EC 3303: Econometrics I

Instrumental Variables

Offers Z_i $\rightarrow$ Attendance X_i $\rightarrow$ Achievement Y_i

Kelvin Seah

AY 2022/2023, Semester 2

Outline

1. Instrumental Variables
2. Application: Causal Effect of Attending a Charter School
3. IV in Practice

Limitations of Fixed Effects

- Fixed effects eliminates bias in the estimator of the coefficient of interest which arise from unobserved variables that:

    - vary from entity to entity but are constant over time

- If the unobserved variables vary across both entity & time, an alternative method is needed.

Instrumental Variables

- One way to obtain a consistent estimator of the effect of interest whenever X_i is correlated with the error term u_i is to use "*Instrumental Variables*".

- IV methods work regardless of the source of correlation between X_i & u_i.

- It works whether the reason for the correlation between X_i & u_i is due to:

    - omitted variables

    - measurement errors in the regressor

    - simultaneous causality

Problem

Regressor of interest, X, is correlated with error term, u.

How IV methods work

- Think of the variation in X as having 2 parts:
  - One part that is correlated with u – problematic part.
  - One part that is uncorrelated with u – part that we want.
  - IV works by providing information to extract those variations in X that are uncorrelated with u.
  - So IVs permit consistent estimation of β.

Single Regressor, Single Instrument

- Consider the most basic case: single regressor X & single instrument Z.

  Regression model:

$$Y_i = \alpha + \beta X_i + u_i$$

- Suppose X_i & u_i are correlated (X_i is "endogenous").
  - Then, $\hat{\beta} \cancel{\xrightarrow{p}} \beta$
- If we are able to find an IV, Z_i , that fulfils 2 conditions:

  a. *Instrument relevance*: $Corr(Z_i, X_i) \neq 0$

  b. *Instrument exogeneity*: $Corr(Z_i, u_i) = 0$

  then we can estimate β consistently

$$\hat{\beta} \xrightarrow{p} \beta$$

a. *Instrument relevance*: $Corr(Z_i, X_i) \neq 0$

- If Z_i is relevant, this means that variation in Z_i is related to variation in X_i.

- Z_i must have a *causal effect* on X_i.

If *in addition*,

b. *Instrument exogeneity*: $Corr(Z_i, u_i) = 0$

- Z_i is exogenous, then the part of the variation in X_i captured by Z_i will be uncorrelated with u_i.

2SLS Estimator

- If an instrument, Z, satisfies both conditions, we can use an IV estimator known as two stage least squares (2SLS) to estimate β consistently.
- 2SLS estimator is calculated in 2 stages

<u>Stage 1</u>

- Decompose X into 2 components:

    1. Problem free component – which is uncorrelated with u
    2. Problematic component – which is correlated with u

- Mathematically,

$$X_i = \pi_0 + \pi_1 Z_i + v_i \quad (1)$$

$$X_i = \pi_0 + \pi_1 Z_i + v_i \quad (1)$$

- X_i is broken into 2 components:

$$\pi_0 + \pi_1 Z_i$$

  - This is the part of X_i that can be predicted by Z_i (part which is uncorrelated with u_i).

$$v_i$$

  - This is the part of X_i that is problematic (part which is correlated with u_i).

Use the problem free component

$$\pi_0 + \pi_1 Z_i$$

and disregard the problematic component

$$v_i$$

- π_0 and π_1 are unknown coefficients & must be estimated first using OLS.
- To do this, apply OLS to equation (1): $X_i = \pi_0 + \pi_1 Z_i + v_i$ (1)
  - this gives $\hat{\pi}_0$ and $\hat{\pi}_1$
  - so we can obtain the predicted value of X_i:
  $$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$
- First stage of 2SLS involves using the instrument Z_i to create predicted values of the regressor of interest $\hat{X}_i$.

Stage 2

- Regress Y_i on $\hat{X}_i$ using OLS

- Resulting estimators from the 2^{nd} stage regression are the 2SLS estimators.

estat firststage /*the F-statistic from the first stage regression shows how relevant the instrument is. The rule of thumb is that the F-statistic should exceed 10. Otherwise, the instrument is "weak" and 2sls will lead to unreliable estimates.*/

IV Model with a Single Endogenous Regressor

- Can extend the previous model to incorporate multiple control variables (W's).

$$Y_i = \alpha + \beta X_i + \delta_1 W_{1i} + \delta_2 W_{2i} + \cdots + \delta_r W_{ri} + u_i \quad (2)$$

Y_i: outcome variable

β: unknown coefficient of interest

X_i: endogenous variable of interest

$\delta_1, \ldots, \delta_r$: unknown coefficients on each of the r control variables

$W_{1i}, \ldots, W_{ri}$: r control variables

u_i: error term

Z_i : instrumental variable

Application: Causal Effect of Attending a Charter School

- See how IV can work to estimate causal effects when there is only partial / imperfect random assignment.
- Charter schools are public schools which operate with autonomy.

Charter	Regular Public
1. Can structure curricular freely	1. Less flexibility to structure curricular
2. Run longer school days	2. Run shorter school days
3. Run school on weekends	3. Run school on weekdays only
4. Teachers rarely belong to unions	4. Most teachers are unionised
5. Selective teacher hiring	5. Less selective teacher hiring

Qn: *Does attending a charter school increase student achievement?*

- Angrist (2010) exploits a lottery influencing enrolment to a charter school to answer this question.

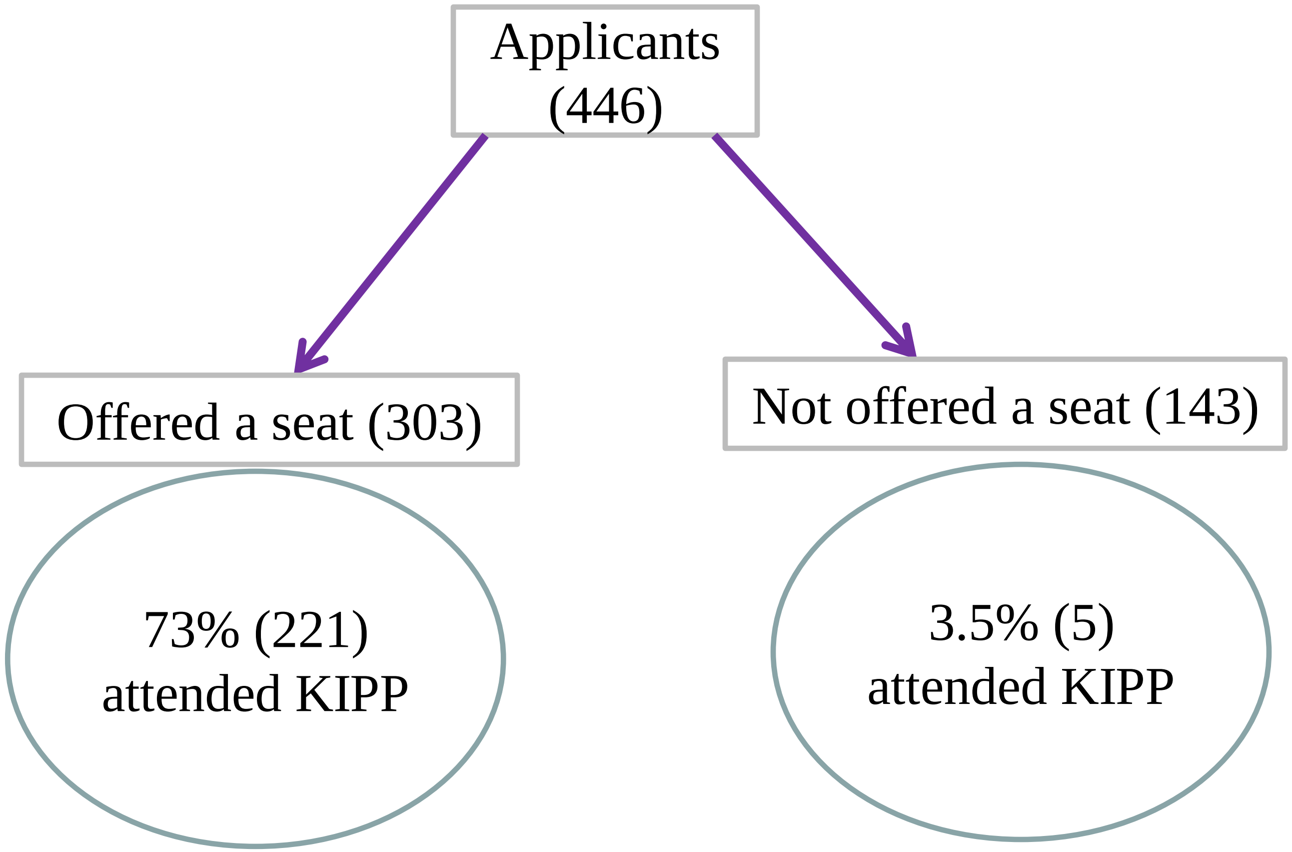
Background

- Focused on one typical charter school in Boston – KIPP.
- Charter had an excess demand for places.
- Law requires scarce charter seats to be allocated by lottery.
- Though seats were allocated randomly, the “experiment” was imperfect

i. some applicants who were offered a seat did not take it up.

ii. some applicants who were not offered a seat initially still wound up there eventually.

Application & Enrolment from KIPP Lotteries



- Applicants who won the lottery and who did not win are similar.

KIPP Applicants		
	Lottery Winners	Winners vs. Losers
Baseline characteristics		
Hispanic	0.510	-0.058 (0.058)
Black	0.257	0.026 (0.047)
Female	0.494	-0.008 (0.059)
Free Lunch	0.814	-0.032 (0.046)
Baseline (4th grade) Math score	-0.290	0.102 (0.120)
Baseline (4th grade) Reading score	-0.386	0.063 (0.125)

- Offer of a charter seat is indeed random, so unrelated to factors influencing student achievement.
- Denote Z_i (instrument): dummy variable, =1 if applicant is offered a seat, =0 otherwise

- However, applicants who actually attended KIPP (enrolled) and who did not attend (non-enrolled), may not be similar.
- Because actual enrolment is not entirely randomly assigned:
  - Lottery winners who chose to go elsewhere may care less about school.
  - Lottery losers who made it into KIPP may care more about the school.
  - So KIPP enrollees may care more about school than non-enrollees.
  - Comparisons of enrolled & non-enrolled likely to overestimate the positive effects of attending KIPP.
- Denote actual enrolment (treatment variable of interest) by X_i: =1 if student attended KIPP, =0 otherwise.
  - X_i is endogenous because it depends on u - factors influencing achievement (e.g. how much student cares about school).

- A regression of Y_i (student achievement) on X_i (enrolment in KIPP) would yield biased & inconsistent estimates of the true effect of charter attendance.

- Even if you include additional observable control variables, you will never be able to measure & include how much a student cares about school.

- Can use Z_i (lottery offer of a seat) to estimate the effect of charter attendance X_i consistently.

Check if the instrument meets the 2 conditions:

1. *Instrument relevance*: $Corr(Z_i, X_i) \neq 0$

- A successful lottery offer increases the probability of KIPP attendance.

- $Corr(Z_i, X_i) \neq 0$ is satisfied

2. *Instrument exogeneity*: $Corr(Z_i, u_i) = 0$

- Lottery offer of a seat is randomly assigned.

- So offer of a seat is unrelated to other factors influencing student achievement.

- $Corr(Z_i, u_i) = 0$ is satisfied

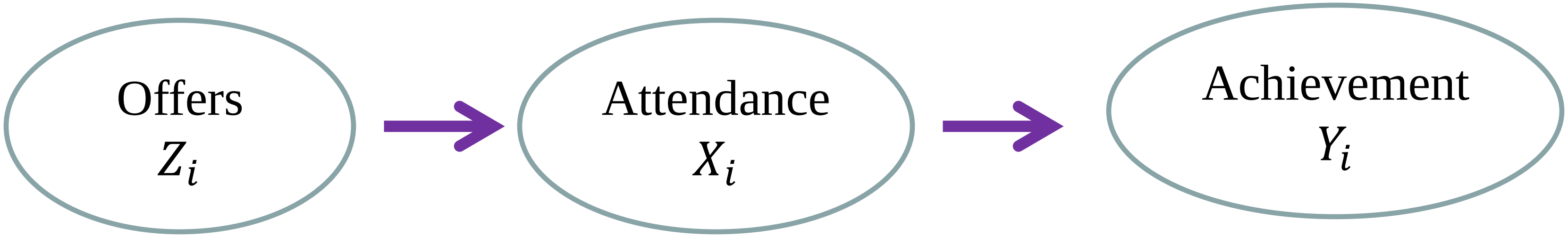
- IV method relies on a *chain reaction* leading from the instrument to the outcome.

Stage 1

- Connects randomly assigned offers with KIPP attendance

Stage 2 (what we are after)

- Connects KIPP attendance with student achievement



- For IV to be valid: The only reason Z_i affects Y_i is through its influence on X_i (exclusion restriction).

IV Estimation in STATA

- Wage & related data on 758 men.
- Variables:
  - lw: log of wage
  - s: years of schooling
  - expr: experience
  - rns: indicator for residency in the U.S. south
  - iq: worker's IQ score (mis-measured ability – so endogenous)
- Suppose, we have as instrument
  - med: mother's level of education

Griliches, Zvi (1976) Wages of Very Young Men. *Journal of Political Economy*, 84(2), S69-S86.

- Some descriptive statistics:
  - `set more off`
  - `use http://www.stata-press.com/data/imeus/griliches, clear`
  - `summarize lw s expr rns iq med`

```
. summarize lw s expr rns iq med
```

Variable	Obs	Mean	Std. Dev.	Min	Max
lw	758	5.686739	.4289494	4.605	7.051
s	758	13.40501	2.231828	9	18
expr	758	1.735429	2.105542	0	11.444
rns	758	.2691293	.4438001	0	1
iq	758	103.8562	13.61867	54	145
med	758	10.91029	2.74112	0	18

  - `keep lw s expr rns iq med`

```
. regress lw s expr rns iq, robust

Linear regression                               Number of obs     =        758
                                                F(4, 753)         =      85.63
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3215
                                                Root MSE          =     .35427
```

lw	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
s	.0939136	.0072657	12.93	0.000	.0796502	.1081771
expr	.0454811	.0064287	7.07	0.000	.0328609	.0581013
rns	-.0997054	.0300299	-3.32	0.001	-.1586577	-.040753
iq	.0037685	.0011556	3.26	0.001	.0014998	.0060371
_cons	3.98435	.116406	34.23	0.000	3.755831	4.212869

- To run IV regression, use the ivregress command.

```
ivregress 2sls depvar [varlist1] (varlist2 =
instrulist) [if] [, options]
```

`depvar` is the dependent variable

`varlist1` is the list of exogenous regressors (i.e. controls)

`varlist2` is the [list of] endogenous regressor(s)

`instrulist` is the list of instruments

e.g.:

```
ivregress 2sls lw s expr rns (iq = med), robust
```

```
. ivregress 2sls lw s expr rns (iq = med), robust

Instrumental variables (2SLS) regression          Number of obs   =        758
                                                  Wald chi2(4)    =     263.20
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.1398
                                                  Root MSE        =     .39757
```

lw	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
iq	.0195403	.0126679	1.54	0.123	-.0052883	.0443689
s	.0462323	.0389475	1.19	0.235	-.0301034	.1225679
expr	.0501608	.0085637	5.86	0.000	.0333761	.0669454
rns	-.0505755	.0528667	-0.96	0.339	-.1541923	.0530414
_cons	2.964178	.8260696	3.59	0.000	1.345112	4.583245

- If the `first` option is specified, STATA will also produce an output showing the first-stage regression, allowing us to evaluate the degree of correlation between the instruments and the endogenous regressor `iq`

```
ivregress 2sls lw s expr rns (iq = med), first r
```

```
. ivregress 2sls lw s expr rns (iq = med), first r

First-stage regressions
-----------------------

                                                Number of obs     =        758
                                                F(    4,     753) =      72.32
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2815
                                                Adj R-squared     =     0.2777
                                                Root MSE          =    11.5741
```

iq	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
s	2.863383	.2132616	13.43	0.000	2.444725	3.282041
expr	-.2517018	.2340601	-1.08	0.283	-.7111896	.2077861
rns	-2.793783	.8988043	-3.11	0.002	-4.558243	-1.029322
med	.4177794	.1733158	2.41	0.016	.0775398	.7580189
_cons	62.10312	3.025267	20.53	0.000	56.16416	68.04208

Instrumental variables (2SLS) regression

Number of obs	=	758
Wald chi2(4)	=	263.20
Prob > chi2	=	0.0000
R-squared	=	0.1398
Root MSE	=	.39757

lw	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
iq	.0195403	.0126679	1.54	0.123	-.0052883	.0443689
s	.0462323	.0389475	1.19	0.235	-.0301034	.1225679
expr	.0501608	.0085637	5.86	0.000	.0333761	.0669454
rns	-.0505755	.0528667	-0.96	0.339	-.1541923	.0530414
_cons	2.964178	.8260696	3.59	0.000	1.345112	4.583245

Instrumented: iq

Instruments: s expr rns med

estat firststage

```
. estat firststage
```

First-stage regression summary statistics

Variable	R-sq.	Adjusted R-sq.	Partial R-sq.	Robust F(1,753)	Prob > F
iq	0.2815	0.2777	0.0084	5.81056	0.0162

- The 4th column marked “F(1, 753)” is an F statistic for the joint significance of the instrument(s).
- F statistic should exceed 10 for inference based on the 2SLS estimator to be reliable.
- Here F-statistic is only 5.81, indicating that mother’s education is only weakly correlated with iq. Hence, mother’s education is a weak instrument.

Problem with IVs

- In the charter e.g., offer of a charter seat acts as a valid instrument for charter school attendance because it isolates variation in charter school attendance that is as good as random.

  - Instrument pulls out the part of the variation in X_i that is random (& therefore uncorrelated with u_i).

- Nice aspect of IV is that the relevance assumption can be tested easily.

- Drawback is that the exogeneity assumption cannot be tested.

  - Major source of controversy surrounding use of instruments.

- In the IQ example, instrument is controversial. Why?

Limitations of IV

- Although IV provides a general solution for obtaining a consistent estimator of the causal effect of interest that holds no matter what the source of correlation between X_i & u_i is, it is difficult to implement because valid instruments are hard to find.

Homework 2

- Homework 2 will be posted on 11 April (Tuesday). Like Homework 1, it will be done through the "**Canvas Quiz**" Platform.

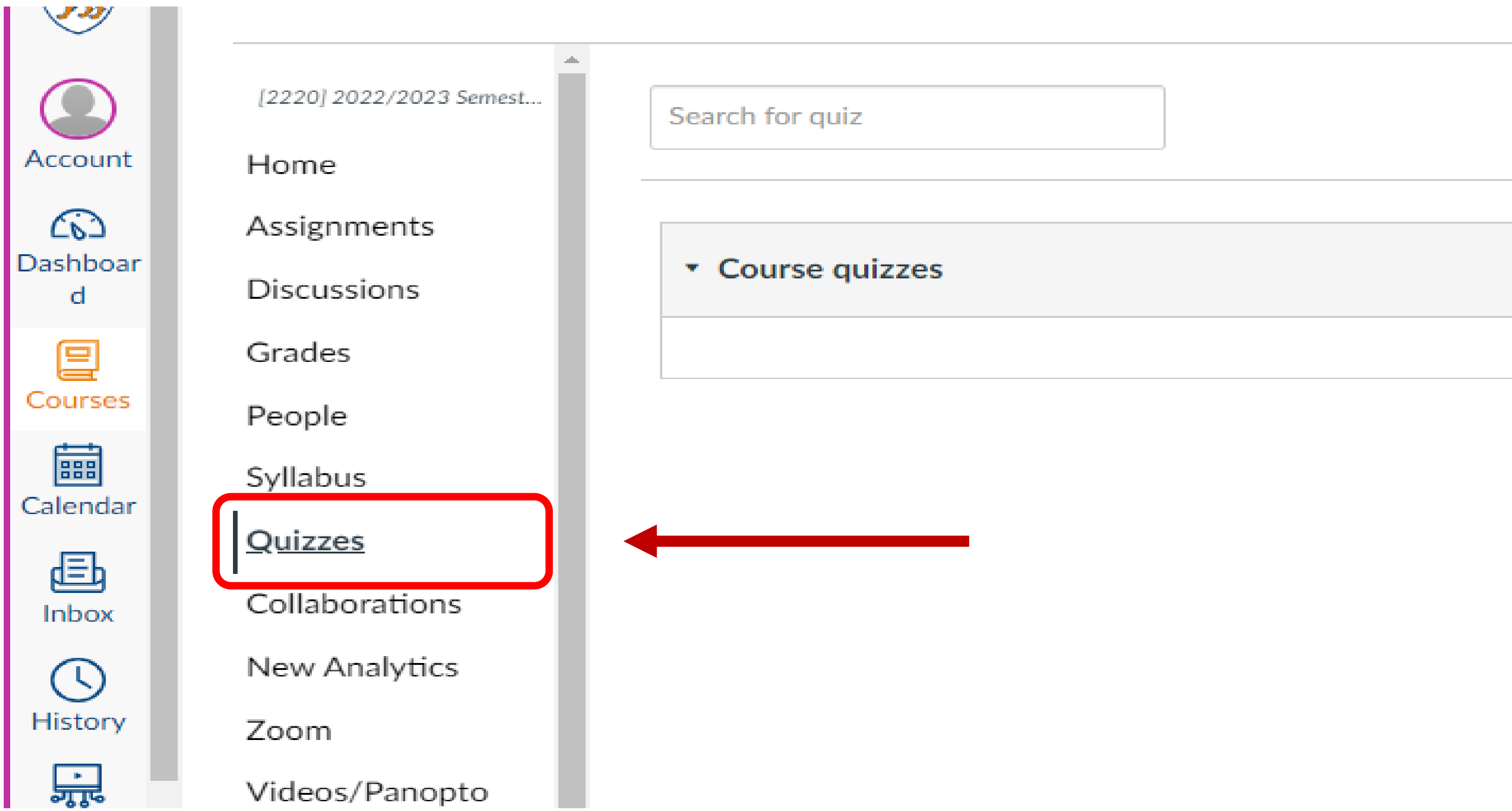
- To access Homework 2, login to Canvas, then on the left panel, click on "Quizzes" (see next slide) and you will be able to access the Homework.

- The Homework will open at 12pm on 11 April (Tuesday) and will **close at 7pm on 14 April (Friday)**.

- You can attempt the HW anytime before 7pm on 14 April. Thereafter, it will not be accessible. No late submission will be accepted.

- Like Homework 1, you will not need to finish the homework in one sitting. You can save the HW and then submit it later.

Canvas Quiz



Account
Dashboard
Courses
Calendar
Inbox
History
[2220] 2022/2023 Semest...
Home
Assignments
Discussions
Grades
People
Syllabus
Quizzes
Collaborations
New Analytics
Zoom
Videos/Panopto
Search for quiz
Course quizzes