Lecture 3: Instrumental Variables

Yu Bai

City University of Macau

The Oregon Trail

- Research Question:
  - Would Medicaid expansion improve health?
- The State of Oregon offered Medicaid to thousands of randomly chosen people in a publicly announced health insurance lottery.
- Winners won the opportunity to apply for the state-run Oregon Health Plan (OHP), the Oregon version of Medicaid.

The Oregon Trail

- The state then reviewed these applications, awarding coverage to Oregon residents who were
  - U.S. citizens OR legal immigrants aged 19–64
  - not otherwise eligible for Medicaid
  - uninsured for at least 6 months
  - income below the federal poverty level
  - few financial assets
- To initiate coverage, lottery winners had to document their poverty status and submit the required paperwork within 45 days.

The Oregon Trail

- Roughly 75,000 lottery applicants registered, and almost 30,000 were randomly selected and invited to apply for OHP.
- What was the first question asked?
  - whether OHP lottery winners were more likely to end up insured as a result of winning
- We find that the probability of Medicaid coverage increased by 26 percentage points for lottery winners.

The Oregon Trail

OHP effects on insurance coverage and health-care use

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
	A. Administrative data			
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)

Recap: The Oregon Trail

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
	A. Health indicators			
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
	B. Financial health			
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

The Oregon Trail

- At that time, we say that lottery winners constitute the OHP treatment group. The other 45,000 constitute the OHP control sample.
- As Oregon's health insurance lottery is to randomly select winners and losers from a pool of registrants, we are confident that treatment status is random assigned.
- However, isn't it better to consider people who ended up insured as the treatment group?

Encouragement designs and non-compliance

- In contrast to RHIE, there is an imperfect match between
  - the units that are assigned the treatment AND
  - the units that received the treatment.
- This is known as **non-compliance**.

Non-compliance

- There are two-types of non-compliance:

One-sided non-compliance

When some units assigned to treatment DO NOT receive the treatment OR some units assigned to control DO receive the treatment

Two-sided non-compliance

When some units assigned to treatment DO NOT receive the treatment AND some units assigned to control DO receive the treatment

- Example of one-sided non-compliance:
  - In field experiment, houses are randomly assigned to canvassers, but some people are not at home.

Dealing with non-compliance

- Use the difference estimator based on treatment intake?

$$
\begin{aligned}
E[Y_{1i}] - E[Y_{0i}] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\
&= \underbrace{E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1]}_{\text{ATT}} \\
&\qquad + \underbrace{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

- Unfortunately,

$$E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \neq 0$$

as treatment intake is not randomized.

Dealing with non-compliance

- Use the difference estimator based on treatment assignment?

$$\begin{aligned} E[Y_{1i}] - E[Y_{0i}] &= E[Y_{1i}|Z_i = 1] - E[Y_{0i}|Z_i = 0] \\ &= E[Y_{1i}|Z_i = 1] - E[Y_{0i}|Z_i = 1] \end{aligned}$$

- There is no selection bias, as treatment is randomly assigned.
- We name the above as the intention to treat (ITT) effect.

Dealing with non-compliance

- What we compute in the Oregon Trail example is actually the **ITT**.
- **ITT** measures the causal effect of the offer of treatment.
  - Not everyone wins the lottery, and not every lottery winner gets the OHP in the end.
- When are we interested in **ITT**? If we
  - care about if some program makes a difference to average outcome
  - are more interested in the effectiveness of the treatment

A new approach

- We will start to learn a new approach, called the instrumental variable (**IV**) regression.
- **IV** approach allows us to use the **ITT** (intention to treat effect) to estimate an **ATT**.
  - Notice that it will not be the usual **ATT** we talked about before, but rather an **ATT** for a particular subgroup.
- It is particularly useful when random treatment intake is impossible and even unethical.

Who invented IV regression?

- It first appeared in Philip G. Wright's book "*The Tariff on Animal and Vegetable Oils*", published in 1928.
- Most of this book is concerned with the question of whether the steep tariffs on farm products imposed in the early 1920s benefited domestic producers.
- Philip G. Wright was not only an economist, but also a poet.

Who invented IV regression?

- Appendix B of the book begins with an elegant statement of the identification problem in simultaneous equations models.
- The appendix then goes on to explain how variables present in one equation but excluded from another solve the identification problem.
- Philip referred to such excluded variables as "external factors", which we call them "instruments" today.
- He (Philip) derived and then used IV to estimate supply and demand curves in markets for butter and flaxseed (flaxseed is used to make linseed oil, an ingredient in paint).

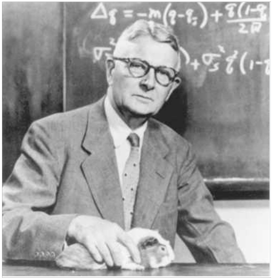
Simultaneous equation models

$$q_t^d(p) = \alpha_0 + \alpha_1 p + \alpha_2 z_t + \epsilon_t^d$$
$$q_t^s(p) = \beta_0 + \beta_1 p + \beta_2 x_t + \epsilon_t^s$$
$$q_t^d(p_t) = q_t^s(p_t) = q_t$$

- The first two equations are the demand and supply equations, while the third equation is the market equilibrium condition.
- Variables determined jointly by solving the system (equilibrium condition) are said to be endogenous.
- Variables determined outside the system, like x_t and z_t, are said to be exogenous.

Who invented IV regression?

- Appendix B was a major breakthrough in econometrics, but much unexpected.
- Some have speculated that this part may be written by Philip's son, Sewall Wright, who was a geneticist and statistician.



Who invented IV regression?

- James Stock and Francesco Trebbi investigated the case for Sewall's authorship using Stylometrics. HERE
- Stylometrics identifies authors by the statistical regularities in their word usage and sentence structure, which confirms Philip's authorship of Appendix B.
- Stock and his student Kerry Clark later uncovered letters between father and son that show the ideas in Appendix B developing jointly in a self-effacing give and take.

Who invented IV regression?

Back to the potential outcome

Instrument

Z_i, a binary instrument for unit i:

- $Z_i = 1$, if unit i is "encouraged" to take treatment
- $Z_i = 0$, if unit i is "encouraged" to take control

Potential treatment

D_{zi}, a potential treatment status, given $Z_i = z$, e.g.

- $D_{1i} = 1$: "encouraged" to take and take it eventually
- $D_{0i} = 1$: not "encouraged" to take but take it anyway

Back to the potential outcome

- Then, the observed treatment D_i, is connected to the potential treatment, (D_{1i}, D_{0i}), as follows

$$D_i = Z_i \cdot D_{1i} + (1 - Z_i) \cdot D_{0i}.$$

- We can now classify units into four categories by their potential treatments:
  - **Compliers**: $D_{1i} > D_{0i}$, i.e. $D_{0i} = 0$ and $D_{1i} = 1$
  - **Always-takers**: $D_{1i} = D_{0i} = 1$
  - **Never-takers**: $D_{1i} = D_{0i} = 0$
  - **Defiers**: $D_{1i} < D_{0i}$, i.e. $D_{0i} = 1$ and $D_{1i} = 0$

Back to the potential outcome

	$Z_i = 0$	$Z_i = 1$
$D_i = 0$	Never-taker (N) or Complier (C)	Never-taker (N) or Defier (D)
$D_i = 1$	Always-taker (A) or Defier (D)	Always-taker (A) or Complier (C)

- The above implies that we have a problem:
  - We **DO NOT** know the type of any given unit i.
- This is because we only observe D_{1i} or D_{0i} for a given unit.
  - When $D_i = 1$, $Z_i = 1$, unit i can be either Always-taker (A) or Complier (C).
- We need certain assumptions to identify the *proportion* of each type.

Identifying assumptions

Assumption I: Independence

$$(Y_i(D_{1i}, 1), Y_i(D_{0i}, 0), D_{1i}, D_{0i}) \perp Z_i$$

Assumption II: First stage validity

$$E[D_{1i} - D_{0i}] \neq 0$$

Assumption III: Monotonicity

$D_{1i} \geq D_{0i}, \forall i$

Assumption IV: Exclusive restriction

For $d = 0, 1$, we have

$$Y_i(d, 0) = Y_i(d, 1) = Y_{d_i}.$$

Interpretations

- Independence assumption implies that treatment Z_i is randomly assigned.
  - IV should not be correlated with both the potential treatment status and potential outcome.
- Notice that

$$
\begin{aligned}
E[D_{1i} - D_{0i}] &= 1 \cdot P(D_{1i} - D_{0i} = 1) + (-1) \cdot \cancel{P(D_{1i} - D_{0i} = -1)} \\
&= P(D_{1i} - D_{0i} = 1) \neq 0,
\end{aligned}
$$

because by Monotonicity assumption the latter term is 0.

Interpretations

- First stage assumption implies that we can identify the proportion of compliers.
- Monotonicity assumption is also not restrictive, as
  - Most people are not that weird.
  - Reminder:
- Exclusive restriction implies that treatment assignment only affects the potential outcome by affecting the potential treatment status.
  - Hard to check empirically
  - Will go back to this later

Decomposition of ITT

- Let π_{type} be the proportion of each type and $\text{LATE}_{\text{type}}$ be the local ATT for each type, where type $= \{\text{C}, \text{D}, \text{AN}\}$
- Then, we have

$$\begin{aligned}
\text{ITT} &= \text{LATE}_\text{C}\pi_\text{C} + \text{LATE}_\text{D}\pi_\text{D} + \text{LATE}_\text{A}\pi_\text{A} + \text{LATE}_\text{N}\pi_\text{N} \\
&= \text{LATE}_\text{C}\pi_\text{C} + \text{LATE}_\text{D}0 + \text{LATE}_\text{A}\pi_\text{A} + \text{LATE}_\text{N}\pi_\text{N} \\
&= \text{LATE}_\text{C}\pi_\text{C} + \text{LATE}_\text{D}0 + 0\pi_\text{A} + 0\pi_\text{N} \\
&= \text{LATE}_\text{C}\pi_\text{C}
\end{aligned}$$

- This implies that under the imposed Assumptions, we can recover the LATE for compliers via

$$\text{LATE}_\text{C} = \frac{\text{ITT}}{\pi_\text{C}}$$

LATE

- We shall drop the subscript C for notation simplicity.
- By plugging in what we have got so far, we obtain

$$\text{LATE} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}.$$

- The denominator $E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$ measures the effects of Z_i on D_i.
- The numerator $E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$ measures the effects of Z_i on Y_i.

LATE

- It can be shown that the above can be simplified to

$$E[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}].$$

- Let us summarize what we have got so far:

Estimand (LATE)

Under Assumptions I-IV, we can identify the local average treatment (**LATE**) effects for compliers:

$$\tau_{\text{LATE}} = E[Y_{1i} - Y_{0i} | D_{1i} > D_{0i}].$$

- τ_{LATE} is also called the **Wald Estimator**.

An equivalent representation

- Still remember the equivalence between the difference estimator and the dummy variable regression?
- Consider the following regression:
  - First stage:
    $$D_i = \alpha_1 + \beta_1 Z_i + \epsilon_{1i}$$
  - Reduced form:
    $$Y_i = \alpha_2 + \beta_2 Z_i + \epsilon_{2i}$$
- This implies that LATE can also be obtained by the "*ratio of coefficient*" approach:
  $$\text{LATE} = \frac{\beta_2}{\beta_1}.$$

Implications of the IV formula

- LATE should be zero if $\beta_2 = 0$: IV has no impact on the potential outcome.
- If $\beta_1 = 0$, β_2 should also be zero, as

$$\beta_2 = \beta_1 \cdot \mathsf{LATE} = 0.$$

  - If IV has no impact on the potential treatment status, it should also not have an impact on the potential outcome.
- What happens if $\hat{\beta}_1 \approx 0$ but $\hat{\beta}_2 \neq 0$?
  - This might be problematic, as IV may affect the potential outcome through other channel, violating the Exclusive restriction.

Back to ATT

- What if we are still interested in ATT?
- Notice that

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E[Y_{1i} - Y_{0i}|D_i = 1, Z_i = 0]P[D_i = 1|Z_i = 0] + \underbrace{E[Y_{1i} - Y_{0i}|D_{1i} > D_{0i}]}_{\text{LATE}} \underbrace{P[D_{1i} > D_{0i}|D_i = 1, Z_i = 1]}_{\text{proportion of C}}$$

- So, ATT is the weighted average of LATE of both Always-takers and Compliers.

Back to ATT

- It is possible to recover ATT with an additional assumption.
- The results are due to Howard Bloom (1986), which are summarized in the following theorem.

Theorem

Suppose that Assumptions I-IV are satisfied and the following condition holds

$$E[D_i|Z_i = 0] = P[D_i = 1|Z_i = 0] = 0.$$

Then, we have

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{P[D_i = 1|Z_i = 1]} = E[Y_{1i} - Y_{0i}|D_i = 1].$$

The Charter Conundrum

- Charter schools are
  - public schools that operate with considerably more autonomy than traditional American public schools;
  - free to structure their curricula and school environments.
- Teachers and staff who work at Charter schools rarely belong to labour unions.
  - Most big-city public school teachers work under teachers' union contracts that regulate pay and working conditions: may improve working conditions for teachers, but they can make it hard to reward good teachers or dismiss bad ones.

The Charter Conundrum

> INTERVIEWER: Have your mom and dad told you about the lottery?
>
> DAISY: The lottery ... isn't that when people play and they win money?
>
> *Waiting for Superman*, 2010

- Among the schools featured in Waiting for Superman is **KIPP** LA College Prep, one of more than 140 schools affiliated with the Knowledge Is Power Program.
- KIPP schools emphasize discipline and comportment and features a long school day, an extended school year, selective teacher hiring, and a focus on traditional reading and math skills.
- Today, the KIPP network serves a student body that is 95% black and Hispanic, with more than 80% of KIPP students poor enough to qualify for the federal government's subsidized lunch program.

The Charter Conundrum

- The American debate over education reform often focuses on the achievement gap, shorthand for uncomfortably large test score differences by race and ethnicity.
- Because of its focus on minority students, KIPP is often central in this debate:
  - Supporters: Nonwhite KIPP students have markedly higher average test scores than nonwhite students from nearby schools.
  - Skeptics: KIPP attracts families whose children are more likely to succeed anyway.

Playing the lottery

- The first KIPP school in New England was a middle school in the town of Lynn, Massachusetts, just north of Boston.
- Although urban charter schools typically enroll many poor, black students, KIPP Lynn is unusual among charters in enrolling a high proportion of Hispanic children with limited English proficiency.
- After 2005, however, demand accelerated, with more than 200 students applying for about 90 seats in fifth grade each year.
- As required by Massachusetts law, scarce charter seats are allocated by lottery.
- Our IV tool uses these admissions lotteries to frame a naturally occurring randomized trial.

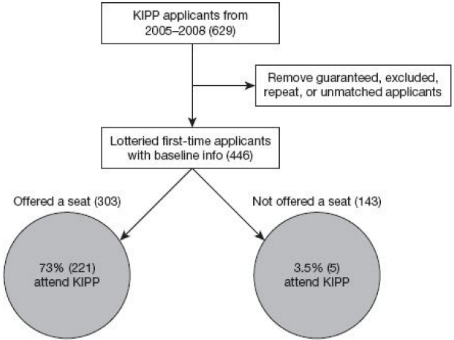
Playing the lottery

- Assuming the only difference created by winning the lottery is in the likelihood of charter enrollment (Assumption ?), IV turns randomized offer effects into causal estimates of the effect of charter attendance.
- Specifically, IV estimates capture causal effects on the sort of child who enrolls in KIPP when offered a seat in a lottery but wouldn't manage to get in otherwise.

Playing the lottery

FIGURE 3.1

Application and enrollment data from KIPP Lynn lotteries



Note: Numbers of Knowledge Is Power Program (KIPP) applicants are shown in parentheses.

Playing the lottery

			KIPP applicants		
	Lynn public fifth graders (1)	KIPP Lynn lottery winners (2)	Winners vs. losers (3)	Attended KIPP (4)	Attended KIPP vs. others (5)
Panel A. Baseline characteristics					
Hispanic	.418	.510	−.058 (.058)	.539	.012 (.054)
Black	.173	.257	.026 (.047)	.240	−.001 (.043)
Female	.480	.494	−.008 (.059)	.495	−.009 (.055)
Free/Reduced price lunch	.770	.814	−.032 (.046)	.828	.011 (.042)
Baseline math score	−.307	−.290	.102 (.120)	−.289	.069 (.109)
Baseline verbal score	−.356	−.386	.063 (.125)	−.368	.088 (.114)
Panel B. Outcomes					
Attended KIPP	.000	.787	.741 (.037)	1.000	1.000 —
Math score	−.363	−.003	.355 (.115)	.095	.467 (.103)
Verbal score	−.417	−.262	.113 (.122)	−.211	.211 (.109)
Sample size	3,964	253	371	204	371

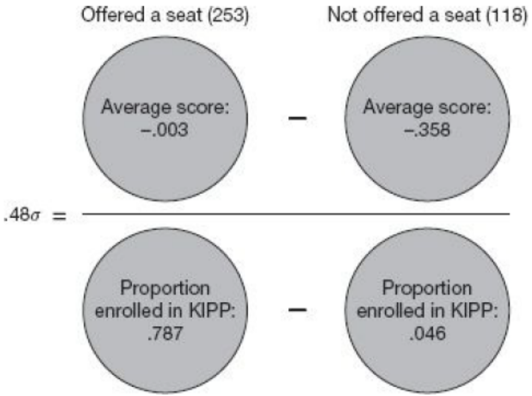
Playing the lottery

- What is the instrumental variable in this context?
- Explain in plain words the meaning of the assumptions in this context:
  1. Independence:
  2. First stage:
  3. Monotonicity:
  4. Exclusive restriction:

Playing the lottery

FIGURE 3.2

IV in school: the effect of KIPP attendance on math scores



LATE for Charter School

- The first stage:

$$\phi = E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$$

  - In the KIPP study, ϕ is the difference in ...

- The reduced form:

$$\rho = E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]$$

  - In the KIPP study, ρ is the difference in ...

LATE for Charter School

- LATE:

$$\lambda = \frac{\rho}{\phi} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

  - In the KIPP study, λ is the difference in ...

- It won't surprise you to learn that there's a formula for IV standard errors and that your econometric software knows it. Problem solved!

LATE for Charter School

TABLE 3.2
The four types of children

		Lottery losers $Z_i = 0$	
		Doesn't attend KIPP $D_i = 0$	Attends KIPP $D_i = 1$
Lottery winners $Z_i = 1$	Doesn't attend KIPP $D_i = 0$	Never-takers (*Normando*)	Defiers "wicked"
	Attends KIPP $D_i = 1$	Compliers (*Camila*)	Always-takers (*Alvaro*)

LATE for Charter School

- The number of charter seats is capped by law in Massachusetts: the consequences of charter expansion is the education policy question.
- Compliers are children likely to attend KIPP were the network to expand and offer additional seats in a lottery: perhaps as a consequence of opening a new school in the same area.
- Luckily, what we care in this example seems to be LATE itself.

LATE for Charter School

- Keep in mind that LATE and ATT are not the same!
- They also not need to be the same over time or in different settings.
- The next question is *external validity*:
  - whether a particular causal estimate has predictive value for times, places, and people beyond those represented in the study that produced.

What we have learned so far...

- With a single instrument Z_i and additional assumptions, we can estimate
  - **ITT**:
$$E[Y_{1i}|Z_i = 1] - E[Y_{0i}|Z_i = 1]$$
  - **proportion of compliers** π_C:
$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0]$$
  - **LATE**:
$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$
- Lotteries are awesome! But, we aim to generalize this framework to include
  - non-binary IV or more than 1 IV
  - additional covariates

An alternative representation of LATE

- The first-stage equation is

$$D_i = \alpha_1 + \beta_1 Z_i + \epsilon_{1i} \tag{1}$$

- The reduced form equation is

$$Y_i = \alpha_2 + \beta_2 D_i + \epsilon_{2i}. \tag{2}$$

  - As $E[\epsilon_{2i}|D_i] \neq 0$, we cannot just run OLS.

An alternative representation of LATE

- Since

$$\text{Cov}\left(Y_i, Z_i\right) = \text{Cov}\left(\alpha_2 + \beta_2 D_i + \epsilon_{2i}, Z_i\right) = \beta_2 \text{Cov}\left(D_i, Z_i\right),$$

which implies that

$$\beta_2 = \frac{\text{Cov}\left(Y_i, Z_i\right)/\text{Var}\left(Z_i\right)}{\text{Cov}\left(D_i, Z_i\right)/\text{Var}\left(Z_i\right)}.$$

- IV estimator is essentially the ratio of reduced-form OLS estimate and first-stage OLS estimate.

2SLS: Two Stage Least Squares

- Same as before, but with multiple and possibly non-binary instruments and controls:

$$D_i = \alpha_1 + \beta_1 Z_i + \lambda_1 W_i + \gamma_1' A_i + \epsilon_{1i}, \tag{3}$$

$$Y_i = \alpha_2 + \beta_2 D_i + \gamma_2' A_i + \epsilon_{2i}, \tag{4}$$

  where

  - Z_i and W_i are instruments;
  - A_i includes controls.

- Instruments often come from deep institutional knowledge and revealing.

2SLS: Two Stage Least Squares

- Substituting D_i in (4) with (3), we have

$$Y_i = \alpha_2 + \beta_2 \left[\alpha_1 + \beta_1 Z_i + \lambda_1 W_i + \gamma_1' A_i\right] + \gamma_2' A_i + \xi_{2i},$$

  where $\xi_{2i} = \beta \epsilon_{1i} + \epsilon_{2i}$.
- Notice that $[\alpha_1 + \beta_1 Z_i + \lambda_1 W_i + \gamma_1' A_i]$ is simply the fitted value of D_i from regression (3).
- The disturbance term ξ_{2i} is also not correlated with Z_i, W_i, and A_i.

2SLS: Two Stage Least Squares

- Let $\pi_1 = \beta_1\beta_2$ so that $\beta_2 = \pi_1/\beta_1$, which is exactly the formula of Wald estimator.
- This also implies that if Z_i is included in the second stage, we will not have the usual Wald estimator representation for β_2 so that the estimator is biased (also inconsistent).

2SLS: Two Stage Least Squares

- Estimation done in two stages:
  1. OLS for (3) to obtain fitted values of $\hat{D}_i$;
  2. OLS for (4) by replacing D_i with $\hat{D}_i$.
- Intuition:
  - As $\hat{D}_i$ reflects the predictive content in D_i caused by a unit value of Z_i (ceteris paribus), which is randomly assigned, coefficient attached to $\hat{D}_i$ measures the effects of D_i on Y_i for compliers (those whose variation of D_i can be explained by Z_i).

Comparison with using controls

- In Lec 2, we consider to relax the random assignment by adding controls:

$$Y_i = \alpha + \beta D_i + \gamma' W_i + \epsilon_i.$$

- So, what are the differences between IVs and controls?
  - β in the above regression measures the relationship between Y_i and part of D_i that is "**not explained**" by W_i. FWL
  - In IV regression, β measures the relationship between Y_i and the **only** part of D_i that is explained by Z_i. First stage

The Population Bomb

- Economists have turned a micro lens on the relationship between family size and living standards.
- On the policy side, we have
  - One Child Policy in China
  - Forced-sterilization program in India
  - Family planning in Mexico and Indonesia
- For the most part, fertility is determined by the choices parents make.
- Not surprisingly, therefore, women with large families differ in many ways from those with smaller families; they tend to be less educated, for example.

The Population Bomb

- How to quantify the causal impact of family size on schooling? Needs to address the omitted variable problem
  - Random assignment?
    1. Draw a sample of families with one child.
    2. In some of these households, randomly distribute an additional child.
    3. Wait 20 years and collect data on the educational attainment of firstborns who did and did not get an extra sibling.
       - aren't likely to see such an experiment any time soon...

The Population Bomb

- The first attempt is on the twins: a family size experiment.
- Based on
  - Joshua D. Angrist, Victor Lavy, and Analia Schlosser, "Multiple Experiments for the Causal Link between the Quantity and Quality of Children," Journal of Labor Economics, vol. 28, no. 4, October 2010, pages 773–824.
- Why Israel?
  - diverse population
  - data availability: family origins and sex of their siblings

The Population Bomb

- We focus here on a group of firstborn adults in a random sample of men and women born to mothers with at least two children.
- On average, such families include 3.6 children. A second twin birth, however, increases average family size by .32, that is, by about one-third of a child.
- What will happens if we regress adult firstborns' highest grade on family size?
  - Negative and significant, perhaps for sure.

The Population Bomb

- The comparison of schooling between firstborns with twin and singleton siblings constitutes the reduced form for an IV estimate that uses twin births as an instrument for family size.
- Not successful: the twins reduced-form and associated IV estimates are close to zero.
- Why?
  - Multiple births are more frequent among mothers who are older and for women in some racial and ethnic groups.

The Population Bomb

- The second attempt is on sibling sex composition.
- Motivated by the fact that families whose first two children are both boys or both girls are more likely to have a third child in some countries.
- Because the sex of a newborn is essentially randomly assigned (male births occur about half the time and, in the absence of sex-selective abortion, little can be done to change this), parental preferences for mixed sibling sex composition generate sex-mix instruments.

The Population Bomb

- IV in this case is a dummy variable that equals 1 for families whose first two children are both male or both female and equals 0 for families with one boy and one girl.
- Still not successful: the educational attainment of firstborn Israeli adults is unaffected by their siblings' sex composition.

Question

Why they do not work? (exclusive restrictions)

The Population Bomb

- Possible remedies:
  - Combining two instruments
  - Adding controls
- This leads to the following Two-stage framework:

$$D_i = \alpha_1 + \phi_t Z_i + \phi_s W_i + \gamma_1 A_i + \delta_1 B_i + e_{1i} \tag{5}$$

$$Y_i = \alpha_0 + \rho_t D_i + \rho_s W_i + \gamma_0 A_i + \delta_0 B_i + e_{0i} \tag{6}$$

where

- W_i: equal to 1 if two boys or two girls and 0 otherwise
- A_i: maternal age
- B_i: equal to 1 for firstborn boys and 0 otherwise

The Population Bomb

- (5) is called first stage regression. (6) is called second stage regression.
- Below is called *reduced form* regression:

$$Y_i = \alpha_2 + \lambda_t Z_i + \lambda_s W_i + \gamma_2 A_i + \delta_2 B_i + e_{2i}.$$

Question

Do controls need to be the same in both stages?

The Population Bomb

TABLE 3.4
Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320 (.052)	.437 (.050)			.449 (.050)
Same-sex sibships			.079 (.012)	.073 (.010)	.076 (.010)
Male		−.018 (.010)		−.020 (.010)	−.020 (.010)
Controls	No	Yes	No	Yes	Yes

The Population Bomb

TABLE 3.5

OLS and 2SLS estimates of the quantity-quality tradeoff

		2SLS estimates		
Dependent variable	OLS estimates (1)	Twins instruments (2)	Same-sex instruments (3)	Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college	−.023	.017	.078	.048
(for age $\geq$ 24)	(.001)	(.052)	(.054)	(.037)
College graduate	−.015	−.021	.125	.052
(for age $\geq$ 24)	(.001)	(.045)	(.053)	(.032)

Weak instruments

Assumption II: First stage validity

$$E[D_{1i} - D_{0i}] \neq 0$$

- This assumption implies that

$$P\left(D_{1i} - D_{0i} = 1\right) \neq 0.$$

- In other words, Z_i has to induce some variation in D_i.
- What happens if
  - in the Oregon Trail example, not so many lottery winners apply for OHP?
  - in the KIPP example, not so many lottery winners attend KIPP?

Weak instruments

- When Z_i explains little variation in D_i, the instrument is said to be **weak**.
- Recall the LATE estimand:

$$\text{LATE} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = \frac{\text{Cov}\,(Y_i, Z_i)}{\text{Cov}\,(D_i, Z_i)}.$$

- If $\text{Cov}\,(D_i, Z_i) = 0$, LATE is undefined as there are no *compliers*.
- Furthermore, if $\text{Cov}\,(D_i, Z_i) \to 0$, it can be shown that
  - 2SLS estimator is biased towards the OLS estimator;
  - the bias does not diminish in large samples;
  - nonstandard asymptotic distribution.

How weak is weak?

- How to assess if Z_i is weak?
- Specify two models:

$$D_i = \alpha_1 + \phi_t Z_i + \phi_s W_i + \gamma' A_i + e_{1i}, \qquad (M_2)$$
$$D_i = \alpha_1 + \phi_s W_i + \gamma' A_i + e_{1i}, \qquad (M_1)$$

- Then, we run a F-test to compare the nested models:

$$F = \frac{\left(R^2_{M_2} - R^2_{M_1}\right) / \left(k_{M_2} - k_{M_1}\right)}{\left(1 - R^2_{M_2}\right) / \left(n - k_{M_2} + 1\right)}.$$

- Intuition: F is bigger when Z_i explains more variation in D_i.
- Rule of thumb: OK if $F \geq 10$.

Matrix algebra

- Let us first write down the model in more compact notation

$$\underbrace{Y}_{n\times 1} = \underbrace{X}_{n\times k}\beta + U$$
$$X = \underbrace{Z}_{n\times g}\underbrace{\gamma}_{g\times k} + V,$$

  where
  - $g \geq k$;
  - Regressors in X are endogenous and Z contains valid instruments.

Matrix algebra

- Using matrix notation, we can write the fitted value of X as $\hat{X} = PX$, where the projection matrix P is defined according to $P = Z(Z'Z)^{-1}Z'$.
- Then, we can write the 2SLS estimator as

$$\begin{aligned}
\hat{\beta}_{\text{2SLS}} &= \left(\hat{X}'\hat{X}\right)^{-1}\left(\hat{X}'Y\right) \\
&= \left(X'PX\right)^{-1}\left(X'PY\right) \\
&= \left(X'Z(Z'Z)^{-1}Z'X\right)^{-1}\left(X'Z(Z'Z)^{-1}Z'Y\right),
\end{aligned}$$

since P is both symmetric and idempotent.

Comparison with OLS

- Recall that, the first stage is to simply do the decomposition

$$x = \hat{x} + \hat{r},$$

which implies that

$$\text{TSS} = \text{ESS} + \text{RSS}, \;\; \text{TSS} \geq \text{ESS}.$$

- Loosely speaking, we have

$$X'X \geq X'PX, \;\; (X'X)^{-1} \leq (X'PX)^{-1}.$$

- This implies that

$$\text{var-cov}\left(\hat{\beta}_{\text{2SLS}}\right) \geq \text{var-cov}\left(\hat{\beta}_{\text{OLS}}\right).$$

- **Implication**: no FREE LUNCH, 2SLS removes the bias, but comes with LARGER variance.