Causal Inference and Regression Interpretation

Mikey Jarrell

MIT

January 31 – February 1, 2023

What is a **potential outcome**?

- Concrete example: effect of college on income

- D_i — treatment status for person i

$$D_i = \begin{cases} 1 & \text{if } i \text{ goes to college: ``is treated''} \\ 0 & \text{if } i \text{ doesn't go to college: ``is untreated''} \end{cases}$$

- **Potential outcomes**: hypothetical, imaginary, Schrödinger's Cat type of thing
  - Y_{1i} — income for i if i were to go to college: "potential outcome if treated"
  - Y_{0i} — income for i were i to not go to college: "potential outcome if untreated"

- Problem: for any student i, we cannot simultaneously observe both Y_{1i} and Y_{0i}.
  - One of the potential outcomes is eventually "observed."
  - Y_i — **observed outcome** for student i

We care about treatment because we care about **treatment effects.**

$$Y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i}) = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

- **Causal effect**: $Y_{1i} - Y_{0i}$
- **Average treatment effect** (ATE): $E[Y_{1i} - Y_{0i}]$
- **Treatment effect on the treated** (TOT):

$$\begin{aligned} E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] &= E[Y_{1i} - Y_{0i}|D_i = 1] \\ \text{observed} - \text{counterfactual} &= \end{aligned}$$

- What we observe: $E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \text{TOT} + \text{selection bias}$ ▸ proof

What should we do about selection bias?

- Problem: people who go to college are different than those who don't.
    - **Selection bias**: $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$
    - E.g., people who go to college are more likely have family connections to good jobs.

- Solution: assign college according to a coin flip.
    - $D_i \perp Y_{0i}, Y_{1i}$
    - Potential outcomes equal on average across two groups.
        - $E[Y_{1i}|D_i = 1] = E[Y_{1i}|D_i = 1]$
        - $E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 1]$

- Now we can compare group means to find causal effect of treatment:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \text{Average Treatment Effect (ATE)}$$ ▸ proof

- Would **random sampling** accomplish the same thing?

Unfortunately, finding causal effects is rarely that simple.

- X_i — **independent variables** (a.k.a., covariates, regressors) ← vector?
- Y_i — **dependent variable** (a.k.a., outcome, regressand)
- Assume conditional expectation function is linear, e.g., individual income is linear function of parents' income *on average.*

$$E[Y_i|X_i] = \alpha + \beta X_i$$

- There is still individual variation: **error**

$$\epsilon_i = Y_i - E[Y_i|X_i]$$

$$E[\epsilon_i] = 0$$ ▸ proof

- Regression finds the combination of intercept (α) and slope (β) ◂ proof that minimize mean squared error for that conditional expectation function. ▸ proof

Control variables

Suppose we are unable to randomize D_i,

- E.g., kids decide on their own whether to go to college; they don't listen to us.
- Selection bias $\neq 0$
- $E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \neq$ ATE

but we have two different independent variables,

- D_i — treatment status (a.k.a., **regressor of interest**)
- W_i — **control**

where W_i is correlated with D_i and with Y_i (once was have accounted for D_i).

- E.g., family wealth (W_i) gives us some information about whether i goes to college (D_i) and i's income (Y_i).

Multivariate regressions

Now we can compare "matched" group means:

$$E[Y_i|D_i = 1, W_i = w] - E[Y_i|D_i = 0, W_i = w]$$

Assuming the conditional expectation function is linear in D_i and W_i, then regression gives us

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

where β is the effect of college, and γ is the effect of family wealth.

- If D_i is random conditional on W_i, then β is ATE.
- **Conditional Independence Assumption (CIA)**: $E[D_i|W_i] \perp \epsilon_i$
- E.g., on average, difference in income between people of equal family wealth is caused by college.

Multivariate regression example

- Suppose our model of the world ("**data generating process**")

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

  where Y is income, D is college, and W is family wealth.

- We get some data, run a regression, and find:
  - $\widehat{\alpha} = \$20,000$
  - $\widehat{\beta} = \$40,000$
  - $\widehat{\gamma} = 0.1$

- Then what is the predicted income for
  - Charlie from *Willy Wonka*? (No college, $W_i = -\$100$)
  - Batman? (No college, $W_i = \$10,000,000$)
  - Preston Bezos? (College for sure, $W_i = \$100,000,000,000$)

Omitted Variable Bias

- Suppose model of the world (**long**) : $Y_i = \alpha + \rho C_i + \gamma A_i + \epsilon_i$
  - Y — income
  - C — college
  - A — ability

- A is unobservable; we can only measure (**short**) : $Y_i = \alpha^* + \rho^* C_i + \nu_i$
  - C — **included** (observable)
  - A — **omitted** (unobservable)

- One could prove the following:

$$\rho^* = \rho + \gamma \delta_{AC}$$

“short = long + effect of omitted × regression of omitted on included”

◂ proof

- For any regression, there are infinite omitted variables. When do they matter?

Omitted variable bias example

- God knows the coefficients from the long: $Y_i = \alpha + \rho C_i + \gamma A_i + \epsilon_i$
    - $\alpha = \$20,000$
    - $\rho = 0$
    - $\gamma = \$80,000$ (Suppose A_i is a dummy.)

- And from the regression of omitted on included: $A_i = \tau + \delta_{AC} C_i + \eta_i$
    - $\tau = 0$
    - $\delta_{AC} = 0.5$

- So what will I find when I estimate the short? $Y_i = \alpha^* + \rho^* C_i + \nu_i$
    - $\widehat{\alpha^*} = \$20,000$
    - $\widehat{\rho^*} = \$40,000$

Reverse Causality

- An omitted variable (a.k.a., a “confounding variable”) is the most common critique of inferring causality from a regression.
  - You claim college caused increased income, but actually, talent is the true cause of both college *and* higher income!
- A less common but equally valid critique is **reverse causality**.
  - You claim college caused increased income, but actually, increased income caused college!
- Randomization clearly solves reverse causality.
  - College attendance was determined by a coin flip; income has no effect on a coin flip.
- When employing other econometric strategies, think about whether they address omitted variable bias and reverse causality.

Difference in Differences (DiD)

$$Y_{it} = \alpha + \beta(D_i \times T_{it}) + \gamma D_i + \delta T_{it} + \epsilon_{it}$$

Where Y_{it} is an outcome for person i at time t, α is a constant, β is the treatment effect, D_i is treatment status, T_{it} is a dummy that equals 1 when time t is post-treatment, and ϵ_{it} is the error term.

- Useful when treatment not randomly assigned.

- Identifying assumption: **parallel trends**
  - Sans treatment, treatment group would have progressed just like the control group.
  - $E[Y_{0,treat,post}] - E[Y_{0,treat,pre}] = E[Y_{0,control,post}] - E[Y_{0,control,pre}]$

- If D_i and T_{it} are dummies, DiD is a difference in means.

$$\begin{aligned}
\alpha &= E[Y_{it}|D_i = 0, T_{it} = 0] \\
\gamma &= E[Y_{it}|D_i = 1, T_{it} = 0] - \alpha \\
\delta &= E[Y_{it}|D_i = 0, T_{it} = 1] - \alpha \\
\beta &= E[Y_{it}|D_i = 1, T_{it} = 1] - \alpha - \gamma - \delta
\end{aligned}$$

DiD Basic Example

- Suppose in 2030, Canada made all college free.

- Given income averages by birth year for those in US and Canada, can we construct a DiD to estimate the causal effect of the policy change on income?
  - Income in US for those born in 2000: \$50,000
  - Income in US for those born in 2020: \$60,000
  - Income in Canada for those born in 2000: \$40,000
  - Income in Canada for those born in 2020: \$55,000

- Use that data to fill in the regression (where $D_i = 1$ for Canada, and $T_i = 1$ for 2020):

$$Y_{it} = \alpha + \beta(D_i \times T_{it}) + \gamma D_i + \delta T_t + \epsilon_{it}$$
$$Y_{it} = \$50,000 + \$5,000(D_i \times T_{it}) - \$10,000 D_i + \$10,000 T_{it} + \epsilon_{it}$$

Duflo (2001) “Schooling and Labor Market Consequences of School Construction in Indonesia”

$$S_{ijk} = c + \alpha_{1j} + \beta_{1k} + (P_j T_i)\gamma_1 + (\mathbf{C_j} T_i)\delta_1 + \epsilon_{ijk} \tag{1}$$

“Where S_{ijk} is the education of individual i born in region j in year k, T_i is a dummy indicating whether the individual belongs to the ‘young’ cohort in the subsample, c_1 is a constant, β_{1k} is a cohort of birth fixed effect, α_{1j} is a district of birth fixed effect, P_j denotes the intensity of the program in the region of birth, and $\mathbf{C_j}$ is a vector of region-specific variables.”

Duflo (2001): Table 3

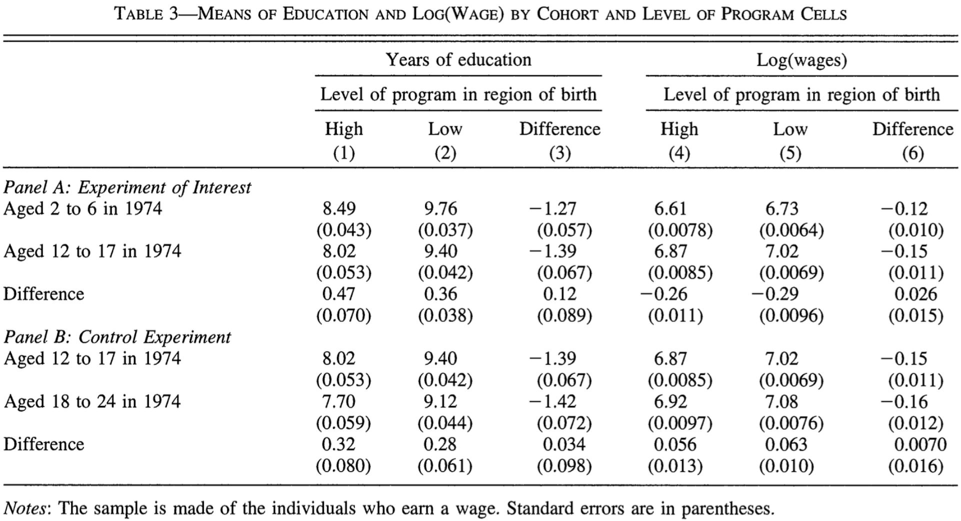

Table 3—Means of Education and Log(Wage) by Cohort and Level of Program Cells

	Years of education			Log(wages)		
	Level of program in region of birth			Level of program in region of birth		
	High (1)	Low (2)	Difference (3)	High (4)	Low (5)	Difference (6)
Panel A: Experiment of Interest						
Aged 2 to 6 in 1974	8.49	9.76	−1.27	6.61	6.73	−0.12
	(0.043)	(0.037)	(0.057)	(0.0078)	(0.0064)	(0.010)
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Difference	0.47	0.36	0.12	−0.26	−0.29	0.026
	(0.070)	(0.038)	(0.089)	(0.011)	(0.0096)	(0.015)
Panel B: Control Experiment						
Aged 12 to 17 in 1974	8.02	9.40	−1.39	6.87	7.02	−0.15
	(0.053)	(0.042)	(0.067)	(0.0085)	(0.0069)	(0.011)
Aged 18 to 24 in 1974	7.70	9.12	−1.42	6.92	7.08	−0.16
	(0.059)	(0.044)	(0.072)	(0.0097)	(0.0076)	(0.012)
Difference	0.32	0.28	0.034	0.056	0.063	0.0070
	(0.080)	(0.061)	(0.098)	(0.013)	(0.010)	(0.016)

Notes: The sample is made of the individuals who earn a wage. Standard errors are in parentheses.

Natura non facit saltus.

- Goal of RCTs is to create two groups that look the same (in expectation):

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

- Sometimes RCTs are impossible, but we find two groups whose potential outcomes should look the same in expectation: **natural experiment.**

- Borders/cutoffs often create “natural” (man-made) experiments. Examples?

Comparison at the cutoff example

- We seek effect of college (D_i) on income (Y_i), but unobservable ability (A_i) confounds:

$$Y_i = \alpha + \beta D_i + \gamma A_i + \eta_i$$

- Imagine a world with only one college: it has a strict SAT score (X_i) requirement, and nobody turns down an acceptance: $D_i = 1$ if $X_i \geq 1400$ and 0 otherwise.

- Then compare those who just barely got in to those who just barely got rejected.

$$\mathrm{E}[Y_i|X_i = 1400] - \mathrm{E}[Y_i|X_i = 1399] = \beta \quad \leftarrow \text{for whom?}$$

  - Assumption: $\mathrm{E}[A_i|X_i = 1400] = \mathrm{E}[A_i|X_i = 1399]$

- But those with 1400 are smarter! Control for X_i. How? Expand **bandwidth** and run:

$$Y_i = \alpha + \beta \mathbb{1}(X_i \geq 1400) + \delta X_i + \epsilon_i$$

Regression Discontinuity Design (RDD (or RD)) setup

- Treatment is determined by a strict cut-off rule:

$$D_i = \begin{cases} 1 & \text{if } x_i \geq x_0 \\ 0 & \text{if } x_i < x_0 \end{cases}$$

- Then the regression equation looks like this:

$$\begin{aligned} Y_i &= \alpha + \beta x_i + \rho D_i + \eta_i \\ Y_i &= \alpha + \beta x_i + \rho \mathbb{1}(x_i \geq x_0) + \eta_i \end{aligned}$$

- Assumption: absent D_i, $E[Y_i]$ is linear function of the **running variable** x_i.

$$E[Y_{0i}] = \alpha + \beta x_i \quad \forall x_i$$

RD near the cut-off

- Might be too ambitious to assume same functional form everywhere.

- In practice, better to just zoom in to the cutoff.

$$
\begin{aligned}
&\lim_{x_i \downarrow x_0} E[Y_i|x_i] - \lim_{x_i \uparrow x_0} E[Y_i|x_i] \\
&= E[Y_{1i}|x_i = x_0] - E[Y_{0i}|x_i = x_0] \\
&= \rho
\end{aligned}
$$

- Relies on assumption that, for small values of δ (bandwidth),

$$
\begin{aligned}
E[Y_i|x_0 - \delta < x_i < x_0] &\approx E[Y_{0i}|x_i = x_0] \\
E[Y_i|x_0 < x_i < x_0 + \delta] &\approx E[Y_{1i}|x_i = x_0]
\end{aligned}
$$

Dell (2009): “The Persistent Effects of Peru’s Mining *Mita*”

$$c_{idb} = \alpha + \gamma \, mita_d + X'_{id}\beta + f(\text{geographic location}_d) + \boldsymbol{\phi}_b + \epsilon_{idb}$$

“where c_{idb} is the outcome variable of interest for observation i in district d along segment b of the *mita* boundary, and $mita_d$ is an indicator equal to 1 if district d contributed to the *mita* and equal to 0 otherwise; X'_{id} is a vector of covariates that includes the mean area weighted elevation and slope for district d and (in regressions with equivalent household consumption on the left-hand side) demographic variables giving the number of infants, children, and adults in the household; $f(\text{geographic location}_d)$ is the RD polynomial, which controls for smooth functions of geographic location. Various forms will be explored. Finally, $\boldsymbol{\phi}_b$ is a set of boundary segment fixed effects that denote which of four equal length segments of the boundary is the closest to the observation’s district capital.”

Dell (2009): Table II

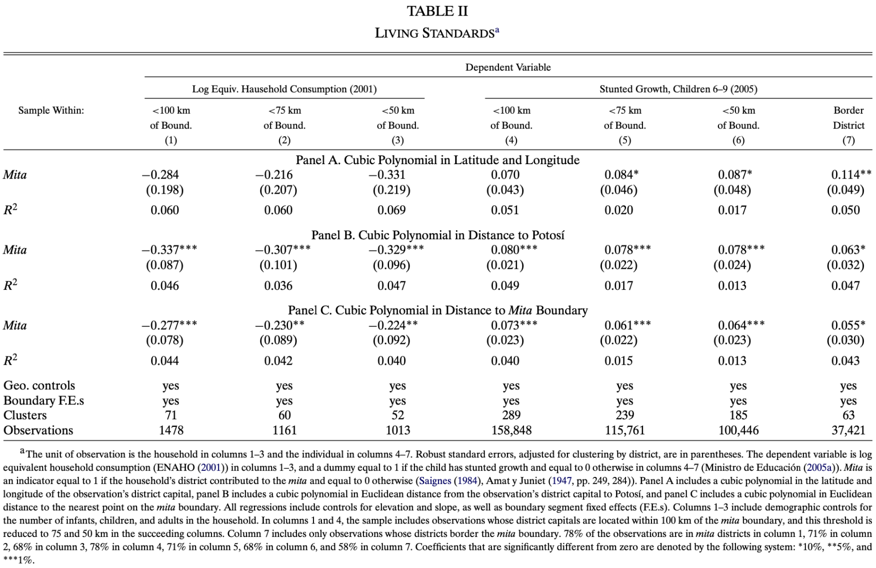

TABLE II

LIVING STANDARDS[a]

	Dependent Variable						
	Log Equiv. Hausehold Consumption (2001)			Stunted Growth, Children 6–9 (2005)			
Sample Within:	<100 km of Bound. (1)	<75 km of Bound. (2)	<50 km of Bound. (3)	<100 km of Bound. (4)	<75 km of Bound. (5)	<50 km of Bound. (6)	Border District (7)
	Panel A. Cubic Polynomial in Latitude and Longitude						
Mita	−0.284	−0.216	−0.331	0.070	0.084*	0.087*	0.114**
	(0.198)	(0.207)	(0.219)	(0.043)	(0.046)	(0.048)	(0.049)
R^2	0.060	0.060	0.069	0.051	0.020	0.017	0.050
	Panel B. Cubic Polynomial in Distance to Potosí						
Mita	−0.337***	−0.307***	−0.329***	0.080***	0.078***	0.078***	0.063*
	(0.087)	(0.101)	(0.096)	(0.021)	(0.022)	(0.024)	(0.032)
R^2	0.046	0.036	0.047	0.049	0.017	0.013	0.047
	Panel C. Cubic Polynomial in Distance to *Mita* Boundary						
Mita	−0.277***	−0.230**	−0.224**	0.073***	0.061***	0.064***	0.055*
	(0.078)	(0.089)	(0.092)	(0.023)	(0.022)	(0.023)	(0.030)
R^2	0.044	0.042	0.040	0.040	0.015	0.013	0.043
Geo. controls	yes	yes	yes	yes	yes	yes	yes
Boundary F.E.s	yes	yes	yes	yes	yes	yes	yes
Clusters	71	60	52	289	239	185	63
Observations	1478	1161	1013	158,848	115,761	100,446	37,421

[a] The unit of observation is the household in columns 1–3 and the individual in columns 4–7. Robust standard errors, adjusted for clustering by district, are in parentheses. The dependent variable is log equivalent household consumption (ENAHO (2001)) in columns 1–3, and a dummy equal to 1 if the child has stunted growth and equal to 0 otherwise in columns 4–7 (Ministro de Educación (2005a)). *Mita* is an indicator equal to 1 if the household's district contributed to the *mita* and equal to 0 otherwise (Saignes (1984), Amat y Juniet (1947, pp. 249, 284)). Panel A includes a cubic polynomial in the latitude and longitude of the observation's district capital, panel B includes a cubic polynomial in Euclidean distance from the observation's district capital to Potosí, and panel C includes a cubic polynomial in Euclidean distance to the nearest point on the *mita* boundary. All regressions include controls for elevation and slope, as well as boundary segment fixed effects (F.E.s). Columns 1–3 include demographic controls for the number of infants, children, and adults in the household. In columns 1 and 4, the sample includes observations whose district capitals are located within 100 km of the *mita* boundary, and this threshold is reduced to 75 and 50 km in the succeeding columns. Column 7 includes only observations whose districts border the *mita* boundary. 78% of the observations are in *mita* districts in column 1, 71% in column 2, 68% in column 3, 78% in column 4, 71% in column 5, 68% in column 6, and 58% in column 7. Coefficients that are significantly different from zero are denoted by the following system: *10%, **5%, and ***1%.

Agte and Bernhardt (2023): “The Economics of Caste Norms: Purity, Status, and Women’s Work in India”

$$y_{i,v} = \alpha + \gamma \mathit{East} + f(\text{location}_v) + \beta X_{i,v} + \epsilon_{i,v}$$

“where $y_{i,v}$ is the outcome of interest for individual i in village v and *East* is an indicator variable equal to 1 if the village is on the eastern side of the Mahanadi River boundary and zero otherwise. $f(\text{location}_v)$ is the RD polynomial, which controls for smooth functions of geographic location for village v. X_i is a vector of covariates for individual i, which include age, survey date fixed effects, and enumerator fixed effects.

Agte and Bernhardt (2023): Tabel 5

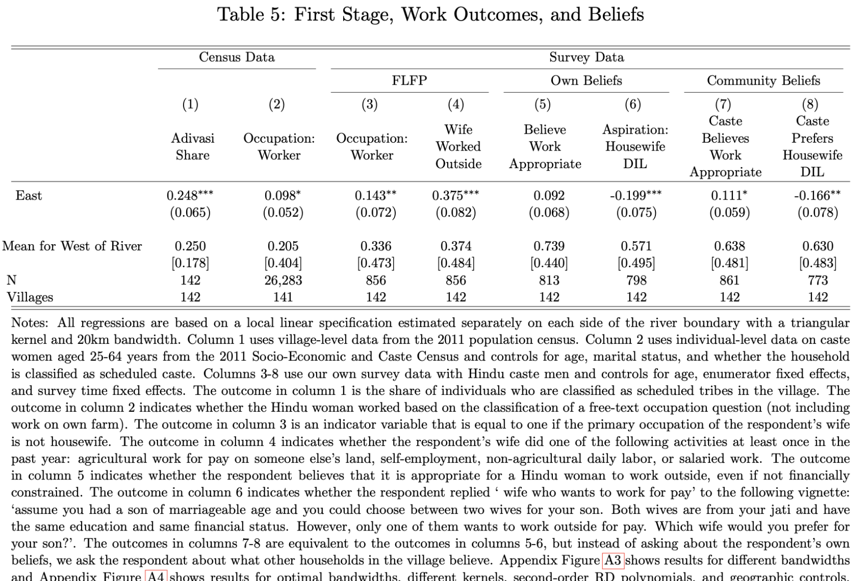

Table 5: First Stage, Work Outcomes, and Beliefs

	Census Data		Survey Data					
			FLFP		Own Beliefs		Community Beliefs	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Adivasi Share	Occupation: Worker	Occupation: Worker	Wife Worked Outside	Believe Work Appropriate	Aspiration: Housewife DIL	Caste Believes Work Appropriate	Caste Prefers Housewife DIL
East	0.248***	0.098*	0.143**	0.375***	0.092	-0.199***	0.111*	-0.166**
	(0.065)	(0.052)	(0.072)	(0.082)	(0.068)	(0.075)	(0.059)	(0.078)
Mean for West of River	0.250	0.205	0.336	0.374	0.739	0.571	0.638	0.630
	[0.178]	[0.404]	[0.473]	[0.484]	[0.440]	[0.495]	[0.481]	[0.483]
N	142	26,283	856	856	813	798	861	773
Villages	142	141	142	142	142	142	142	142

Notes: All regressions are based on a local linear specification estimated separately on each side of the river boundary with a triangular kernel and 20km bandwidth. Column 1 uses village-level data from the 2011 population census. Column 2 uses individual-level data on caste women aged 25-64 years from the 2011 Socio-Economic and Caste Census and controls for age, marital status, and whether the household is classified as scheduled caste. Columns 3-8 use our own survey data with Hindu caste men and controls for age, enumerator fixed effects, and survey time fixed effects. The outcome in column 1 is the share of individuals who are classified as scheduled tribes in the village. The outcome in column 2 indicates whether the Hindu woman worked based on the classification of a free-text occupation question (not including work on own farm). The outcome in column 3 is an indicator variable that is equal to one if the primary occupation of the respondent's wife is not housewife. The outcome in column 4 indicates whether the respondent's wife did one of the following activities at least once in the past year: agricultural work for pay on someone else's land, self-employment, non-agricultural daily labor, or salaried work. The outcome in column 5 indicates whether the respondent believes that it is appropriate for a Hindu woman to work outside, even if not financially constrained. The outcome in column 6 indicates whether the respondent replied ' wife who wants to work for pay' to the following vignette: 'assume you had a son of marriageable age and you could choose between two wives for your son. Both wives are from your jati and have the same education and same financial status. However, only one of them wants to work outside for pay. Which wife would you prefer for your son?'. The outcomes in columns 7-8 are equivalent to the outcomes in columns 5-6, but instead of asking about the respondent's own beliefs, we ask the respondent about what other households in the village believe. Appendix Figure A3 shows results for different bandwidths and Appendix Figure A4 shows results for optimal bandwidths, different kernels, second-order RD polynomials, and geographic controls.

De-meaned regression

- We seek the effect of studying philosophy (P_i) on income, so we estimate

$$Y_i = \alpha + \beta P_i + \epsilon_i$$

- We find $\widehat{\beta} = \$50,000$. Not what we expected. Why?
  - Hypothesis: Rich kids study philosophy, rich kids go to fancy schools, ↑ income.
  - Test: subset to people within just one school, rerun regression.

$$Y_i = \alpha_{\text{MIT}} + \beta_{\text{MIT}} P_i + \epsilon_i \quad \forall i \in \text{MIT}$$

- Similarly, we can subtract the average income at school s $(\overline{Y_s})$ from the income of each person i at school s (Y_{is}).

$$Y_{is} - \overline{Y_s} = \beta_{\text{de-meaned}} P_{is} + \epsilon_{is}$$

Fixed Effects (FEs): de-meaning or controls?

- Remember, including control W_i in regression allows us to compare within groups:

$$\mathrm{E}[Y_i|D_i = 1, W_i = w] = \mathrm{E}[Y_i|D_i = 0, W_i = w]$$

- We could run our within-school comparisons by including dummies for each school.

$$Y_i = \alpha + \beta P_i + \gamma_1 \mathrm{MIT}_i + \gamma_2 \mathrm{Harvard}_i + \gamma_3 \mathrm{BU}_i + \gamma_4 \mathrm{Berklee} + \gamma_5 \mathrm{Tufts}_i + \gamma_6 \mathrm{BC}_i \ldots$$

- Software does this automatically with **fixed effects** (FEs), which we write like this,

$$Y_{is} = \alpha_s + \beta P_{is} + \epsilon_{is}$$

where α_s signifies a series of dummies, one for each school in the dataset, essentially letting each school have its own intercept.

Dell, Jones, Olken (2012): Temperature and Growth

- Levels regression: $g_i = \alpha + \beta T_i + \epsilon_i$

- De-meaned regression: $g_{it} - \overline{g_i} = \alpha + \beta(T_{it} - \overline{T_i}) + \epsilon_{it}$

- Fixed effects regression: $g_{it} = \alpha_i + \beta T_{it} + \epsilon_{it}$

- What they do in the paper:

$$g_{it} = \theta_i + \theta_{rt} + \sum_{j=0}^{L} \rho_j T_{it-j} + \epsilon_{it}$$

"where θ_i are country fixed effects, θ_{rt} are time fixed effects (interacted separately with region dummies and a poor country dummy in our main specifications), ϵ_{it} is an error term clustered simultaneously by country and region-year, and T_{it} is a vector of annual average temperature and precipitation with up to L lags included."

Interactions

- Adding an interaction term to the fixed effects regression:

$$g_{irt} = \alpha_i + \gamma_{rt} + \beta T_{rt} + \tau T_{rt} \times POOR_i + \epsilon_{irt}$$

- Suppose $POOR_i$ is binary. Interpret these hypothetical results:
  - $\beta \approx 0$, $\tau < 0$
  - $\beta > 0$, $\tau \approx 0$
  - $\beta < 0$, $\tau > 0$

- Suppose $POOR_i$ is continuous. How does that change things?

de Mel et al. (2008) “Return to capital in microenterprises”

- RCT: give cash to businesses.

$$Y_{it} = \alpha + \sum_{g=1}^{4} \beta_g \mathit{Treatment}_{git} + \sum_{t=2}^{9} \delta_t + \lambda_i + \epsilon_{it}$$

  “where Y represents the outcome of interest; $g = 1$ to 4, the four treatment types granted to enterprise i any time before wave t; δ_t are wave fixed effects and λ_i are enterprise fixed effects.”

- Why include controls (in this case fixed effects) in an RCT?

- Why does one of the sums start from 2?

de Mel et al. (2008): Table II

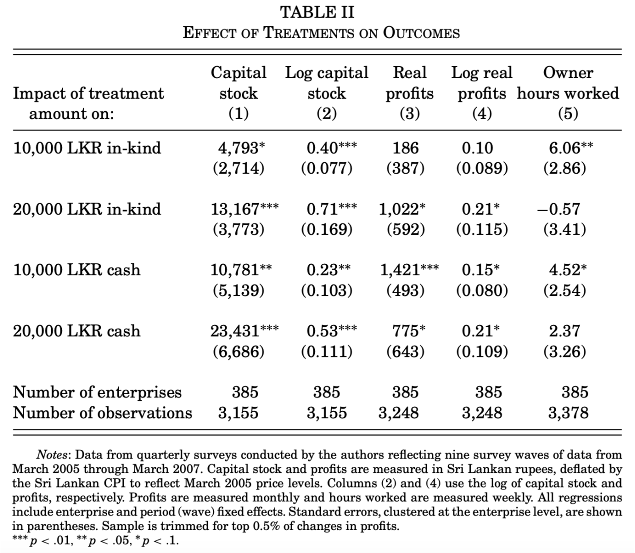

TABLE II
EFFECT OF TREATMENTS ON OUTCOMES

Impact of treatment amount on:	Capital stock (1)	Log capital stock (2)	Real profits (3)	Log real profits (4)	Owner hours worked (5)
10,000 LKR in-kind	4,793*	0.40***	186	0.10	6.06**
	(2,714)	(0.077)	(387)	(0.089)	(2.86)
20,000 LKR in-kind	13,167***	0.71***	1,022*	0.21*	−0.57
	(3,773)	(0.169)	(592)	(0.115)	(3.41)
10,000 LKR cash	10,781**	0.23**	1,421***	0.15*	4.52*
	(5,139)	(0.103)	(493)	(0.080)	(2.54)
20,000 LKR cash	23,431***	0.53***	775*	0.21*	2.37
	(6,686)	(0.111)	(643)	(0.109)	(3.26)
Number of enterprises	385	385	385	385	385
Number of observations	3,155	3,155	3,248	3,248	3,378

Notes: Data from quarterly surveys conducted by the authors reflecting nine survey waves of data from March 2005 through March 2007. Capital stock and profits are measured in Sri Lankan rupees, deflated by the Sri Lankan CPI to reflect March 2005 price levels. Columns (2) and (4) use the log of capital stock and profits, respectively. Profits are measured monthly and hours worked are measured weekly. All regressions include enterprise and period (wave) fixed effects. Standard errors, clustered at the enterprise level, are shown in parentheses. Sample is trimmed for top 0.5% of changes in profits.
$^{***}p < .01$, $^{**}p < .05$, $^{*}p < .1$.

Instrumental Variables (IV) or Two-Stage Least Square (2SLS)

- We want to estimate $Y_i = \alpha + \beta D_i + \epsilon_i$, but D_i is not randomly assigned. We can use Z_i as an **instrument** for D_i if two assumptions hold:

- **Relevance**: $\text{Cov}[Z_i, D_i] \neq 0$
  - If this holds, then we have a valid **first stage** regression: $D_i = \tau + \phi Z_i + \eta_i$

- **Excludability/exogeneity**: $\text{Cov}[Z_i, \epsilon_i] = 0$
  - Excludability can be broken down into:
    - **As-good-as-random**: Z_i does not need to be randomly assigned, but it needs to be uncorrelated with any other X_i that is unobserved (not controlled for) that affects on Y_i.
    - **Exclusion restriction**: Z_i affects Y_i only through its effect on D_i.
  - If this holds, then we have a valid **reduced form** regression: $Y_i = \theta + \rho Z_i + \nu_i$

- Wald: $\beta_{IV} = \frac{\rho}{\phi}$

IV example

- Model of the world: $Y_i = \alpha + \beta D_i + \epsilon_i$
- D_i (treatment, i.e., college) not randomly assigned; selection bias.
- Randomly assign lottery winners (Z_i): free college!
  - Winners $(Z_i = 1)$: 70% go to college, make \$70,000 on average.
  - Losers $(Z_i = 0)$: 50% go to college, make \$60,000.

$$\begin{aligned}
\text{First Stage: } D_i &= \tau + \phi Z_i + \eta_i \\
D_i &= 0.5 + 0.2 Z_i + \eta_i \\
\text{solve for predicted } Z_i\text{: } \widehat{Z}_i &= -2.5 + 5 D_i \\
\text{Reduced Form: } Y_i &= \theta + \rho Z_i + \nu_i \\
Y_i &= \$60,000 + \$10,000 Z_i + \nu_i \\
\text{plug in Z: } Y_i &= \$35,000 + \$50,000 D_i + \epsilon_i
\end{aligned}$$

How to think about IV/2SLS

- Z_i causes some people i to take up treatment D_i — **compliers**
  - The scholarship caused a some people to go to college who otherwise wouldn't have.

- Some people would have taken up D_i even if they hadn't gotten Z_i — **always-takers**
  - Rich people don't need the scholarship.

- Others wouldn't have taken up D_i even if they had gotten Z_i — **never-takers**
  - If you want be a mechanic, you have no use for a scholarship.

- For the latter two groups, we have no way of finding out the counterfactual:
  - For always-takers, what is $E[Y_{0i}]$?
  - For never-takers, what is $E[Y_{1i}]$?

- β_{IV} tells us the effect of D_i on compliers: the **local average treatment effect** (LATE).
  - The effect of college for those who wouldn't have gone if not for the scholarship.

More ways to think about IV/2SLS

- Model of the world — effect (β) of college (D_i) on income (Y_i): $Y_i = \alpha + \beta D_i + \epsilon_i$

- Estimate **first stage** — effect (ϕ) of lottery (Z_i) on college (D_i): $D_i = \tau + \phi Z_i + \eta_i$

- From here, can go one of two different directions to arrive at the same destination:
  1. Two-stage least squares (2SLS)
     - Predict treatment (college, $\widehat{D}_i$) with instrument (lottery, Z_i): $\widehat{D}_i = \widehat{\tau} + \widehat{\phi} Z_i$
     - Plug predicted treatment into model, estimate **second stage**: $Y_i = \alpha + \beta_{2SLS} \widehat{D}_i + \epsilon_i$
  2. Instrumental variables (IV)
     - Estimate **reduced form**, effect of instrument (lottery, Z_i) on outcome (income, Y_i): $Y_i = \theta + \rho Z_i + \nu_i$
     - Scale up reduced form by first stage to get Wald: $\beta_{IV} = \frac{\rho}{\phi}$

- Luckily, $\beta_{IV} = \beta_{2SLS}$, so only one of these two intuitions needs to make sense to you.
  1. 2SLS: *The effect of college for people caused by the lottery to go to college.*
  2. IV: *If the lottery worked on everybody, this is how big the effect would be.*

Selection Bias Proof

$$
\begin{align}
& E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \tag{1} \\
& = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \tag{2} \\
& = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 1] \tag{3} \\
& = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \tag{4} \\
& = E[Y_{1i} - E_{0i}|D_i = 1] + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \tag{5} \\
& = \text{TOT} + \text{Selection Bias} \tag{6}
\end{align}
$$

◂ Go Back

Proof that random assignment eliminates selection bias

We begin by comparing group means:

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

Because $D_i \perp Y_{0i}, Y_{1i}$, the potential outcomes of the groups are the same in expectation, then,

$$\begin{aligned}
&= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] \\
&= E[Y_{1i} - Y_{0i}|D_i = 1] \\
&= E[Y_{1i} - Y_{0i}] \\
&= \text{Average treatment effect (ATE)}
\end{aligned}$$

◂ go back

Proof that ϵ is mean zero

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{1}$$

$$\epsilon_i = Y_i - \alpha - \beta X_i \tag{2}$$

$$\epsilon_i = Y_i - E[Y_i|X_i] \tag{3}$$

$$E[\epsilon_i] = E[Y_i] - E\big[E[Y_i|X_i]\big] \tag{4}$$

$$E[\epsilon_i] = E[Y_i] - E[Y_i] \tag{5}$$

$$E[\epsilon_i] = 0 \tag{6}$$

◂ Go Back

Finding α and β in bivariate regression

$$E[\epsilon_i] = 0 \tag{1}$$

$$E[Y_i - \alpha - \beta X_i] = 0 \tag{2}$$

$$\alpha = E[Y_i] - \beta E[X_i] \tag{3}$$

Referring back to (2)...

$$E\big[X_i(Y_i - \alpha - \beta X_i)\big] = 0 \tag{4}$$

$$E\big[X_i(Y_i - \big(E[Y_i] - \beta E[X_i]\big) - \beta X_i)\big] = 0 \tag{5}$$

$$E[X_i Y_i] - E[X_i]E[Y_i] + \beta\big(E[X_i]\big)^2 - \beta E[X_i{}^2] = 0 \tag{6}$$

$$\beta = \frac{E[X_i Y_i] - E[X_i]E[Y_i]}{E[X_i{}^2] - \big(E[X_i]\big)^2} \tag{7}$$

$$\beta = \frac{\mathrm{Cov}(X_i, Y_i)}{\mathrm{V}(X_i)} \tag{8}$$

◂ Go Back

Proof that α and β minimize MSE

$$\text{argmin}_{a,b}\, E\left[(Y_i - \epsilon_i)^2\right] \tag{1}$$

$$= \text{argmin}_{a,b}\, E\left[\left(Y_i - (a + bX_i)\right)^2\right] \tag{2}$$

$$= \text{argmin}_{a,b}\, E[Y_i^2] - 2aE[Y_i] - 2bE[X_i Y_i] + a^2 + 2abE[X_i] + b^2 E[X_i^2] \tag{3}$$

$$\tag{4}$$

First order condition for a:

$$0 = \frac{\partial E[Y_i^2] - 2aE[Y_i] - 2bE[X_i Y_i] + a^2 + 2abE[X_i] + b^2 E[X_i^2]}{\partial a} \tag{5}$$

$$= 0 - 2E[Y_i] - 0 + 2a + 2bE[X_i] + 0 \tag{6}$$

$$a = E[Y_i] - bE[X_i] \tag{7}$$

▸ continued...

Proof that α and β minimize MSE

First order condition for b:

$$
\begin{aligned}
0 &= \frac{\partial E[{Y_i}^2] - 2aE[Y_i] - 2bE[X_i Y_i] + a^2 + 2abE[X_i] + b^2 E[{X_i}^2]}{\partial b} && (8)\\
&= 0 - 0 - 2E[X_i Y_i] + 0 + 2aE[X_i + 2bE[{X_i}^2]] && (9)\\
&= -E[X_i Y_i] + E[X_i Y_i] - b\big(E[X_i]\big)^2 + bE[{X_i}^2] && (10)\\
b &= \frac{E[X_i Y_i] - E[X_i]E[Y_i]}{E[{X_i}^2] - \big(E[X_i]\big)^2} && (11)\\
&= \beta && (12)
\end{aligned}
$$

Plugging back into (7)...

$$
\begin{aligned}
a &= E[Y_i] - \beta E[X_i] && (13)\\
&= \alpha && (14)
\end{aligned}
$$

◂ Go Back

Proof that Long $=$ Short $+$ blah $\times$ blah

- Long: $Y_i = \alpha + \rho C_i + \gamma A_i + \epsilon_i$
- Short: $Y_i = \alpha^* + \rho^* C_i + \nu_i$
- Regression of omitted on included: $A_i = \tau + \delta_{AC} C_i + \eta_i$, so $\delta_{AC} = \frac{\text{Cov}(A_i, C_i)}{\text{V}(C_i)}$

$$\begin{aligned}
\rho^* &= \frac{\text{Cov}(Y_i, C_i)}{\text{V}(C_i)} \\
&= \frac{\text{Cov}(\alpha + \rho C_i + \gamma A_i + \epsilon_i, C_i)}{\text{V}(C_i)} \\
&= \frac{\text{Cov}(\alpha, C_i) + \text{Cov}(\rho C_i, C_i) + \text{Cov}(\gamma A_i, C_i) + \text{Cov}(\epsilon_i, C_i)}{\text{V}(C_i)} \\
&= \frac{0 + \rho\,\text{V}(C_i) + \gamma\,\text{Cov}(A_i, C_i) + 0}{\text{V}(C_i)} \\
&= \rho + \gamma \delta_{AC}
\end{aligned}$$

◂ Go Back