Instrumental Variables

Teppei Yamamoto

Keio University

Introduction to Causal Inference
Spring 2016

Noncompliance in Randomized Experiments

- Often we cannot force subjects to take specific treatments
- Units choosing to take the treatment may differ in unobserved characteristics from units that refrain from doing so

Example: Non-compliance in JTPA Experiment

	Not Enrolled in Training	Enrolled in Training	Total
Assigned to Control	3,663	54	3,717
Assigned to Training	2,683	4,804	7,487
Total	6,346	4,858	11,204

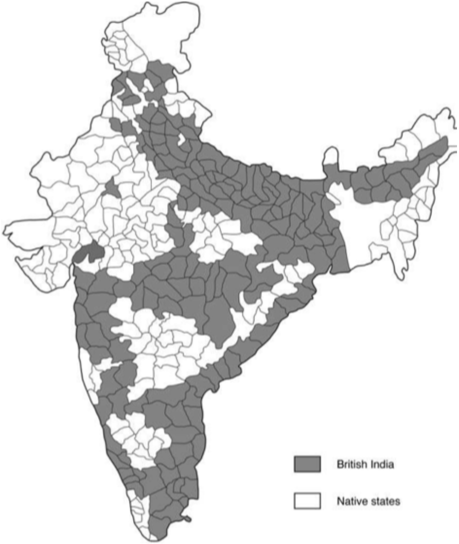
Partial Compliance in Randomized Experiments

- Unable to force all experimental subjects to take the (randomly) assigned treatment/control
- Intention-to-Treat (ITT) effect $\neq$ treatment effect
- Selection bias: self-selection into the treatment/control groups

- Political information bias: effects of campaign on voting behavior
- Ability bias: effects of education on wages
- Healthy-user bias: effects of exercises on blood pressure

- Encouragement design: randomize the encouragement to receive the treatment rather than the receipt of the treatment itself
- Instrumental variables can be regarded as an observational-study analogue of randomized encouragements

Iyer (2010):



Comparing Annexed States and Native States

Variable	British Empire	Native States
Mean Annual Rainfall	1503.41	1079.16
Log (population)	14.42	13.83
Population Density	279.47	169.20
Proportion Illiterate	.32	.28

- Naive comparison would suggest that districts directly ruled by British did better than districts formerly part of native states.
- Clear evidence that the British selectively annexed districts, making any comparison confounded.

The Doctrine of Lapse

Lord Dalhousie, Governor-General of India from 1848-1856, enacted a new policy regarding annexation:

> *I hold that on all occasions where heirs natural shall fail, the territory should be made to lapse and adoption should not be permitted, excepting in those cases in which some strong political reason may render it expedient to depart from this general rule.*

Deaths of Indian Rulers without Natural Heirs

- Number of districts where rulers died without an heir: 20
- Number of districts annexed due to the doctrine of lapse: 16
- Number of districts annexed due to other reasons: 19
- Annexation conditional on ruler dying without an heir: 16/20
- Annexation conditional on ruler not dying without an heir: 19/161

Potential Outcomes Framework for IV

Setup:

- Randomized encouragement: $Z_i \in \{0, 1\}$
- Potential treatment variables: $D_i(1), D_i(0)$
  1. $D_i(z) = 1$: would receive the treatment if $Z_i = z$
  2. $D_i(z) = 0$: would not receive the treatment if $Z_i = z$
- Observed treatment receipt indicator: $D_i = D_i(Z_i)$
- Observed and potential outcomes: $Y_i = Y_i(Z_i, D_i(Z_i))$
- Observed outcome can also be written as $Y_i = Y_i(Z_i)$

Identification of Intention-to-Treat Effect

Assumptions:

- SUTVA for $D_i(z)$ and $Y_i(z, d)$
- Randomization of encouragement:

$$\{Y_i(1), Y_i(0), D_i(1), D_i(0)\} \perp\!\!\!\perp Z_i$$

- But $\{Y_i(1), Y_i(0)\} \not\perp\!\!\!\perp D_i \mid Z_i = z$

Identification of Intention-to-Treat Effect

Assumptions:

- SUTVA for $D_i(z)$ and $Y_i(z, d)$
- Randomization of encouragement:

$$\{Y_i(1), Y_i(0), D_i(1), D_i(0)\} \perp\!\!\!\perp Z_i$$

- But $\{Y_i(1), Y_i(0)\} \not\perp\!\!\!\perp D_i \mid Z_i = z$

A quantity of interest: Intention-to-treat (ITT) effect:

$$\mathrm{ITT} \equiv \mathbb{E}[Y_i(1) - Y_i(0)]$$

- Effect of encouragement itself on outcome (regardless of *actual* treatment)

Identification of Intention-to-Treat Effect

Assumptions:

- SUTVA for $D_i(z)$ and $Y_i(z, d)$
- Randomization of encouragement:

$$\{Y_i(1), Y_i(0), D_i(1), D_i(0)\} \perp\!\!\!\perp Z_i$$

- But $\{Y_i(1), Y_i(0)\} \not\!\perp\!\!\!\perp D_i \mid Z_i = z$

A quantity of interest: Intention-to-treat (ITT) effect:

$$\text{ITT} \equiv \mathbb{E}[Y_i(1) - Y_i(0)]$$

- Effect of encouragement itself on outcome (regardless of *actual* treatment)

Because Z_i is randomized, ITT is identified by difference in means between the encouraged and unencouraged:

$$\text{ITT} = \mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]$$

Principal Stratification and Compliance Types

- Four principal strata (or compliance types):
  - compliers: $D_i(1) = 1$ and $D_i(0) = 0$
  - non-compliers $\begin{cases} \text{always-takers:} & D_i(1) = D_i(0) = 1 \\ \text{never-takers:} & D_i(1) = D_i(0) = 0 \\ \text{defiers:} & D_i(1) = 0 \text{ and } D_i(0) = 1 \end{cases}$

Principal Stratification and Compliance Types

- Four principal strata (or compliance types):
  - compliers: $D_i(1) = 1$ and $D_i(0) = 0$
  - non-compliers $\begin{cases} \text{always-takers:} & D_i(1) = D_i(0) = 1 \\ \text{never-takers:} & D_i(1) = D_i(0) = 0 \\ \text{defiers:} & D_i(1) = 0 \text{ and } D_i(0) = 1 \end{cases}$

- Correspondence between observed and principal strata:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier/Always-taker	Defier/Always-taker
$D_i = 0$	Defier/Never-taker	Complier/Never-taker

Principal Stratification and Compliance Types

- Four principal strata (or compliance types):
  - compliers: $D_i(1) = 1$ and $D_i(0) = 0$
  - non-compliers $\begin{cases} \text{always-takers:} & D_i(1) = D_i(0) = 1 \\ \text{never-takers:} & D_i(1) = D_i(0) = 0 \\ \text{defiers:} & D_i(1) = 0 \text{ and } D_i(0) = 1 \end{cases}$

- Correspondence between observed and principal strata:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier/Always-taker	Defier/Always-taker
$D_i = 0$	Defier/Never-taker	Complier/Never-taker

- Without further assumptions, compliance types cannot be identified from observed strata

Example: Indirect vs. Direct Rule in India

Type	Heir	No Heir	Explanation
1. Always Annexed	Annexed	Annexed	Rich princely state?
2. Annexed if no heir	Not Annexed	Annexed	Somewhat desirable?
3. Never annexed	Not Annexed	Not Annexed	Hard to Rule?
4. Annex if Heir	Annexed	Not Annexed	Rebellious family?

- Death without an heir is randomly assigned
- Districts annexed are a mix of type 1, 2, and 4. Those that are not annexed are a mix of types 2, 3, and 4. Thus, the two groups aren't comparable.
- Randomization ensures that the proportion of types are same over randomizations.

More Examples: Who Are the Compliers?

Study	Outcome	Treatment	Instrument
Angrist and Evans (1998)	Earnings	More than 2 Children	Multiple Second Birth (Twins)
Angrist and Evans (1998)	Earnings	More than 2 Children	First Two Children are Same Sex
Levitt (1997)	Crime Rates	Number of Policemen	Mayoral Elections
Angrist and Krueger (1991)	Earnings	Years of Schooling	Quarter of Birth
Angrist (1990)	Earnings	Veteran Status	Vietnam Draft Lottery
Miguel, Satyanath and Sergenti (2004)	Civil War Onset	GDP per capita	Lagged Rainfall
Acemoglu, Johnson and Robinson (2001)	Economic performance	Current Institutions	Settler Mortality in Colonial Times
Cleary and Barro (2006)	Religiosity	GDP per capita	Distance from Equator

Instrumental Variables Assumptions

Additional identification assumptions (Angrist, Imbens & Rubin 1996):

1. Monotonicity: No defiers

$$D_i(1) \geq D_i(0) \quad \text{for all } i.$$

Instrumental Variables Assumptions

Additional identification assumptions (Angrist, Imbens & Rubin 1996):

1. Monotonicity: No defiers

$$D_i(1) \geq D_i(0) \quad \text{for all } i.$$

2. Exclusion restriction: Instrument (encouragement) affects outcome only through treatment

$$Y_i(1,d) \ = \ Y_i(0,d) \quad \text{for } \ d = 0, 1$$

Zero ITT effect for always-takers and never-takers

Instrumental Variables Assumptions

Additional identification assumptions (Angrist, Imbens & Rubin 1996):

1. Monotonicity: No defiers

$$D_i(1) \geq D_i(0) \quad \text{for all } i.$$

2. Exclusion restriction: Instrument (encouragement) affects outcome only through treatment

$$Y_i(1, d) \ = \ Y_i(0, d) \quad \text{for } d = 0, 1$$

   Zero ITT effect for always-takers and never-takers

3. Relevance, or nonzero average encouragement effect:

$$\mathbb{E}[D_i(1) - D_i(0)] \neq 0$$

   Empirically testable

Decomposing the ITT Effect

- ITT effect can be decomposed into combination of subgroup ITTs:

$$\begin{aligned} \text{ITT} &= \text{ITT}_c \times \Pr(\text{compliers}) + \text{ITT}_a \times \Pr(\text{always-takers}) \\ &\quad + \text{ITT}_n \times \Pr(\text{never-takers}) + \text{ITT}_d \times \Pr(\text{defiers}) \end{aligned}$$

where

$$\begin{aligned} \text{ITT}_c &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = 1, D_i(0) = 0], \\ \text{ITT}_a &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = D_i(0) = 1], \text{ etc.} \end{aligned}$$

Decomposing the ITT Effect

- ITT effect can be decomposed into combination of subgroup ITTs:

$$\begin{aligned}\mathrm{ITT} &= \mathrm{ITT}_c \times \Pr(\text{compliers}) + \mathrm{ITT}_a \times \Pr(\text{always-takers}) \\ &\quad + \mathrm{ITT}_n \times \Pr(\text{never-takers}) + \mathrm{ITT}_d \times \Pr(\text{defiers})\end{aligned}$$

where

$$\begin{aligned}\mathrm{ITT}_c &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = 1, D_i(0) = 0], \\ \mathrm{ITT}_a &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = D_i(0) = 1], \text{ etc.}\end{aligned}$$

- Under monotonicity and exclusion restriction, this simplifies as:

$$\begin{aligned}\mathrm{ITT} &= \mathrm{ITT}_c \times \Pr(\text{compliers}) + \mathrm{ITT}_a \times \Pr(\text{always-takers}) \\ &\quad +\mathrm{ITT}_n \times \Pr(\text{never-takers}) + 0 \quad [\because \text{ monotonicity}]\end{aligned}$$

Decomposing the ITT Effect

- ITT effect can be decomposed into combination of subgroup ITTs:

$$\begin{aligned}\mathrm{ITT} &= \mathrm{ITT}_c \times \Pr(\text{compliers}) + \mathrm{ITT}_a \times \Pr(\text{always-takers}) \\ &\quad + \mathrm{ITT}_n \times \Pr(\text{never-takers}) + \mathrm{ITT}_d \times \Pr(\text{defiers})\end{aligned}$$

where

$$\begin{aligned}\mathrm{ITT}_c &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = 1, D_i(0) = 0], \\ \mathrm{ITT}_a &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = D_i(0) = 1], \text{ etc.}\end{aligned}$$

- Under monotonicity and exclusion restriction, this simplifies as:

$$\begin{aligned}\mathrm{ITT} &= \mathrm{ITT}_c \times \Pr(\text{compliers}) + \mathrm{ITT}_a \times \Pr(\text{always-takers}) \\ &\quad + \mathrm{ITT}_n \times \Pr(\text{never-takers}) + 0 \quad [\because \text{monotonicity}] \\ &= \mathrm{ITT}_c \times \Pr(\text{compliers}) + 0 \times \Pr(\text{always-takers}) \\ &\quad + 0 \times \Pr(\text{never-takers}) \quad [\because \text{exclusion restriction}]\end{aligned}$$

Decomposing the ITT Effect

- ITT effect can be decomposed into combination of subgroup ITTs:

$$\begin{aligned} \text{ITT} &= \text{ITT}_c \times \Pr(\text{compliers}) + \text{ITT}_a \times \Pr(\text{always-takers}) \\ &\quad + \text{ITT}_n \times \Pr(\text{never-takers}) + \text{ITT}_d \times \Pr(\text{defiers}) \end{aligned}$$

  where

$$\begin{aligned} \text{ITT}_c &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = 1, D_i(0) = 0], \\ \text{ITT}_a &= \mathbb{E}[Y_i(1, D_i(1)) - Y_i(0, D_i(0)) \mid D_i(1) = D_i(0) = 1], \text{ etc.} \end{aligned}$$

- Under monotonicity and exclusion restriction, this simplifies as:

$$\begin{aligned} \text{ITT} &= \text{ITT}_c \times \Pr(\text{compliers}) + \text{ITT}_a \times \Pr(\text{always-takers}) \\ &\quad + \text{ITT}_n \times \Pr(\text{never-takers}) + 0 \quad [\because \text{monotonicity}] \\ &= \text{ITT}_c \times \Pr(\text{compliers}) + 0 \times \Pr(\text{always-takers}) \\ &\quad + 0 \times \Pr(\text{never-takers}) \quad [\because \text{exclusion restriction}] \\ &= \text{ITT}_c \times \Pr(\text{compliers}) \end{aligned}$$

IV Estimand and Interpretation

- Therefore, ITT_c can be nonparametrically identified:

$$
\begin{aligned}
\text{ITT}_c &= \frac{\text{ITT}}{\Pr(\text{compliers})} \\
&= \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} \\
&= \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}
\end{aligned}
$$

IV Estimand and Interpretation

- Therefore, ITT_c can be nonparametrically identified:

$$\begin{aligned}
\text{ITT}_c &= \frac{\text{ITT}}{\Pr(\text{compliers})} \\
&= \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} \\
&= \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}
\end{aligned}$$

- ITT_c can be interepreted as Local Average Treatment Effect (LATE) for compliers:

$$\text{ITT}_c = \text{LATE}_c = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0]$$

IV Estimand and Interpretation

- Therefore, ITT_c can be nonparametrically identified:

$$\begin{aligned}\text{ITT}_c &= \frac{\text{ITT}}{\Pr(\text{compliers})} \\ &= \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} \\ &= \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}\end{aligned}$$

- ITT_c can be interepreted as Local Average Treatment Effect (LATE) for compliers:

$$\text{ITT}_c = \text{LATE}_c = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) = 1, D_i(0) = 0]$$

- LATE has a clear causal meaning, but interpretation is often tricky:
  - Compliers are defined in terms of principal strata, so we can never identify who they actually are
  - Different encouragement (instrument) yields different compliers

Special Case: One-sided Noncompliance

- Sometimes, control units have no access to the treatment
- If so, we have one-sided noncompliance where $D_i(0) = 0$ for all i

Special Case: One-sided Noncompliance

- Sometimes, control units have no access to the treatment
- If so, we have one-sided noncompliance where $D_i(0) = 0$ for all i

One-sided noncompliance makes things easier:

- Rules out always-takers and defiers
  $\Longrightarrow$ Monotonicity is *guaranteed* to hold!

Special Case: One-sided Noncompliance

- Sometimes, control units have no access to the treatment
- If so, we have one-sided noncompliance where $D_i(0) = 0$ for all i

One-sided noncompliance makes things easier:

- Rules out always-takers and defiers
  $\Longrightarrow$ Monotonicity is *guaranteed* to hold!
- Some individuals can be identified to be compliers or never-takers:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier/~~Always-taker~~	~~Defier/Always-taker~~
$D_i = 0$	~~Defier/~~Never-taker	Complier/Never-taker

Special Case: One-sided Noncompliance

- Sometimes, control units have no access to the treatment
- If so, we have one-sided noncompliance where $D_i(0) = 0$ for all i

One-sided noncompliance makes things easier:

- Rules out always-takers and defiers
  $\Longrightarrow$ Monotonicity is *guaranteed* to hold!
- Some individuals can be identified to be compliers or never-takers:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier/~~Always-taker~~	~~Defier/Always-taker~~
$D_i = 0$	~~Defier/~~Never-taker	Complier/Never-taker

- The LATE is now equal to the ATT, which is easier to interpret:

$$\begin{aligned}\text{LATE}_c &= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i(1) = 1, \cancel{D_i(0) = 0}] \\ &= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \;=\; \text{ATT}\end{aligned}$$

Violations of Instrumental Variables Assumptions

- Exclusion restriction: Violated when there are alternative causal paths
- Violation implies:

$$\text{Large sample bias} \quad = \quad \text{ITT}_{\text{noncomplier}} \frac{\Pr(\text{noncomplier})}{\Pr(\text{complier})}$$

- A weak instrument exacerbates the bias (more on this later)

Violations of Instrumental Variables Assumptions

- Exclusion restriction: Violated when there are alternative causal paths
- Violation implies:

$$\text{Large sample bias} \quad = \quad \text{ITT}_{\text{noncomplier}} \frac{\Pr(\text{noncomplier})}{\Pr(\text{complier})}$$

- A weak instrument exacerbates the bias (more on this later)

- Monotonicity: Violated when defiers exist
- Violation implies:

$$\text{Large sample bias} \quad = \quad \frac{\{\text{LATE}_c - \text{LATE}_d\} \Pr(\text{defier})}{\Pr(\text{complier}) - \Pr(\text{defier})}$$

- Bias becomes large when:
  - the proportion of defiers is large
  - causal effects are heterogenous between compliers and defiers

Example: The Vietnam Draft Lottery (Angrist 1990)

- Effect of military service on civilian earnings
- Simple comparison between Vietnam veterans and non-veterans are likely to be a biased measure
- Angrist (1990) used draft-eligibility, determined by the Vietnam era draft lottery, as an instrument for military service in Vietnam
- Draft eligibility is random and affected the probability of enrollment
- Estimate suggest a 15% effect of veteran status on earnings in the period 1981-1984 for white veterans born in 1950-51; although the estimators are quite imprecise

Wald Estimates for Vietnam Draft Lottery

		Draft-Eligibility Effects in Current $				
Cohort	Year	FICA Earnings (1)	Adjusted FICA Earnings (2)	Total W-2 Earnings (3)	$\hat{p}^e - \hat{p}^n$ (4)	Service Effect in 1978 $ (5)
1950	1981	−435.8	−487.8	−589.6	0.159	−2,195.8
		(210.5)	(237.6)	(299.4)	(0.040)	(1,069.5)
	1982	−320.2	−396.1	−305.5		−1,678.3
		(235.8)	(281.7)	(345.4)		(1,193.6)
	1983	−349.5	−450.1	−512.9		−1,795.6
		(261.6)	(302.0)	(441.2)		(1,204.8)
	1984	−484.3	−638.7	−1,143.3		−2,517.7
		(286.8)	(336.5)	(492.2)		(1,326.5)
1951	1981	−358.3	−428.7	−71.6	0.136	−2,261.3
		(203.6)	(224.5)	(423.4)	(0.043)	(1,184.2)
	1982	−117.3	−278.5	−72.7		−1,386.6
		(229.1)	(264.1)	(372.1)		(1,312.1)
	1983	−314.0	−452.2	−896.5		-2,181.8
		(253.2)	(289.2)	(426.3)		(1,395.3)
	1984	−398.4	−573.3	−809.1		−2,647.9
		(279.2)	(331.1)	(380.9)		(1,529.2)
1952	1981	−342.8	−392.6	−440.5	0.105	−2,502.3
		(206.8)	(228.6)	(265.0)	(0.050)	(1,556.7)
	1982	−235.1	−255.2	−514.7		−1,626.5
		(232.3)	(264.5)	(296.5)		(1,685.8)
	1983	−437.7	−500.0	−915.7		−3,103.5
		(257.5)	(294.7)	(395.2)		(1,829.2)
	1984	−436.0	−560.0	−767.2		−3,323.8
		(281.9)	(330.1)	(376.0)		(1,959.3)

Estimating the Size of the Complier Group

- Since we never observe both $D_i(0)$ and $D_i(1)$ for the same i, we cannot identify individual units as compliers
- However, we can easily identify the proportion of compliers in the population using the "first stage" effect:

$$\begin{aligned}\Pr(\text{complier}) &= \Pr(D_i(1) - D_i(0) = 1) \\ &= \mathbb{E}[D_i(1) - D_i(0)] \\ &= \mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]\end{aligned}$$

- Using a similar logic we can identify the proportion of compliers among the treated or controls only. For example:

$$\begin{aligned}\Pr(\text{complier}|D_i = 1) &= \frac{\Pr(D_i = 1 \mid \text{complier}) \Pr(\text{complier})}{\Pr(D_i = 1)} \\ &= \frac{\Pr(Z_i = 1)(\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0])}{\Pr(D_i = 1)}\end{aligned}$$

Size of Complier Group

TABLE 4.4.2
Probabilities of compliance in instrumental variables studies

Source (1)	Endogenous Variable (D) (2)	Instrument (Z) (3)	Sample (4)	$P[D=1]$ (5)	First Stage, $P[D_1 > D_0]$ (6)	$P[Z=1]$ (7)	Compliance Probabilities $P[D_1 > D_0 \mid D=1]$ (8)	$P[D_1 > D_0 \mid D=0]$ (9)
Angrist (1990)	Veteran status	Draft eligibility	White men born in 1950	.267	.159	.534	.318	.101
			Non-white men born in 1950	.163	.060	.534	.197	.033
Angrist and Evans (1998)	More than two children	Twins at second birth	Married women aged 21–35 with two or more children in 1980	.381	.603	.008	.013	.966
		First two children are same sex		.381	.060	.506	.080	.048
Angrist and Krueger (1991)	High school graduate	Third- or fourth-quarter birth	Men born between 1930 and 1939	.770	.016	.509	.011	.034
Acemoglu and Angrist (2000)	High school graduate	State requires 11 or more years of school attendance	White men aged 40–49	.617	.037	.300	.018	.068

Notes: The table computes the absolute and relative size of the complier population for a number of instrumental variables. The first stage, reported in column 6, gives the absolute size of the complier group. Columns 8 and 9 show the size of the complier population relative to the treated and untreated populations.

Estimators for the LATE

- The LATE formula:

$$\text{LATE} = \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}$$

Estimators for the LATE

- The LATE formula:

$$\text{LATE} = \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} = \frac{\text{Cov}(Y_i, Z_i)}{\text{Cov}(D_i, Z_i)}$$

- A plug-in estimator is called the Wald estimator:

$$\widehat{\text{LATE}} = \frac{\frac{1}{n_1}\sum_{i=1}^n Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^n (1 - Z_i) Y_i}{\frac{1}{n_1}\sum_{i=1}^n Z_i D_i - \frac{1}{n_0}\sum_{i=1}^n (1 - Z_i) D_i} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(D_i, Z_i)}$$

  where $n_1 = \#$ assigned to treatment and $n_0 = n - n_1$

  - The Wald estimator is consistent, but not unbiased in finite samples
  - The small sample bias can be considerable when the instrument is weak (i.e. when $\widehat{\text{Cov}}(D_i, Z_i) \simeq 0$)
  - The relationship needs to be *more* than just statistically significant at a conventional level; a rule of thumb is $F > 10$

Estimators for the LATE

- The LATE formula:

$$\mathrm{LATE} = \frac{\mathbb{E}(Y_i \mid Z_i = 1) - \mathbb{E}(Y_i \mid Z_i = 0)}{\mathbb{E}(D_i \mid Z_i = 1) - \mathbb{E}(D_i \mid Z_i = 0)} = \frac{\mathrm{Cov}(Y_i, Z_i)}{\mathrm{Cov}(D_i, Z_i)}$$

- A plug-in estimator is called the Wald estimator:

$$\widehat{\mathrm{LATE}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i) Y_i}{\frac{1}{n_1}\sum_{i=1}^{n} Z_i D_i - \frac{1}{n_0}\sum_{i=1}^{n}(1 - Z_i) D_i} = \frac{\widehat{\mathrm{Cov}}(Y_i, Z_i)}{\widehat{\mathrm{Cov}}(D_i, Z_i)}$$

  where $n_1 = \#$ assigned to treatment and $n_0 = n - n_1$

  - The Wald estimator is consistent, but not unbiased in finite samples
  - The small sample bias can be considerable when the instrument is weak (i.e. when $\widehat{\mathrm{Cov}}(D_i, Z_i) \simeq 0$)
  - The relationship needs to be *more* than just statistically significant at a conventional level; a rule of thumb is $F > 10$

- $\widehat{\mathrm{LATE}}$ can also be calculated via two-stage least squares (2SLS), a traditional instrumental variables method in econometrics

Traditional Instrumental Variable Framework

Traditional framework considers a general setting with multiple treatments and instruments:

- Endogeneous regressors: $D_i = [D_{i1} \ \cdots \ D_{iJ}]^\top$
- Instruments: $Z_i = [Z_{i1} \ \cdots \ Z_{iL}]^\top$

Traditional Instrumental Variable Framework

Traditional framework considers a general setting with multiple treatments and instruments:

- Endogeneous regressors: $D_i = [D_{i1} \ \cdots \ D_{iJ}]^\top$
- Instruments: $Z_i = [Z_{i1} \ \cdots \ Z_{iL}]^\top$
- Let $\begin{cases} \boldsymbol{D} & = & [\mathbf{1} \ D] & \text{(2nd stage model matrix)} \\ \boldsymbol{Z} & = & [\mathbf{1} \ Z] & \text{(1st stage model matrix)} \end{cases}$
- The model:

$$Y \ = \ \boldsymbol{D}\beta + \varepsilon \quad \text{where} \quad \mathbb{E}[\varepsilon_i] = 0 \text{ and } Var(\varepsilon_i) = \sigma^2$$

- D_i is allowed to be endogeneous: $D_i \not\perp\!\!\!\perp \varepsilon_i$
- Z_i must be exogenous and excluded from the model: $Z_i \perp\!\!\!\perp \varepsilon_i$

Traditional Instrumental Variable Framework

Traditional framework considers a general setting with multiple treatments and instruments:

- Endogeneous regressors: $D_i = [D_{i1} \ \cdots \ D_{iJ}]^\top$
- Instruments: $Z_i = [Z_{i1} \ \cdots \ Z_{iL}]^\top$
- Let $\begin{cases} \boldsymbol{D} & = & [\mathbf{1} \ D] & \text{(2nd stage model matrix)} \\ \boldsymbol{Z} & = & [\mathbf{1} \ Z] & \text{(1st stage model matrix)} \end{cases}$
- The model:

$$Y \ = \ \boldsymbol{D}\beta + \varepsilon \quad \text{where} \quad \mathbb{E}[\varepsilon_i] = 0 \text{ and } Var(\varepsilon_i) = \sigma^2$$

- D_i is allowed to be endogeneous: $D_i \not\perp\!\!\!\perp \varepsilon_i$
- Z_i must be exogenous and excluded from the model: $Z_i \perp\!\!\!\perp \varepsilon_i$
- Condition for identification: At least as many instruments as endogenous regressors
- Can also incorporate observed covariates ($\boldsymbol{X}$) that is assumed to be exogenous; just include them in both stages

Estimation via Two Stage Least Squares

- The two-stage least squares (2SLS) estimator:

$$\hat{\beta}_{2SLS} = (\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top Y$$

Estimation via Two Stage Least Squares

- The two-stage least squares (2SLS) estimator:

$$\hat{\beta}_{2SLS} = (\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top Y$$

- This can be calculated by running OLS twice (hence the name):
  - Stage 1: Regress D on $\boldsymbol{Z}$ and obtain fitted values

$$\hat{D} = P_Z D \equiv \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top D$$

    where P_Z is the projection (or "hat") matrix
  - Stage 2: Regress Y on $\hat{\boldsymbol{D}} \equiv [\mathbf{1}\ \hat{D}]$

$$\begin{aligned}(\hat{\boldsymbol{D}}^\top \hat{\boldsymbol{D}})^{-1}\hat{\boldsymbol{D}}^\top Y &= (\boldsymbol{D}^\top P_Z^\top P_Z \boldsymbol{D})^{-1}\boldsymbol{D}^\top P_Z^\top Y \\ &= (\boldsymbol{D}^\top P_Z \boldsymbol{D})^{-1}\boldsymbol{D}^\top P_Z Y = \hat{\beta}_{2SLS}\end{aligned}$$

Estimation via Two Stage Least Squares

- The two-stage least squares (2SLS) estimator:
$$\hat{\beta}_{2SLS} \ = \ (\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top Y$$
- This can be calculated by running OLS twice (hence the name):
  - Stage 1: Regress D on $\boldsymbol{Z}$ and obtain fitted values
  $$\hat{D} \ = \ P_Z D \ \equiv \ \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top D$$
  where P_Z is the projection (or “hat”) matrix
  - Stage 2: Regress Y on $\hat{\boldsymbol{D}} \equiv [\mathbf{1} \ \hat{D}]$
  $$\begin{aligned}(\hat{\boldsymbol{D}}^\top \hat{\boldsymbol{D}})^{-1}\hat{\boldsymbol{D}}^\top Y \ &= \ (\boldsymbol{D}^\top P_Z^\top P_Z \boldsymbol{D})^{-1}\boldsymbol{D}^\top P_Z^\top Y \\ &= \ (\boldsymbol{D}^\top P_Z \boldsymbol{D})^{-1}\boldsymbol{D}^\top P_Z Y \ = \ \hat{\beta}_{2SLS}\end{aligned}$$
- Can show that $\hat{\beta}_{2SLS} \xrightarrow{P} \beta$ (consistency) given $\mathbb{E}[Z_i \varepsilon_i] = 0$

Estimation via Two Stage Least Squares

- The two-stage least squares (2SLS) estimator:
$$\hat{\beta}_{2SLS} \;=\; (\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D})^{-1}\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top Y$$
- This can be calculated by running OLS twice (hence the name):
  - Stage 1: Regress D on $\boldsymbol{Z}$ and obtain fitted values
  $$\hat{D} \;=\; P_Z D \;\equiv\; \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top D$$
  where P_Z is the projection (or "hat") matrix
  - Stage 2: Regress Y on $\hat{\boldsymbol{D}} \equiv [\mathbf{1}\ \hat{D}]$
  $$\begin{aligned}(\hat{\boldsymbol{D}}^\top \hat{\boldsymbol{D}})^{-1}\hat{\boldsymbol{D}}^\top Y &= (\boldsymbol{D}^\top P_Z^\top P_Z \boldsymbol{D})^{-1}\boldsymbol{D}^\top P_Z^\top Y \\ &= (\boldsymbol{D}^\top P_Z \boldsymbol{D})^{-1}\boldsymbol{D}^\top P_Z Y \;=\; \hat{\beta}_{2SLS}\end{aligned}$$
- Can show that $\hat{\beta}_{2SLS} \xrightarrow{P} \beta$ (consistency) given $\mathbb{E}[Z_i \varepsilon_i] = 0$
- Under homoskedasticity: $\mathbb{V}(\hat{\beta}_{2SLS} \mid \boldsymbol{D}, \boldsymbol{Z}) \;=\; \sigma^2 \left\{\boldsymbol{D}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D}\right\}^{-1}$
- Estimate σ^2 based on $\hat{\varepsilon} \equiv Y - \boldsymbol{D}\hat{\beta}_{2SLS}$ (don't use stage 2 residuals!)
- Can be made robust to heteroskedasticity and clustering

What's "Wrong" with the Traditional Framework?

The traditional framework appears to be more versatile because:

- It allows for multiple treatments that are not binary
- It allows for multiple instruments that are not binary
- It allows for incorporating observed covariates easily

What's "Wrong" with the Traditional Framework?

The traditional framework appears to be more versatile because:

- It allows for multiple treatments that are not binary
- It allows for multiple instruments that are not binary
- It allows for incorporating observed covariates easily

However, these come at the following costs:

- It inherently makes the constant treatment effect assumption
- It obscures the fact that the effects are identified off of the particular subpopulation who "comply" with the treatment assignment

What's "Wrong" with the Traditional Framework?

The traditional framework appears to be more versatile because:

- It allows for multiple treatments that are not binary
- It allows for multiple instruments that are not binary
- It allows for incorporating observed covariates easily

However, these come at the following costs:

- It inherently makes the constant treatment effect assumption
- It obscures the fact that the effects are identified off of the particular subpopulation who "comply" with the treatment assignment

Note that one can still use 2SLS as a nonparametric estimator of LATE with a single binary treatment and instrument

Example: Job Training Partnership Act (JTPA)

- Largest randomized training evaluation ever undertaken in the U.S.; started in 1983 at 649 sites throughout the country
- Sample: Disadvantaged persons in the labor market (previously unemployed or low earnings)
- Z_i: Assignment to the program, consisting of one of three general service strategies (`assignmt`)
  - classroom training in occupational skills
  - on-the-job training and/or job search assistance
  - other services (e.g. probationary employment)
- D_i: Actual enrollment in the assigned program (`training`)
- Y_i: Earnings 30 month after assignment (`earnings`)
- X_i: Characteristics measured before assignment (age, gender, previous earnings, race, etc.)

JTPA Example: Naïve Estimation of ATE via OLS

R Code

```
> d <- read.dta("jtpa.dta")
> summary(lm(earnings ~ training, data = d))
Call:
lm(formula = earnings ~ training, data = d)

Residuals:
   Min      1Q Median      3Q     Max
-17396 -13587  -4955    8776 141155

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14605.1      209.8  69.624   <2e-16 ***
training      2791.1      318.6   8.761   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16710 on 11202 degrees of freedom
Multiple R-squared: 0.006806,       Adjusted R-squared: 0.006717
F-statistic: 76.76 on 1 and 11202 DF,  p-value: < 2.2e-16
```

JTPA Example: Compliance Probability

R Code

```
> summary(lm(training ~ assignmt, data = d))

Call:
lm(formula = training ~ assignmt, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64165 -0.01453 -0.01453  0.35835  0.98547

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.014528   0.006529   2.225   0.0261 *
assignmt    0.627118   0.007987  78.522   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.398 on 11202 degrees of freedom
Multiple R-squared: 0.355,      Adjusted R-squared: 0.355
F-statistic:  6166 on 1 and 11202 DF,  p-value: < 2.2e-1
```

JTPA Example: Estimation of ITT Effect

R Code

```
> summary(lm(earnings ~ assignmt, data = d))

Call:
lm(formula = earnings ~ assignmt, data = d)

Residuals:
   Min     1Q Median     3Q    Max 
-16200 -13803  -4817   8950 139560 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  15040.5      274.9  54.716  < 2e-16 ***
assignmt      1159.4      336.3   3.448 0.000567 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16760 on 11202 degrees of freedom
Multiple R-squared: 0.00106,	Adjusted R-squared: 0.000971 
F-statistic: 11.89 on 1 and 11202 DF,  p-value: 0.000566
```

JTPA Example: Wald Estimator for Complier LATE

$$\text{Wald estimator for LATE} = \frac{\mathrm{Cov}(Y_i, Z_i)}{\mathrm{Cov}(D_i, Z_i)}$$

R Code

```
> cov(d[,c("earnings","training","assignmt")])

             earnings    training    assignmt
earnings 2.811338e+08 685.5254685 257.0625061
training 6.855255e+02   0.2456123   0.1390407
assignmt 2.570625e+02   0.1390407   0.221713

> 257.0625061/0.1390407
[1] 1848.829
```

JTPA Example: 2SLS for Complier LATE

R Code

```
> training_hat <- lm(training ~ assignmt, data = d)$fitted
> summary(lm(earnings ~ training_hat, data = d))

Call:
lm(formula = earnings ~ training_hat, data = d)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   15013.6      281.3  53.375  < 2e-16 ***
training_hat   1848.8      536.2   3.448 0.000567 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16760 on 11202 degrees of freedom
Multiple R-squared: 0.00106,       Adjusted R-squared: 0.000971
F-statistic: 11.89 on 1 and 11202 DF,  p-value: 0.0005669
```

Note that the standard errors are not quite right

JTPA Example: 2SLS for Complier LATE

R Code

```
> library(AER)
> summary(ivreg(earnings ~ training | assignmt, data = d))

Call:
ivreg(formula = earnings ~ training | assignmt, data = d)

Residuals:
   Min      1Q Median      3Q     Max
-16862 -13716  -4943    8834 140746

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  15013.6      280.6  53.508  < 2e-16 ***
training      1848.8      534.9   3.457 0.000549 ***
---
Residual standard error: 16720 on 11202 degrees of freedom
Multiple R-Squared: 0.00603,        Adjusted R-squared: 0.005941
Wald test: 11.95 on 1 and 11202 DF,  p-value: 0.0005491
```

Extensions

- The LATE-based interpretation can be extended in several important ways
- In each case, the 2SLS estimate is a certain weighted average of complier LATEs

Extensions

- The LATE-based interpretation can be extended in several important ways
- In each case, the 2SLS estimate is a certain weighted average of complier LATEs

1. Multiple instruments: Use $Z_i = [Z_{i1}, Z_{i2}, ..., Z_{iK}]$ to instrument a single treatment D_i
   $\Longrightarrow \hat{\beta}_{2SLS} =$ weighted average of K instrument-specific LATEs
2. Continuous or multivalued treatment: $D_i \in \mathbb{R}$ or $D_i = 0, 1, ..., J$
   - Monotonicity now becomes: $D_i(1) \geq D_i(0)$ for all i
   - Compliers are now defined for each value d in the support of D_i

   $\Longrightarrow \hat{\beta}_{2SLS} =$ weighted average of LATEs at each d
3. Covariates: Often Z_i is ignorable only after conditioning on X_i
   $\Longrightarrow \hat{\beta}_{2SLS} =$ weighted average of covariate-specific LATEs

For details, see Angrist and Pischke, Chapter 4.5.

JTPA Example: 2SLS with Covariates

R Code

```
> summary(ivreg(earnings ~ training + prevearn + sex + age + married
+             | prevearn + sex + age + married +assignmt,data = d))
Call:
ivreg(formula = earnings ~ training + prevearn + sex + age +
    married | prevearn + sex + age + married + assignmt, data = d)
Residuals:
   Min      1Q Median      3Q     Max
-58052 -10916  -4050   8316 117239

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.162e+04  6.042e+02  19.238  < 2e-16 ***
training     1.927e+03  4.998e+02   3.855 0.000116 ***
prevearn     1.270e+00  3.885e-02  32.675  < 2e-16 ***
sex          3.760e+03  3.053e+02  12.316  < 2e-16 ***
age         -9.592e+01  1.543e+01  -6.215  5.3e-10 ***
married      2.707e+03  3.488e+02   7.760  9.2e-15 ***
---
Residual standard error: 15600 on 11198 degrees of freedom
Multiple R-Squared: 0.1348,        Adjusted R-squared: 0.1344
Wald test:   335 on 5 and 11198 DF,  p-value: < 2.2e-16
```