb	fraction of locations with minimum daily temperature between 10ºF and 20ºF
w_minc	fraction of locations with minimum daily temperature between 20ºF and 32ºF
Instruments	
pr_attend_v	predicted attendance violent movies
pr_attend_m	predicted attendance moderately violent movies
pr_attend_n	predicted attendance nonviolent movies
attend_v_f	attendance violent movies one week in the future
attend_m_f	attendance moderately violent movies one week in the future
attend_n_f	attendance nonviolent movies one week in the future
attend_v_b	attendance violent movies one week in the past
attend_m_b	attendance moderately violent movies one week in the past
attend_n_b	attendance nonviolent movies one week in the past

EE12.2 a(i.)

To detect whether there is time trend or not:

$$log(assaults) = \beta_0 + \psi_1 year + \psi_2 month + u$$

Or just simply graph a twoway plot with the time indicator.

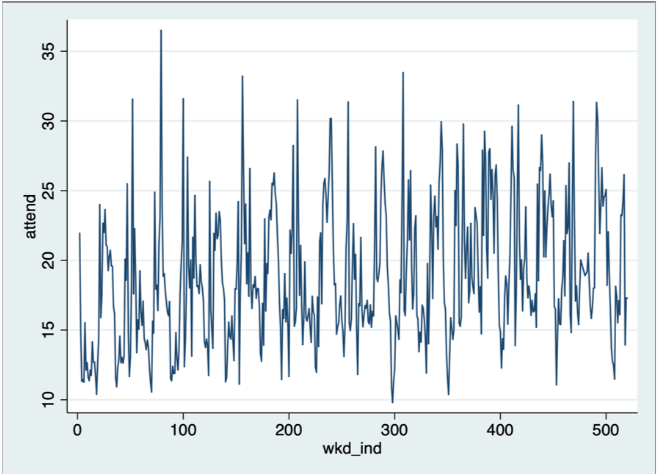


EE12.2 a(ii.)

To detect whether there is time trend or not:

$$attendance = \beta_0 + \phi_1 year + \phi_2 month + \tilde{u}$$

Or just simply graph a twoway plot with the time indicator.



EE12.2 b

The original model is:

$$\mathit{log}(\mathit{assaults}) = \beta_0 + \beta_1 \mathit{attend\_v} + \beta_2 \mathit{attend\_m} + \beta_3 \mathit{attend\_n} + \phi \mathit{Time} + \psi \mathit{Controls}$$

And the variables of interest are `attend_v`, `attend_m`, `attend_n`

Till now, we do not say the variables of interest are exog. or endog.

EE12.2 b(iii.)

The original model is:

$$log(assaults) = \beta_0 + \beta_1 attend\_v + \beta_2 attend\_m + \beta_3 attend\_n + \phi Time + \psi Controls$$

What we want to know is the change in the dependent variable: $\Delta assaults$.

Let's start from:

$$\Delta log(assaults) = \beta_1 \Delta attend\_v + \beta_2 \Delta attend\_m + \beta_3 \Delta attend\_n$$

Then,

$$\Delta log(\widehat{assaults}) = \hat{\beta}_1 \Delta attend\_v + \hat{\beta}_2 \Delta attend\_m + \hat{\beta}_3 \Delta attend\_n$$

With the description in the questions:

$$\Delta log(\widehat{assaults}) = 6\hat{\beta}_1 - 2\hat{\beta}_2 - \hat{\beta}_3$$

EE12.2 b(iii.)

$$\Delta log(\widehat{assaults}) = 6\hat{\beta}_1 - 2\hat{\beta}_2 - \hat{\beta}_3$$

And we can easily calculate the estimate: $\Delta log(\widehat{assaults}) = -.01063206$

That is: $\Delta \widehat{assaults} \approx e^{-.01063206}$

EE12.2 b(iii.) conti.

Given $\Delta log(\widehat{assaults}) = 6\hat{\beta}_1 - 2\hat{\beta}_2 - \hat{\beta}_3$

To obtain $se(\Delta log(\widehat{assaults})) = se(6\hat{\beta}_1 - 2\hat{\beta}_2 - \hat{\beta}_3)$,

we need to apply approaches in section 7.3.

We may simplify the model by:

$$y = \beta_0 + \beta_1 v + \beta_2 m + \beta_3 n + u$$

Now let's focus on $\beta_1 v + \beta_2 m + \beta_3 n$ only.

Our goal is to have a explanatory variable which has a coefficient equals to the above number.

EE12.2 b(iii.) conti.

Simplified model:

$$y = \beta_0 + \beta_1 v + \beta_2 m + \beta_3 n + u$$

$$y = \beta_1 v + \beta_2 m + \beta_3 n - 6\beta_1 n + 2\beta_2 n + 6\beta_1 n - 2\beta_2 n + u$$

$$y = \beta_1(v + 6n) + \beta_2(m - 2n) + (\beta_3 - 6\beta_1 + 2\beta_2)n + u$$

Now we can simply use OLS to obtain the s.e.

EE12.2 c

Now we want to use IVs.

Notations:

$Y: dependentvariable$

$X: possiblyendog.expanatoryvariables$

$W: exog.expanatoryvariables$

$Z: instrumentvariables$

Recall the 2SLS:

- Regress X on Z, W
- Obtain $\hat{X}$
- Regress Y on $\hat{X}, W$
- Obtain the coef. of $\hat{X}$

EE12.2 c

Denote

$Y: log(assaults)$

$X: attend\_v, attend\_m, attend\_n$

$W: Time, Holiday, Weather$

$Z: pr\_attend\_v, pr\_attend\_m, pr\_attend\_n$

which are predictions based on historical attendance patterns.

THINK: Why are these variable instruments?

We'll demonstrate the manual 2SLS approach first.

EE12.2 c

We may simply use `ivreg` or `ivregress`

The format would be:

```
ivreg Y (X=Z) W, r
```

or:

```
ivregress 2sls Y (X=Z) W, r
ivregress gmm Y (X=Z) W, r
```

EE12.2 d

Now change the IVs:

$$Z: attend\_v\_f, attend\_m\_f, attend\_n\_f,$$

attend_v_b, attend_m_b, attend_n_b

which are the lagged terms and the future terms.

EE12.2 e

For Overidentifying Restriction Test, we first apply J-Test in textbook p.449, then we'll demonstrate a simple and more general command in Stata.

Note that the model now is:

$Y = \beta X + \gamma W + u$ This is (12.12) in p.438

Denote the residuals obtained in the above regression as $\hat{u}_{2SLS}$.

Then we regress $\hat{u}_{2SLS}$ on Z, W, that is:

$$\hat{u}_{2SLS} = \delta Z + \tilde{\gamma} W + e$$

EE12.2 e

And we can obtain the J-statistic from $J = mF$ where

$m =$ the number of instruments

$F =$ the joint test statistic of $\delta = 0$

$k =$ the number of exog. variables

Under homosk., $J \sim \chi^2(m - k)$

EE12.2 e

An easy command is:

```
ivregress Y (X=Z) W, r
estat overid
```

will give you the test statistic under robust variance covariance estimator.

Quasi Experiment EE13.1

EE12.2

我們直接從課本的 Empirical Exercise 13.1 來複習如何用 Stata 實作 quasi-experimental analysis，並做出解釋。

EE13.1 Questions

E13.1 A prospective employer receives two resumes: a resume from a white job applicant and a similar resume from an African American applicant. Is the employer more likely to call back the white applicant to arrange an interview? Marianne Bertrand and Sendhil Mullainathan carried out a randomized controlled experiment to answer this question. Because race is not typically included on a resume, they differentiated resumes on the basis of "white-sounding names" (such as Emily Walsh or Gregory Baker) and "African American–sounding names" (such as Lakisha Washington or Jamal Jones). A large collection of fictitious resumes was created, and the presupposed "race" (based on the "sound" of the name) was randomly assigned to each resume. These resumes were sent to prospective employers to see which resumes generated a phone call (a call-back) from the prospective employer. Data from the experiment and a detailed data description are on the text website, **http://www.pearsonglobaleditions.com**, in the files **Names** and **Names_Description**.[8]

EE13.1 Questions

a. Define the *callback rate* as the fraction of resumes that generate a phone call from the prospective employer. What was the callback rate for whites? For African Americans? Construct a 95% confidence interval for the difference in the callback rates. Is the difference statistically significant? Is it large in a real-world sense?

b. Is the African American/white callback rate differential different for men than for women?

c. What is the difference in callback rates for high-quality versus low-quality resumes? What is the high-quality/low-quality difference for white applicants? For African American applicants? Is there a significant difference in this high-quality/low-quality difference for whites versus African Americans?

d. The authors of the study claim that race was assigned randomly to the resumes. Is there any evidence of nonrandom assignment?

EE13.1 Data Descrptions

Names Data

1. Observations: 4870 resumes
2. Time Period : 2001

Variable Descriptions

Variable Name	Description
	Key Variables
firstname	applicant's first name
female	1 = female
black	1 = black
high	1= high quality resume
call_back	1= applicant was called back
chicago	1 = data from Chicago
	Detailed Information on Resume
ofjobs	number of jobs listed on resume
yearsexp	number of years of work experience on the resume
honors	1=resume mentions some honors
volunteer	1=resume mentions some volunteering experience
military	1=applicant has some military experience
empholes	1=resume has some employment holes
workinschool	1=resume mentions some work experience while at school
email	1=email address on applicant's resume
computerskills	1=resume mentions some computer skills
specialskills	1=resume mentions some special skills
college	applicant has college degree or more
	Detailed Information Concerning Employer
expminreq	min experience required, if any
eoe	1=ad mentions employer is EOE
manager	1=manager wanted
supervisor	1=supervisor wanted
secretary	1=secretary wanted
offsupport	1=office support
salesrep	1=sales representative wanted
retailsales	1=retail sales worker wanted
req	1=ad mentions any requirement for job
expreq	1=ad mentions some experience requirement
comreq	1=ad mentions some communication skills requirement
educreq	1=ad mentions some educational requirement
compreq	1=ad mentions some computer skill requirement
orgreq	1=ad mentions some organizational skills requirement
manuf	1=employer industry is manufacturing
transcom	1=employer industry is transport/communication
bankreal	1=employer industry is finance, insurance, real estate
trade	1=employer industry is wholesale or retail trade
busservice	1=employer industry is business and personal services
othservice	1=employer industry is health, educ. and social services
missind	1=employer industry is other/unknown

EE13.1 a

The command:

```
mean callback
```

gives us the fraction of resumes that generate a phone call from prospective employer.

The expression:

```
ci mean callback
```

gives us the same result.

EE13.1 a

For whites, simply add a condition:

```
ci mean callback if black==0
```

For blacks:

```
ci mean callback if black==1
```

EE13.1 a conti.

To understand how Stata gives us the 95% C.I., recall the point estimator in the previous semester.

Let

$$y_i = \begin{cases} 1, \text{if receive phone call} \\ 0, \text{o.w.} \end{cases}$$

Clearly, y_i follows $Bernoulli(p)$ where p is the unknown fraction of resumes that generate a phone call from prospective employer.

Naturally, we would like to use $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ to estimate p. Where $n = 4870$.

We need to know $se(\bar{y})$ in order to build up a interval estimator.

EE13.1 a conti.

Now, our concern would be: What is the sampling distribution of $\bar{y}$?

Given

$$y_i \overset{i.i.d.}{\sim} Bernoulli(p)$$

then

$$\sum_{i=1}^{n} y_i \sim Binomial(n, p)$$

which can be approximated by normal distribution

$$\sum_{i=1}^{n} y_i \overset{a}{\sim} N(np, np(1-p))$$

as $n \to \infty$. And let $X = \frac{1}{n}\sum_{i=1}^{n} y_i$, then

$$X \sim N(p, \frac{p(1-p)}{n})$$

That is, $se(\bar{y})$ can be calculated easily from the approximated normal distribution.

EE13.1 a conti.

With the normal quantile $Z_{0.05} = -1.96$, we know the interval estimator for *p* is:

$[\bar{y} - 1.96 \times se(\bar{y}), \bar{y} + 1.96 \times se(\bar{y})]$

And we know
$se(\bar{y}) \approx (0.0804928(1 - .0804928)/4870)^{\frac{1}{2}} = 0.003898446728$

Which is nearly the `Std. Err.` calculated by Stata.

After that, we may construct the C.I. under every given quantile.

EE13.1 a conti.

Some may know that the command:

`ci proportion callback`

gives us a very similar result, and wonder the difference between the two.

Actually, you may see the text "Binomial Exact" in the latter command.

That is, Stata calculate the exact sampling distribution from "categories" we're interested in, instead of using the approximated normal distribution.

```
. ci proportion call_back
```

Variable	Obs	Proportion	Std. Err.	— Binomial Exact — [95% Conf.	Interval]
call_back	4,870	.0804928	.0038984	.0730025	.0884904

EE13.1 a conti.

Also, if the variable is binary, and it is valued at 1 and 0, then the two command will yield the same result as the number of observations is large.

And the option `proportion` is designed for category variables. You may notice that there are Binomial distribution, Trinomial distribution, and Multinomial distribution.

For example,

$$(X, Y) \sim Trinomial(n, p_1, p_2)$$

$$f_{XY}(x, y) = \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y (1 - p_1 - p_2)^{n-x-y}$$

EE13.1 b

Given the model:

$$\textit{call\_back} = \beta_0 + \beta_1 \textit{black} + \beta_2 \textit{female} + \beta_3 \textit{blackfemale} + u$$

If *black* $= 0$ & *female* $= 0$, then the effect is β_0

If *black* $= 1$ & *female* $= 0$, then the effect is $\beta_0 + \beta_1$

If *black* $= 0$ & *female* $= 1$, then the effect is $\beta_0 + \beta_2$

If *black* $= 1$ & *female* $= 1$, then the effect is $\beta_0 + \beta_1 + \beta_2 + \beta_3$

EE13.1 c

Given the model:

$$call\_back = \beta_0 + \beta_1 black + \beta_2 high + \beta_3 blackhigh + u$$

If $black = 0$ & $high = 0$, then the effect is β_0

If $black = 1$ & $high = 0$, then the effect is $\beta_0 + \beta_1$

If $black = 0$ & $high = 1$, then the effect is $\beta_0 + \beta_2$

If $black = 1$ & $high = 1$, then the effect is $\beta_0 + \beta_1 + \beta_2 + \beta_3$

EE13.1 Table

	(1)	(2)	(3)
VARIABLES	a	b	c
black	-0.0320***	-0.0304*	-0.0231**
	(0.00778)	(0.0155)	(0.0106)
female		0.0102	
		(0.0137)	
blackFemale		-0.00224	
		(0.0179)	
high			0.0229*
			(0.0120)
blackHigh			-0.0178
			(0.0156)
Constant	0.0965***	0.0887***	0.0850***
	(0.00599)	(0.0119)	(0.00801)
Observations	4,870	4,870	4,870
R-squared	0.003	0.004	0.004

Robust standard errors in parentheses

Stata Homework 5 Announcement

HW

1. Textbook Empirical Exercise 12.1 (a.)-(f.)
2. Notice that (g.) is not included.

HW 格式要求

1. Upload only one pdf file.
2. All formula should be expressed in LaTeX format.
3. Deadline is 6/2 Tue. 14:10

Example Answers

Check Stata handout.

EE12.1 Questions

E12.1 How does fertility affect labor supply? That is, how much does a woman's labor supply fall when she has an additional child? In this exercise, you will estimate this effect using data for married women from the 1980 U.S. Census.[6] The data are available on the text website, **http://www.pearsonglobaleditions.com**, in the file **Fertility** and described in the file **Fertility_Description**. The data set contains information on married women aged 21–35 with two or more children.

EE12.1 Questions

a. Regress *weeksworked* on the indicator variable *morekids*, using OLS. On average, do women with more than two children work less than women with two children? How much less?

b. Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (*morekids*) on labor supply (*weeksworked*).

c. The data set contains the variable *samesex*, which is equal to 1 if the first two children are of the same sex (boy–boy or girl–girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child? Is the effect large? Is it statistically significant?

d. Explain why *samesex* is a valid instrument for the IV regression of *weeksworked* on *morekids*.

---

[6]These data were provided by Professor William Evans of the University of Maryland and were used in his paper with Joshua Angrist, "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review*, 1998, 88(3): 450–477.

e. Is *samesex* a weak instrument?

f. Estimate the IV regression of *weeksworked* on *morekids*, using *samesex* as an instrument. How large is the fertility effect on labor supply?

g. Do the results change when you include the variables *agem1*, *black*,