Econometrics II: Econometric Modelling

Jürgen Meinecke

Research School of Economics, Australian National University

21 September, 2018

Welcome Back!

I hope you guys had a relaxing and energizing break!

Assignment 1

You can collect your assignments today during your tutorials

You can keep your assignment and take it with you

However, if you want to qualify for a remark you need to:

- raise and explain any concerns regarding the marking with your tutor as soon as possible during today's tutorial
- hand your assignment back to your tutor by the end of today's tutorial

Once you leave the tutorial room with your assignment, you cannot as for a remark

Assignment 1 (con't)

Uncollected assignments end up on my desk, if you want to collect your assignment after today you will need to see me (just come by or send me an e-mail for an appointment)

If you want to be able to get a remark on an uncollected assignment, please collect it from me before 5 October

Assignments collected later than that cannot get a remark!

Assignment 2

Assignment 2 is already available online
(deadline: 17 October, 12:00pm very sharp!)

Exclusively covers binary dependent variable models
(covered in lectures 9 and 10)

Two exercises:

1. empirical application
2. reading and disseminating a research paper
   (you could already start this exercise now)

Roadmap for next 6 Lectures

- 2 weeks of randomized controlled trials
- 2 weeks of binary dependent variables models
- 2 weeks of panel data models

Roadmap

Remember this important result from lecture 2:

The OLS estimator is internally valid only under OLS Assumption 1.

We have obsessed about OLS Assumption 1 for the most part of EMET2007 and EMET3004

Let's obsess some more…

Recall from lecture 5 of EMET2007:

Assumption (OLS Assumption 1)

The error term u_i is conditionally mean independent (CMI) of X_i

$$E[u_i|X_i] = E[u_i] = \mu_u \quad (= 0)$$

Assumption 1 says that X_i is not informative about the expected value of u_i

We said that this would be the case if u_i can be regarded as purely random

Since lecture 5 of EMET2007 we have learned a lot about the error term u_i

For example, u_i may contain omitted variables that are still correlated with X_i

That would directly break OLS Assumption 1

In other words, we have realized during the past few weeks that u_i can be a complicated object and it is not always easy to wrap our heads around it

That also means that we are often not in a position to simply state that u_i is "purely random"

If we don't have a good concept of u_i then we cannot simply state that it is a random error term

If we cannot trust u_i to be random, what else can we do?

There is a really compelling answer to that:
Make X_i random instead!

If, by some magic, X_i is random then u_i will be conditionally mean independent of X_i by construction
(and therefore OLS Assumption 1 holds)

If X_i is truly random, then it does not matter whether u_i contains omitted variables

Because X_i assigned randomly, all other individual characteristics – the things that make up u_i – are distributed independently of X_i, so u_i and X_i are independent

What do I mean by “making X_i random”?

Typical examples come from medical research:
Say you want to study the effect of a new medication on people’s blood pressure

There are two principal ways to go about it:

1. observational study
2. randomized controlled trial

Let’s look at these in turn

Observational Study

This is what we have done, essentially, throughout EMET2007 and EMET3004

As the medical researcher you go out and collect data on people who may be prone to blood pressure problems

You hand out questionnaires and ask them important questions about age, gender, race, medical history and whatever else you deem important

Most important, you also measure their blood pressure and ask them if they take blood pressure medication

You can then run a regression of blood pressure against a blood pressure medication dummy variable (plus a lot of other variables on age, gender, race, etc.)

Crossing your fingers, you declare your estimate on the blood pressure medication dummy variable to capture the causal effect of blood pressure medication on blood pressure

Here's the problem:
your colleagues will ridicule your for poor research

There's a plethora of endogeneity problems with this approach

(By now this should be too obvious to everyone in this course!)

Randomized Controlled Trial

Your colleagues will tell you to run an RCT instead

You randomly sample 1,000 subjects that have blood pressure problems

You randomly divide your subjects in two groups:

- 500 subjects randomly assigned to "treatment group"
  They receive the actual blood pressure medication
- 500 subjects randomly assigned to the "control group"
  They receive a "placebo", that is, a fake medication that has no effect whatsoever (and the subjects are made to believe that they receive the actual medication)

After some time (let's say one year) you compare the blood pressure levels between the treatment and control groups

This simple comparison of treatment and control groups at the end of the trial gives you the best estimate of the causal effect of the blood pressure medication

Technically, there are two ways of estimating this:

1. Simple comparison of sample averages
   Compute the sample average of blood pressure for the treatment and the control groups and compare the two
2. Using OLS: $Y_i = \beta_0 + \beta_1 X_i + u_i$,
   where Y_i is blood pressure and X_i is a dummy variable that is equal to one if person i received the actual treatment;
   OLS estimator $\hat{\beta}_1$ will be the causal effect estimate

Both ways of estimating will yield exactly the same result
(I hope this surprises nobody!)

Using OLS preferable because Stata will also spit our standard errors and confidence intervals for the causal effect plus we can include additional variables on the rhs

Also, the OLS model demonstrates nicely that OLS Assumption 1 is satisfied:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

What could be hidden inside u_i?

- age, gender, race, overall healthiness, fitness, etc.
- but all these things will be uncorrelated with X_i because X_i was assigned randomly
- in other words: both treatment and control groups will have comparable distributions of age, gender, race, healthiness, fitness, etc.

In economics, RCT are becoming more popular, but examples in economics are quite different to medical research

- Job training program
  - Y = has a job (yes or no)
  - X = went through training program (yes or no)
- Class size effect
  - Y = test score (Stanford Achievement Test)
  - X = class size treatment group
- Job discrimination
  - Y= gets a job interview (yes or no)
  - X= has Australian sounding name (yes or no)
- Bus driver discrimination
  This one was recently in the local news in Brisbane; UQ researcher sent grad students on local buses pretending to have forgotten their bus pass
  - Y= gets a free ride on Brisbane bus (yes or no)
  - X= looks foreign (yes or no)

Roadmap

In a randomized controlled trial (RCT) the ATE can be estimated by OLS or TSLS

Whether you use OLS or TSLS depends on the precise way you assign the treatment

There are two ways:

1. You randomize treatment
2. You randomize *eligibility* for treatment

What's the difference?

Randomizing treatment

You offer treatment to people (randomly) and somehow find an credible way of withholding treatment from the control group

The obvious example of such a case is a double-blind medical study in which medication is offered to subjects (some are placebos)

That's what most people would want from an RCT

Randomizing eligibility for treatment

Sometimes you cannot always persuade people that the treatment is good for them

Especially in economics this is often the case

Example: treating unemployed people with job training

As the researcher, you cannot randomly assign that treatment and make sure that people will actually "take" the treatment

Notice difference to medical example from previous slide:
If you have a sick patients and you offer them a potentially life-saving treatment, they will not say no

But if you offer a job training program to an unemployed person, they may tell you to mind your own business

So in the case of treating unemployed with job training, what can you do?

Re-define the treatment: instead of saying to an unemployed person that they will be treated with job training (which they may laugh off) you tell them that they are *eligible* for job training and they then can decide whether or not to take up the offer

The treatment then becomes *eligibility* for job training

Notice that whether or not a person actually receives the treatment is not randomly assigned (because people can choose, after being made eligible for the treatment, whether or not they are being treated)

Therefore, actually receiving the treatment may be an endogenous variable

In terms of estimation, these two different treatment approaches are quite different

1. Randomizing treatment
   - Outcome variable Y: blood pressure (for example)
   - Explanatory variable X: treatment dummy variable
   - Estimation: OLS of Y on X
2. Randomizing eligibility for treatment
   - Outcome variable Y: employment status (for example)
   - Explanatory variable X:
     dummy if the actually received the treatment
   - Instrumental variable Z:
     dummy if person was eligible for treatment
   - Estimation: TSLS of Y on X using instrument Z
   - This will estimate the effect of actually receiving the treatment (which is not randomly assigned) on employment outcomes using the randomly assigned treatment eligibility

Roadmap

Randomizing Treatment: STAR Experiment

- 4-year study conducted in Tennessee in late 1980s, $12 million
- Upon entering the school system, a student was randomly assigned to one of three groups:
  - regular class (22 - 25 students)
  - regular class + aide (helps out teacher)
  - small class (13 - 17 students)
- teachers were also randomly assigned to class sizes (within their schools)
- STAR project paid for the additional teachers needed to guarantee small classes
- Over all 4 years, approximately 11,600 students participated at 80 schools
- Y = Stanford Achievement Test scores

- The regression model:

$$Y_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 RegAide_i + u_i$$

  where,

$$SmallClass_i = 1 \text{ if in a small class}$$

$$RegAide_i = 1 \text{ if in regular class with aide}$$

- Additional regressors (W's):
  - teacher experience,
  - free lunch eligibility,
  - gender and race

OLS estimates (no additional regressors)

TABLE 13.1 Project STAR: Differences Estimates of Effect on Standardized Test Scores of Class Size Treatment Group

	Grade			
Regressor	**K**	**1**	**2**	**3**
Small class	13.90** (2.45)	29.78** (2.83)	19.39** (2.71)	15.59** (2.40)
Regular size with aide	0.31 (2.27)	11.96** (2.65)	3.48 (2.54)	−0.29 (2.27)
Intercept	918.04** (1.63)	1039.39** (1.78)	1157.81** (1.82)	1228.51** (1.68)
Number of observations	5786	6379	6049	5967

The regressions were estimated using the Project STAR Public Access Data Set described in Appendix 13.1. The dependent variable is the student's combined score on the math and reading portions of the Stanford Achievement Test. Standard errors are given in parentheses under the coefficients. **The individual coefficient is statistically significant at the 1% significance level using a two-sided test.

TABLE 13.2 Project STAR: Differences Estimates with Additional Regressors for Kindergarten

Regressor	(1)	(2)	(3)	(4)
Small class	13.90** (2.45)	14.00** (2.45)	15.93** (2.24)	15.89** (2.16)
Regular size with aide	0.31 (2.27)	−0.60 (2.25)	1.22 (2.04)	1.79 (1.96)
Teacher's years of experience		1.47** (0.17)	0.74** (0.17)	0.66** (0.17)
Boy				−12.09** (1.67)
Free lunch eligible				−34.70** (1.99)
Black				−25.43** (3.50)
Race other than black or white				−8.50 (12.52)
Intercept	918.04** (1.63)	904.72** (2.22)		
School indicator variables?	no	no	yes	yes
$\overline{R}^2$	0.01	0.02	0.22	0.28
Number of observations	5786	5766	5766	5748

The regressions were estimated using the Project STAR Public Access Data Set described in Appendix 13.1. The dependent variable is the combined test score on the math and reading portions of the Stanford Achievement Test. The number of observations differ in the different regressions because of some missing data. Standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

Randomizing Eligibility: Job Counseling

Paper by Behagel, Crépon and Gurgand in *American Economic Journal: Applied Economics*, 2014

Newly unemployed job seekers in France have three options:

- Use standard public job counseling
- Use intensive public job counseling
- Use intensive private job counseling

Main difference b/w standard and intensive job counseling: Caseload of counsellers much reduced (40 unemployed vs 120)

Main difference b/w intensive public and private: Private counselors get incentive payments for successful placement

Newly unemployed were randomly made *eligible* to one of the three options

Main variables:

- Z_{1i}: person i eligible intensive public counseling
- Z_{2i}: person i eligible intensive private counseling
- Z_{3i}: person i eligible standard public counseling
- X_{1i}: person i received intensive public counseling
- X_{2i}: person i received intensive private counseling
- X_{3i}: person i received standard public counseling
- Y: employment outcome after 6 months

Each person will be eligible for exactly one: $Z_{1i} + Z_{2i} + Z_{3i} = 1$

Each person will receive exactly one: $X_{1i} + X_{2i} + X_{3i} = 1$
(if a person rejects intensive counseling they automatically move to standard counseling)

Authors first estimate the so-called *intention to treat* (ITT) effect:

$$Y_i = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + w_i$$

Here are the estimates: $\hat{\alpha}_1 = 0.022$ and $\hat{\alpha}_2 = 0.020$

In words: being eligible for treatment in the intensive public program increases the chances of finding a long-lasting job within 6 months by 2.2 percentage points

Being eligible for the intensive private program increases chances by 2 percentage points

What does the ITT capture?
ITT measures potential effect of treatment (it was our intention to treat everyone who was made eligible for treatment)

ITT underestimates ATE (actual treatment effect) (why?)

Next, the authors estimate the actual ATE via TSLS:

$$\text{Main equation:} \quad Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

$$\text{First stage:} \quad X_{1i} = \pi_0 + \pi_1 Z_{1i} + \pi_2 Z_{2i} + v_{1i}$$

$$X_{2i} = \phi_0 + \phi_1 Z_{1i} + \phi_2 Z_{2i} + v_{2i}$$

The authors get the following TSLS estimates:
$\hat{\beta}_1 = 0.072$ and $\hat{\beta}_2 = 0.050$

In words: actually participating in the intensive public program increases the chances of finding a long-lasting job within 6 months by 7.2 percentage points

The intensive private program improves chances by 5 percentage points