ECON 7710

Correlation, Causation & Potential Outcomes

Chris Cornwell

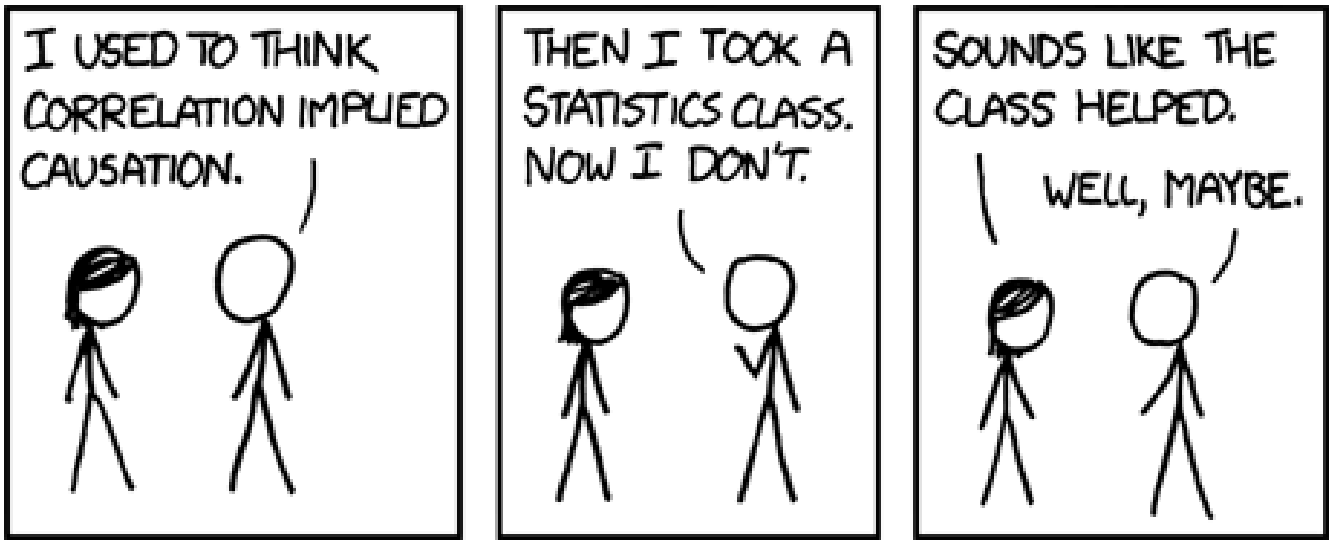
Terry College of Business

Fall 2021

Section 1

Correlation vs Causation

Correlation is not causation



Source: xkcd

Causation does not imply correlation

If X causes Y, the pdf of Y conditional on X, $f(y|x)$, must be a function of x.

It is possible for $f(y|x)$ to depend on x while $\text{corr}(x, y) = 0$.

Clearly, there can be no causal relationship if $f(y|x) = f(y)$, i.e. if X and Y are independent.

Homicides and ice cream

CRIME

When Ice Cream Sales Rise, So Do Homicides. Coincidence, or Will Your Next Cone Murder You?

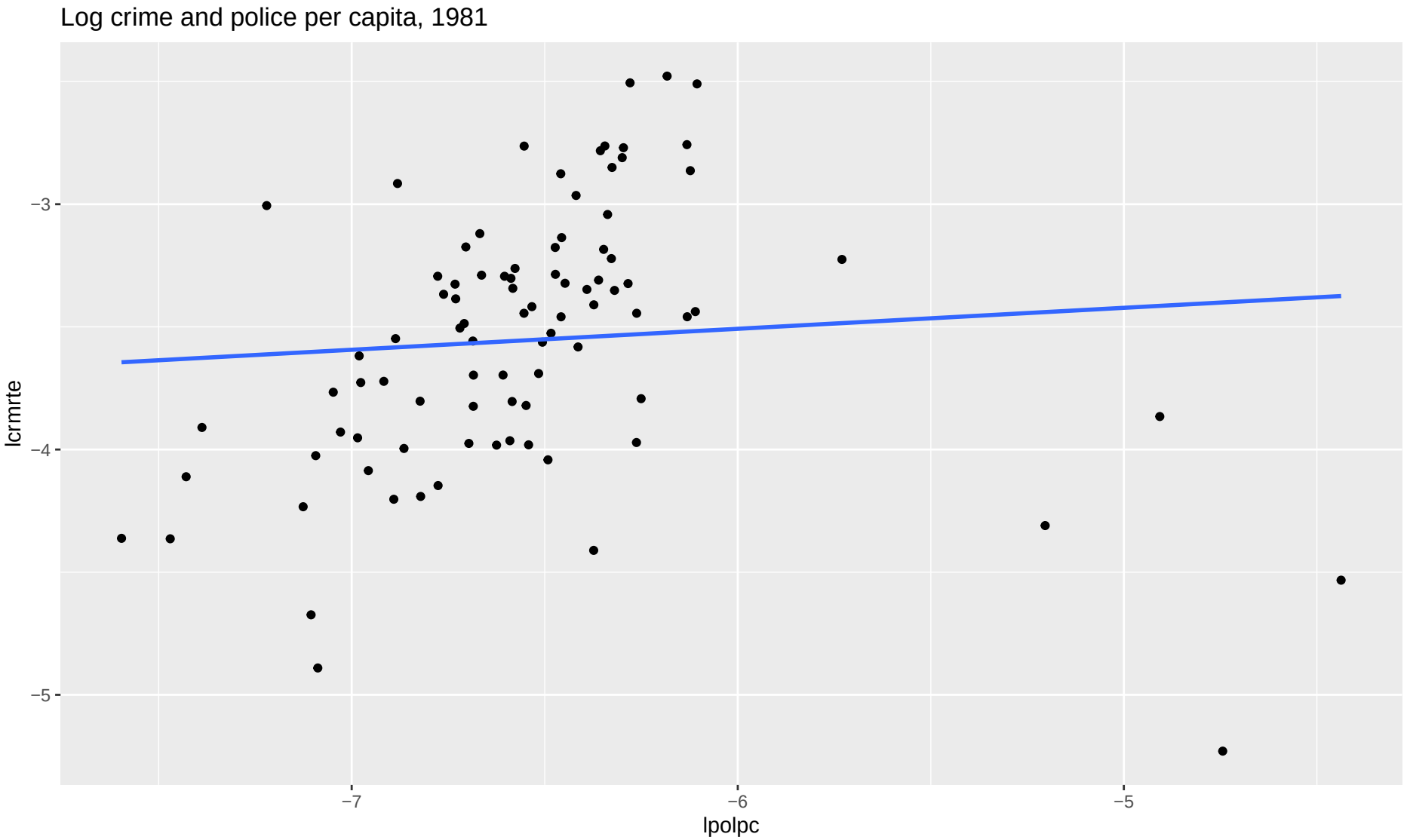
BY JUSTIN PETERS JULY 09, 2013 • 2:59 PM

Selling a boy an ice cream cone, or a murder magnet?

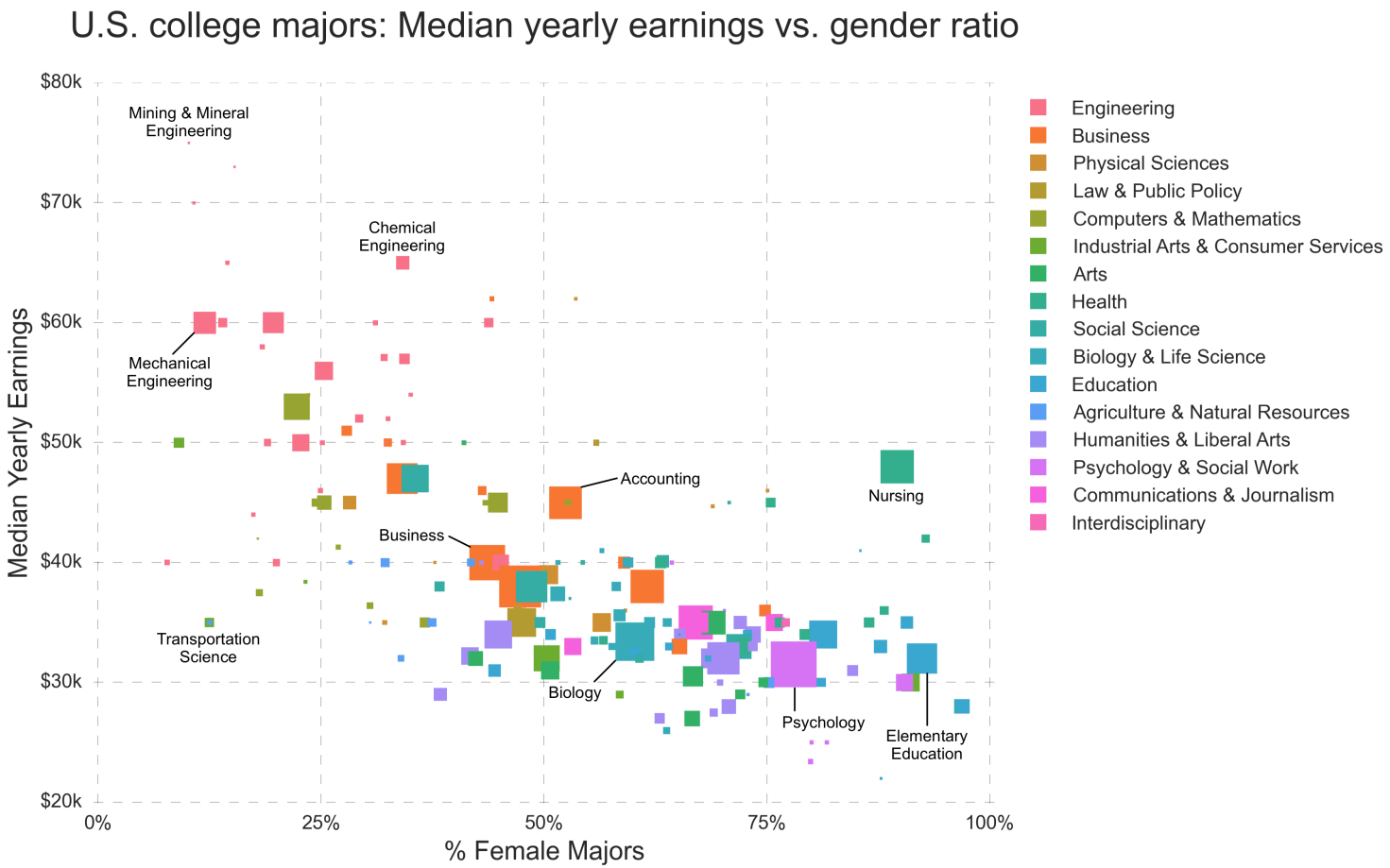
Photo by Andrew Burton/Getty Images

Source: Slate

Crime and policing

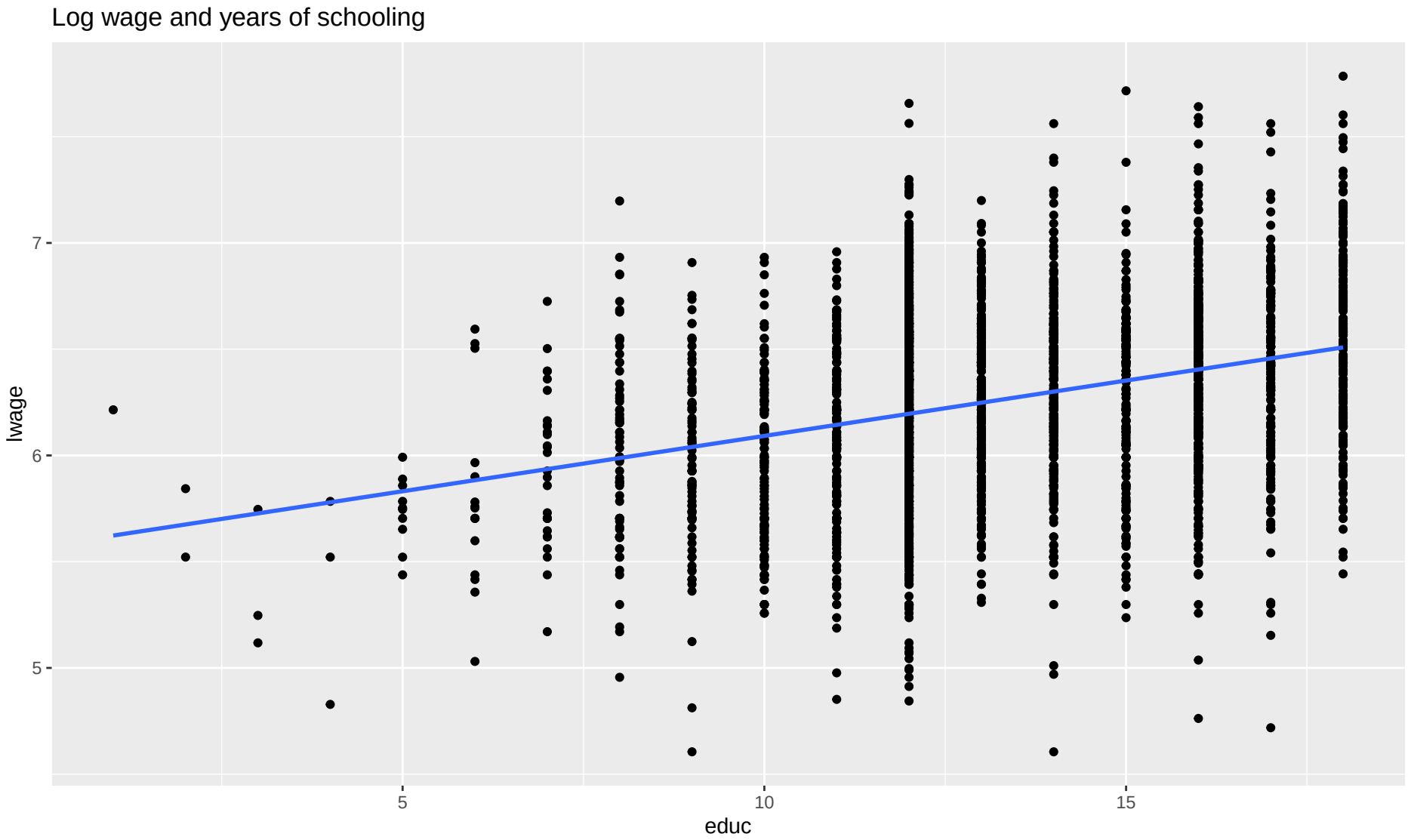


Earnings and female share of college major

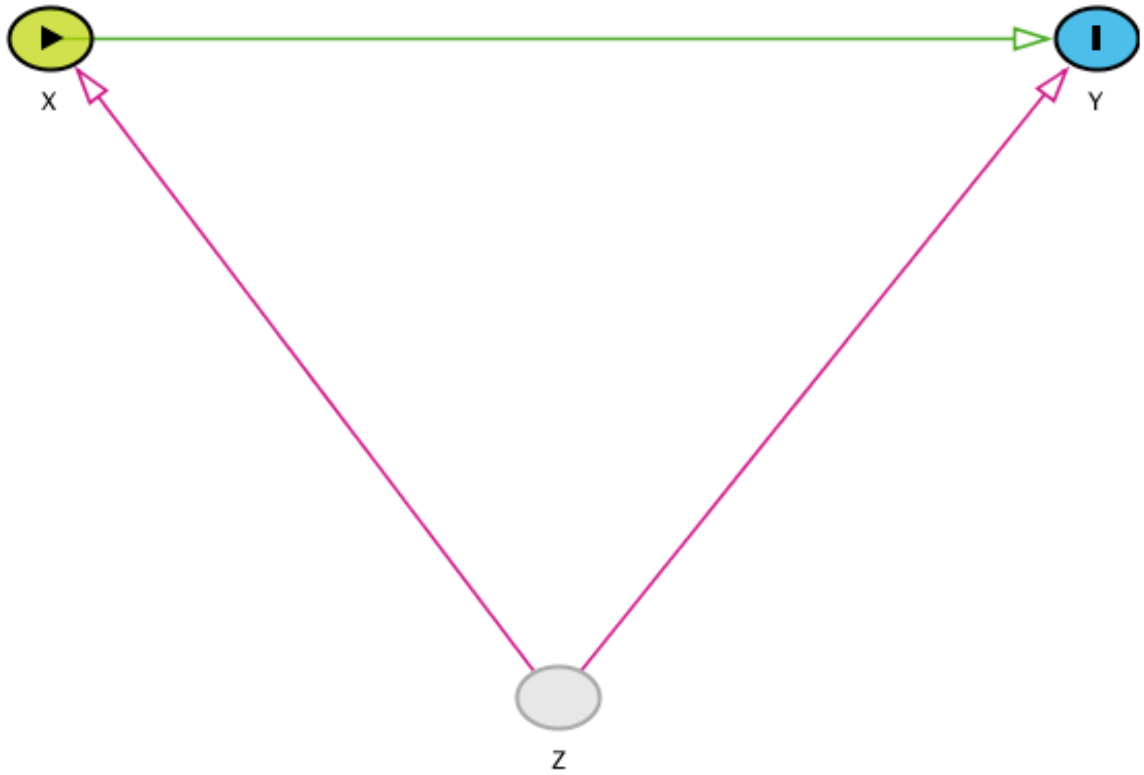


Data source: github.com/fivethirtyeight/data/tree/master/college-majors
Author: Randy Olson (randalolson.com / @randal_olson)
Notes: Each square is a college major sized by the square root of the number of recent graduates
The largest major, Psychology, has had 393,735 recent graduates

Wages and education



Confounding



Section 2

Potential outcomes

What is the Endgame?

14,000,605 possibilities

The road not taken

Storytelling

Causal inference does not happen without a good story. That story starts economic theory or background knowledge and ends with an identifiable estimand:

Economic theory, background knowledge

$\rightarrow$ Assumptions

$\rightarrow$ Causal Model

$\rightarrow$ Testable Implications

$\rightarrow$ Estimand

$\rightarrow$ Econometric methods

See Pearl's "Inference Engine" in *Book of Why*.

Potential outcomes set up (Rubin Causal Model)

Let D represent a binary *treatment*,

$$D = \begin{cases} 1, \text{if treated} \\ 0, \text{otherwise,} \end{cases}$$

which may have a causal relationship with outcome y. For a given individual, there are two possible states of the world or *potential outcomes*:

$$y_1 = \text{outcome if } D = 1$$
$$y_0 = \text{outcome if } D = 0.$$

Unless you have the Time Stone, you only observe one or the other:

$$\text{observed } y = \begin{cases} y_1, \text{if } D = 1 \\ y_0, \text{if } D = 0, \text{or} \end{cases}$$

$$y = (1 - D)y_0 + D\,y_1 = y_0 + \underbrace{(y_1 - y_0)}_{\text{unit-specific } TE} D.$$

More on PO notation

Now let's introduce some notation to make the unit-specificity clear. To each variable, we are going to tack on a "sub-i" to explicitly indicate we are talking about the observed outcome, potential outcomes and treatment for a specific unit or individual:

$$y_i = \begin{cases} y_{1i}, \text{if } D_i = 1 \\ y_{0i}, \text{if } D_i = 0, \text{and} \end{cases}$$

$$y_i = (1 - D_i)y_{0i} + D_i\, y_{1i} = y_{0i} + \underbrace{(y_{1i} - y_{0i})}_{TE_i} D_i.$$

The "1" and "0" to the left of the sub-i will always refer to a state of the world or potential outcome. Unfortunately, there is no consensus on PO notation. You will find others using superscripts (e.g. Y_i^1) and still others putting the state of world in parentheses (e.g. $Y_i(1)$).

Causal inference problem

The *causal inference problem* is that you don't observe counterfactuals:

		Potential Outcomes	
		y_{1i}	y_{0i}
Treated	$D_i = 1$	observed	counterfactual
Untreated	$D_i = 0$	counterfactual	observed

So you can view it as a missing data problem:

i	D_i	y_i	y_{1i}	y_{0i}	$y_{1i} - y_{0i}$
1	0	0		0	?
2	1	1	1		?
3	1	0	0		?
4	0	0		0	?
5	0	1		1	?
6	1	1	1		?

From (y_i, D_i) to the ATE

Because the counterfactual outcomes are never observed, we can never observe TE_i. However, it may be possible to identify the *average treatment effect (ATE)*:

$$ATE = E(y_{1i} - y_{0i}) = E(y_{1i}) - E(y_{0i}).$$

What assumptions do we need on the distribution of (y_i, D_i) to do this?

i	D_i	y_i	y_{1i}	y_{0i}	$y_{1i} - y_{0i}$
1	0	0		0	?
2	1	1	1		?
3	1	0	0		?
4	0	0		0	?
5	0	1		1	?
6	1	1	1		?

Why will this probably not work?

$$\widehat{ATE} = \frac{1}{N_1}\sum_{i=1}^{N_1} y_{1i} - \frac{1}{N_0}\sum_{i=1}^{N_0} y_{0i} = \frac{2}{3} - \frac{1}{3} = \frac{1}{3}$$

Selection bias

Recall we can write the observed outcome as

$$y_i = (1 - D_i)y_{0i} + D_i\, y_{1i} = y_{0i} + (y_{1i} - y_{0i})D_i.$$

Now let's compare average observed outcomes by treatment status:

$$\begin{aligned} E(y_i|D_i = 1) &= E(y_{0i}|D_i = 1) + E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 1) \\ &= E(y_{1i}|D_i = 1) \\ E(y_i|D_i = 0) &= E(y_{0i}|D_i = 0) \end{aligned}$$

Adding and subtracting $E(y_{0i}|D_i = 1)$, we obtain

$$\begin{aligned} E(y_i|D_i = 1) - E(y_i|D_i = 0) = &\underbrace{E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 1)}_{ATT} \\ &+ \underbrace{E(y_{0i}|D_i = 1) - E(y_{0i}|D_i = 0)}_{\text{selection bias}}, \end{aligned}$$

where ATT = average treatment effect on the treated.

Randomization and independence

So the problem is that in general y_{0i} differs systematically by treatment status.

But, what if we *randomize* treatment? Then, treatment status is *independent* of the potential outcomes,

$$(y_{0i}, y_{1i}) \perp\!\!\!\perp D_i,$$

and the assignment mechanism is *ignorable*. In this case, the difference in average observed outcomes can be written as

$$\begin{aligned} E(y_i|D_i = 1) - E(y_i|D_i = 0) &= E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 0) \\ &= E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 1) \end{aligned}$$

because $E(y_{0i}|D_i = 0)$ and $E(y_{0i}|D_i = 1)$ are interchangeable. Randomization also implies

$$\begin{aligned} ATT = E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 1) &= E(y_{1i} - y_{0i}|D_i = 1) \\ &= E(y_{1i} - y_{0i}) \\ &= ATE. \end{aligned}$$

Randomization solves the selection problem.

Conditional independence

Random experiments are often infeasible, unethical or both. But what if there were a set of *covariates* that once you conditioned on them, you could proceed as if treatment assignment was as good as random?

Conditional independence assumption (CIA). Conditional on the covariate vector $\mathbf{x}_i$, the potential outcomes are independent of treatment assignment:

$$\{y_{0i}, y_{1i}\} \perp\!\!\!\perp D_i \,|\, \mathbf{x}_i.$$

Under the CIA,

$$E(y_i|\mathbf{x}_i, D_i = 1) - E(y_i|\mathbf{x}_i, D_i = 0) = E(y_{1i} - y_{0i|}|\mathbf{x}_i) \equiv ATE(\mathbf{x}_i),$$

which is the ATE conditional on $\mathbf{x}_i$.

The CIA is also referred to as *ignorability* or *unconfoundedness*.

Whatever you call it, the bad news is that the CIA is *untestable*. So you need a good story.

Back to the ATE

So how do you get from $ATE(\mathbf{x}_i)$ to the unconditional ATE? By the LIE,

$$\begin{aligned} ATE &= E[E(y_{1i} - y_{0i}|\mathbf{x}_i)] \\ &= \sum_x E(y_{1i} - y_{0i}|\mathbf{x}_i = \mathbf{x})P(\mathbf{x}_i = \mathbf{x}) \\ &= \sum_x [E(y_i|\mathbf{x}_i = \mathbf{x}, D_i = 1) - E(y_i|\mathbf{x}_i = \mathbf{x}, D_i = 0)]P(\mathbf{x}_i = \mathbf{x}) \\ &= \sum_x E(y_i|\mathbf{x}_i = \mathbf{x}, D_i = 1)P(\mathbf{x}_i = \mathbf{x}) \\ &\quad - \sum_x E(y_i|\mathbf{x}_i = \mathbf{x}, D_i = 0)P(\mathbf{x}_i = \mathbf{x}). \end{aligned}$$

Now what do you need to estimate the ATE?

Overlap

To estimate the *ATE*, we need to be able to observe treated and untreated units for every outcome on $\mathbf{x}_i$. Formally, this is called the *overlap* assumption:

Overlap. For all $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is the support of the covariates, $0 < P(D_i = 1|\mathbf{x}_i) < 1$.

If overlap is satisfied, then there is a chance of observing treated and untreated units for any set of covariate values. If there is no overlap for a set of covariate values, then we cannot estimate an *ATE* for a population that includes those values.

As a practical matter, we also desire *covariate balance* in the sense that the covariate distributions are similar between treated and untreated units.

Together, the CIA and overlap are often referred to as *strong ignorability*.

Regression

Let's say we have a good argument for the CIA and a random sample that provides overlap and covariate balance. Then what?

The next step is to estimate the CEF, $E(y_i|\mathbf{x}_i, D_i)$. This is where *regression* comes in.

Suppose

$$E(y_i|\mathbf{x}_i, D_i) = \beta_0 + \delta D_i + \mathbf{x}_i\beta.$$

Then

$$ATE(\mathbf{x}_i) = E(y_i|\mathbf{x}_i, D_i = 1) - E(y_i|\mathbf{x}_i, D_i = 0) = \delta,$$

which is also the *ATE* if δ is a constant. (Should we treat δ as a constant?)

Under the CIA, δ can be consistently estimated by *OLS*.

Tennessee STAR class-size experiment (Krueger 1999)

"Does class size matter?" is the question addressed by the Tennessee Student/Teacher Achievement Ratio (STAR) experiment, the largest randomized experiment of its kind at the time. Project STAR began in the 1985-86 school year randomly assigning kindergarteners and teachers to one of three class configurations, and then followed them for four years. The sample included 11,600 students from 80 schools.

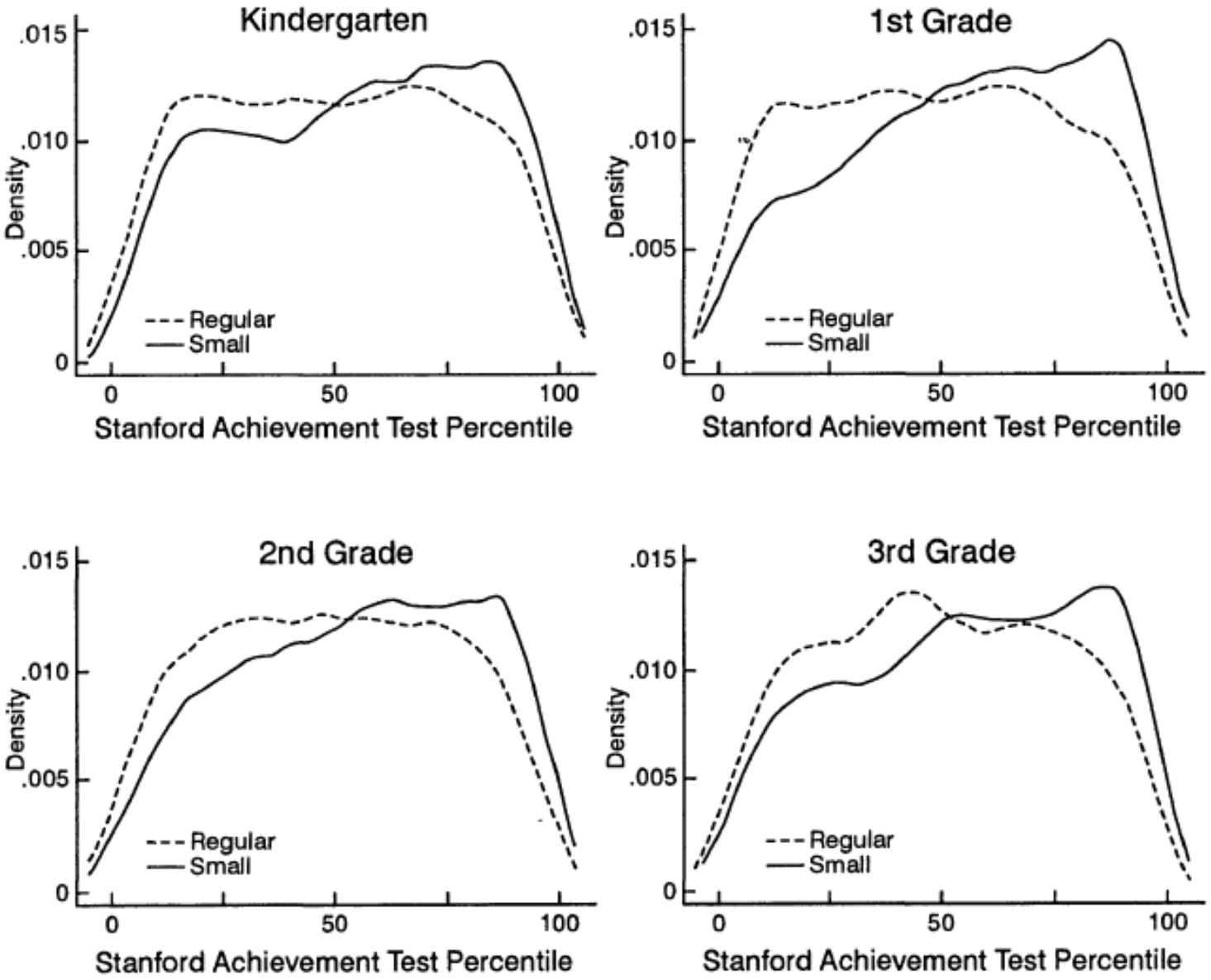
TABLE I

COMPARISON OF MEAN CHARACTERISTICS OF TREATMENTS AND CONTROLS: UNADJUSTED DATA

A. Students who entered STAR in kindergarten[b]

Variable	Small	Regular	Regular/Aide	Joint *P*-Value[a]
1. Free lunch[c]	.47	.48	.50	.09
2. White/Asian	.68	.67	.66	.26
3. Age in 1985	5.44	5.43	5.42	.32
4. Attrition rate[d]	.49	.52	.53	.02
5. Class size in kindergarten	15.1	22.4	22.8	.00
6. Percentile score in kindergarten	54.7	49.9	50.0	.00

Test-score distributions



Experimental evidence

TABLE V

OLS AND REDUCED-FORM ESTIMATES OF EFFECT OF CLASS-SIZE ASSIGNMENT ON AVERAGE PERCENTILE OF STANFORD ACHIEVEMENT TEST

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
A. Kindergarten								
Small class	4.82	5.37	5.36	5.37	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)	(2.19)	(1.25)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44	—	—	8.35	8.44
			(1.35)	(1.36)			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39	—	—	4.48	4.39
			(.63)	(.63)			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07	—	—	−13.15	−13.07
			(.77)	(.77)			(.77)	(.77)
White teacher	—	—	—	−.57	—	—	—	−.57
				(2.10)				(2.10)
Teacher experience	—	—	—	.26	—	—	—	.26
				(.10)				(.10)
Master's degree	—	—	—	−.51	—	—	—	−.51
				(1.06)				(1.06)
School fixed effects	No	Yes	Yes	Yes	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31	.01	.25	.31	.31

ECON 7710

Regression

Chris Cornwell

Terry College of Business

Fall 2021

Section 1

SIMPLE REGRESSION

CEF formulation

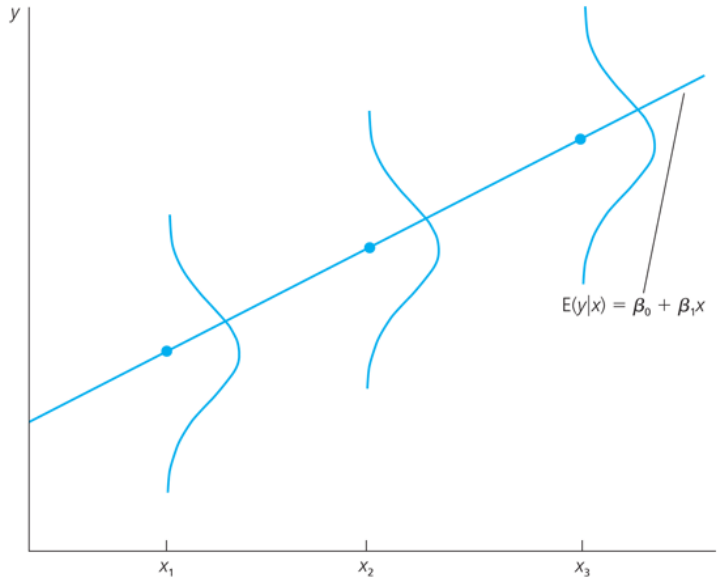
CEF prediction property. For random variables Y and X, the CEF, $E(Y|X) = \mu(X)$, minimizes the MSE of prediction, $E\{[Y - \mu(X)]^2|X\}$.

If the CEF is linear, the *population regression function (PRF)* is the CEF:

$$E(y|x) = \beta_0 + \beta_1 x.$$

Figure 2.1

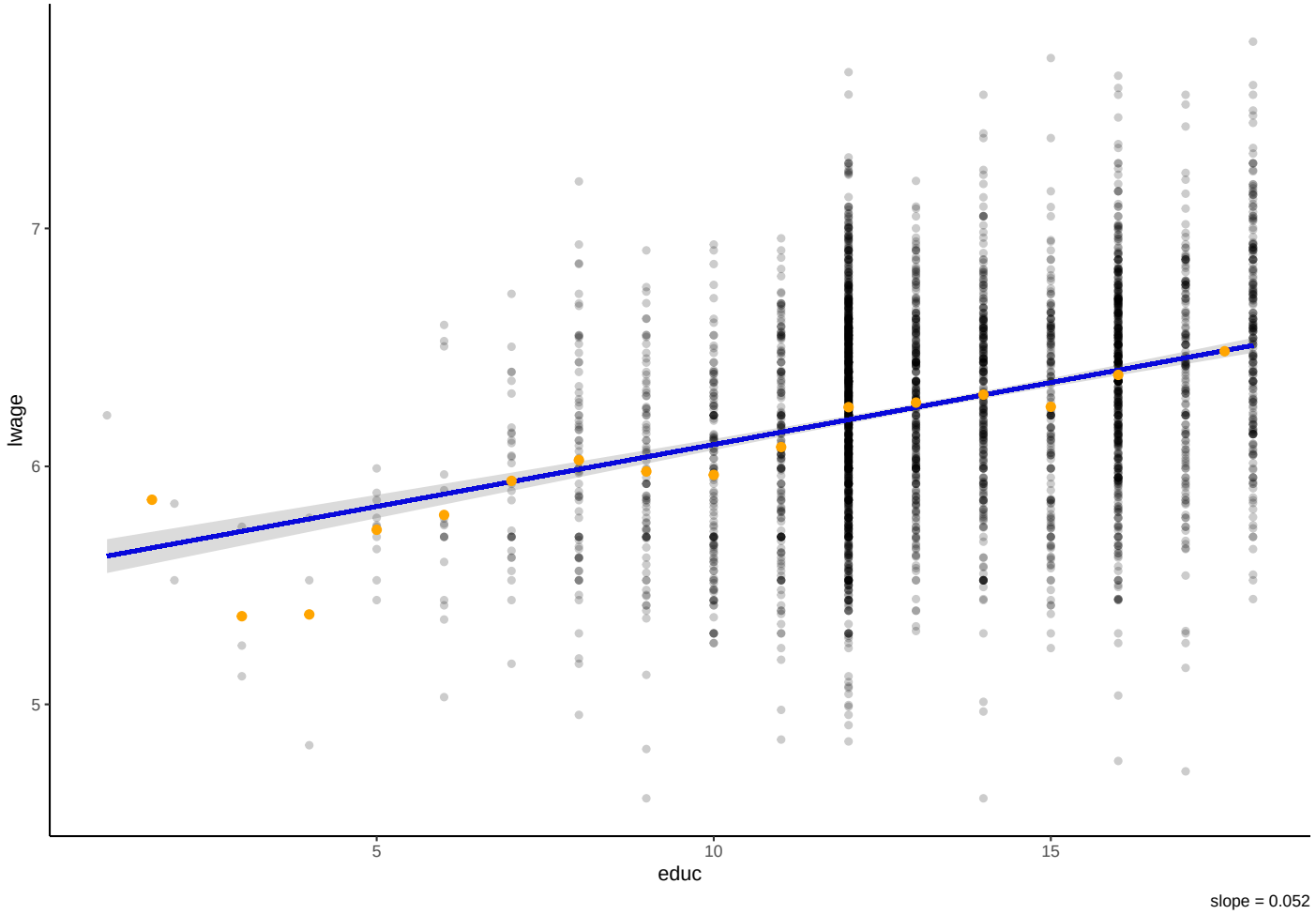
$\mathrm{E}(y|x)$ as a linear function of x.



Source: Wooldridge

Best linear approximation to the CEF

Regression provides the *best linear predictor* of y and the **best linear approximation** to the CEF, in a minimum MSE sense.



Population regression model

Using the linear approximation to the CEF, we have the standard expression of the population regression model in terms of y:

$$y = \beta_0 + \beta_1 x + u,$$

where y is the *dependent variable* or *regressand*, x is an *explanatory variable* or *regressor*, and u is an *error term* capturing unobserved factors.

Taking expectations conditional on x, we get

$$E(y|x) = \beta_0 + \beta_1 x + E(u|x).$$

How do you square the CEF formulation and the standard expression in terms of y?

Interpretation

The mechanical interpretation of β_1 is the change in y associated with a unit change in x, holding the unobserved factors constant:

$$\Delta y = \beta_1 \Delta x, \quad \Delta u = 0.$$

If y is the wage rate and x is years of schooling, then β_1 measures the average change in the wage associated with an additional year of schooling.

If y is the log wage, then β_1 measures the average percentage change in the wage associated with an additional year of schooling, or the rate of return.

Wages or log wages

Table 1: Simple wage and log-wage regressions, Card (1995) sample

	Dependent variable:	
	wage	lwage
	(1)	(2)
educ	29.655***	0.052***
	(1.708)	(0.003)
Constant	183.949***	5.571***
	(23.104)	(0.039)
Observations	3,010	3,010
R^2	0.091	0.099
Note:	*p<0.1; **p<0.05; ***p<0.01	

Conditional mean assumption

A necessary condition for β_1 to have a causal interpretation is that x and u are *mean independent*:

$$E(u|x) = E(u).$$

Whenever there is a constant term in the regression, we can always assert that $E(u) = 0$.

Thus, mean independence is commonly expressed as a conditional mean assumption (CMA):

$$E(u|x) = 0.$$

Implications of the CMA

If the CMA is true,

$$E(y|x) = \beta_0 + \beta_1 x + E(u|x) = \beta_0 + \beta_1 x$$

and β_1 measures the incremental effect of a unit change in x on the CEF.

The CMA is a high bar that is rarely met. In the simple log-wage regression suppose u is ability. The CMA implies

$$E(ability|educ = 12) = E(ability|educ = 16),$$

which is clearly false. To make matters worse, ability is not directly observable and good *proxy* controls are hard to come by.

What can we do in situations like this?

Ordinary least squares

The OLS estimators of β_0 and β_0 minimize:

$$\min_{\tilde{\beta}_0, \tilde{\beta}_1} Q = \sum_i^N (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2,$$

Differentiating Q with respect to $\tilde{\beta}_0$ and $\tilde{\beta}_1$ gives the first-order conditions:

$$\frac{\partial Q}{\partial \tilde{\beta}_0} = \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_i \hat{u}_i = 0$$
$$\frac{\partial Q}{\partial \tilde{\beta}_0} = \sum_i x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_i x_i \hat{u}_i = 0,$$

where $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ is an *OLS residual*. Solving the first-order conditions, we obtain

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$
$$\hat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2}.$$

Method of moments

The CMA implies that x and u are uncorrelated in the population. This assumption, together with the fact that u has zero expected value, imply two population moment conditions:

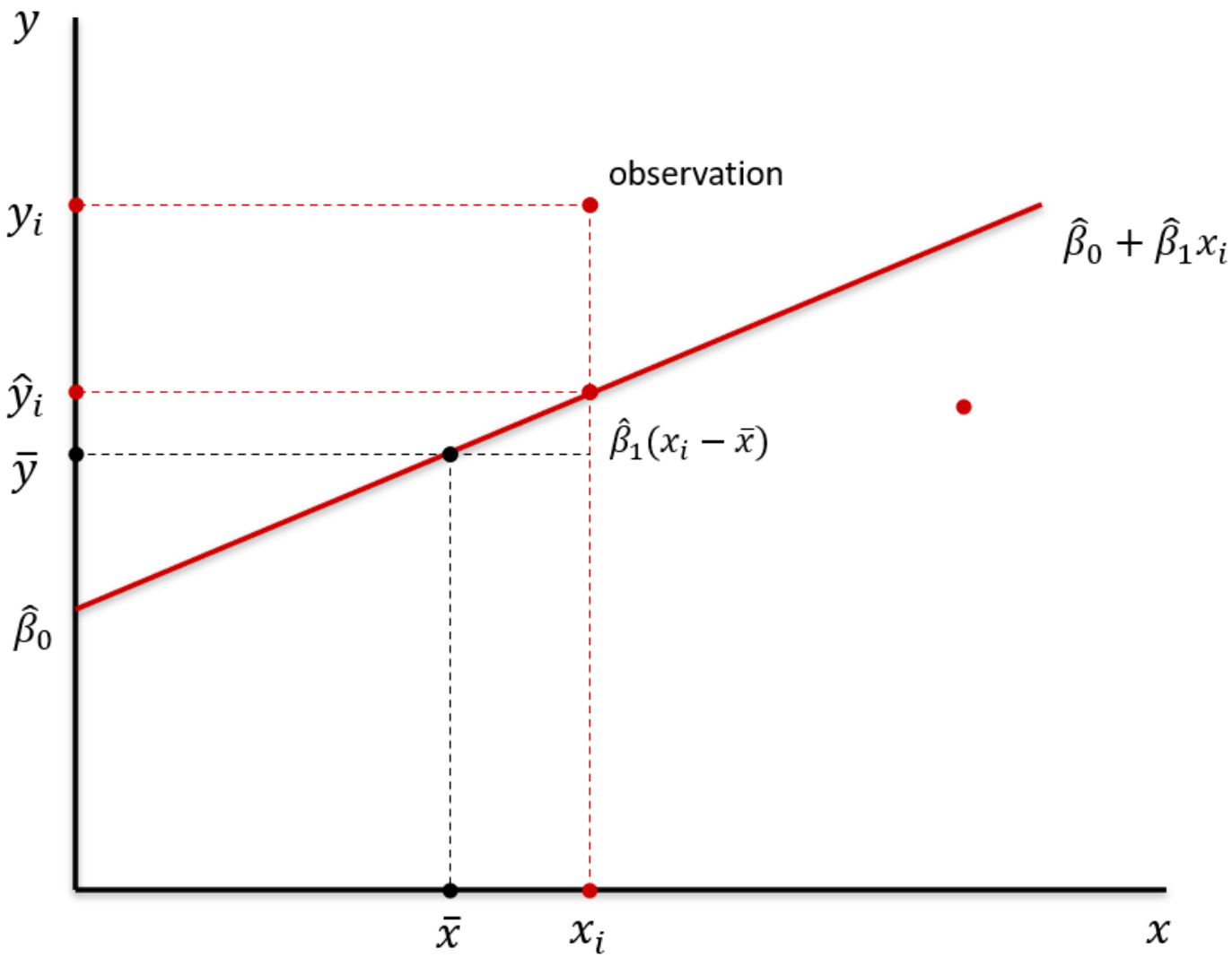
$$E(u) = E(y - \beta_0 - \beta_1 x) = 0$$
$$E(xu) = E[x(y - \beta_0 - \beta_1 x)] = 0.$$

The sample counterparts are just the OLS first-order conditions:

$$\frac{1}{N}\sum_i \hat{u}_i = \frac{1}{N}\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$
$$\frac{1}{N}\sum_i x_i \hat{u}_i = \frac{1}{N}\sum_i x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0.$$

Thus, the OLS and MM estimators of β_0 and β_1 are the same.

Fitted values and residuals



Variance decomposition and regression fit

Because $y_i = \hat{y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i + \hat{u}_i$, we can write

$$y_i - \overline{y} = \hat{\beta}_1 (x_i - \overline{x}) + \hat{u}_i,$$

and the variation in y or *total sum of squares (SST)* as

$$\sum_i (y_i - \overline{y})^2 = \hat{\beta}_1^2 \sum_i (x_i - \overline{x})^2 + \sum_i \hat{u}_i^2 + 2\hat{\beta}_1 \sum_i (x_i - \overline{x}) \hat{u}_i,$$

The cross-product term is zero by the FOC, so

$$SST = \hat{\beta}_1^2 \sum_i (x_i - \overline{x})^2 + \sum_i \hat{u}_i^2 = \underbrace{SSE}_{\text{explained}} + \underbrace{SSR}_{\text{residual}}$$

and we obtain the *goodness-of-fit* measure,

$$R^2 = \frac{SSE}{SST}.$$

Facts about R^2

Here are some facts about R^2:

- R^2 is equal to the squared correlation between y and x.
- As long as the regression contains a constant,

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

- R^2 is usually computed as $1 - \frac{SSR}{SST}$, which has meaning even in nonlinear models.
- If a regression doesn't contain a constant, SSR could be bigger than SST and R^2 would be negative. In this case, $\bar{y}$ is a better predictor of y than a regression through the origin.
- Adding explanatory variables to the model cannot lower R^2.

For causal inference, R^2 will not be of primary importance. Why would we say that?

Units of measurement

Suppose we *scale* each y_i by a constant c_1 and x_i by a constant c_2, run a regression of $c_1 y_i$ on $c_2 x_i$. Then, the OLS estimator of β_1 is

$$\begin{aligned}\tilde{\beta}_1 &= \frac{\sum_i (c_2 x_i - c_2 \overline{x})(c_1 y_i - c_1 \overline{y})}{\sum_i (c_2 x_i - c_2 \overline{x})^2} \\ &= \frac{\sum_i c_1 c_2 (x_i - \overline{x})(y_i - \overline{y})}{\sum_i c_2^2 (x_i - \overline{x})^2} \\ &= (c_1 / c_2)\hat{\beta}_1.\end{aligned}$$

The effect of this scaling on the estimated intercept is:

$$\tilde{\beta}_0 = c_1 \overline{y} - \tilde{\beta}_1 (c_2 \overline{x}) = c_1 (\overline{y} - \hat{\beta}_1 \overline{x}) = c_1 \hat{\beta}_0$$

What would be the effect of the scaling if instead we regressed $\log(c_1 y_i)$ on $\log(c_2 x_i)$?

Boston house price data (1990)

Table 2: Descriptive statistics, 1990 Boston houses sample

Statistic	N	Mean	St. Dev.	Min	Max
price	88	293,546.000	102,713.400	111,000	725,000
assess	88	315.736	95.314	198.700	708.600
bdrms	88	3.568	0.841	2	7
lotsize	88	9,019.864	10,174.150	1,000	92,681
sqrft	88	2,013.693	577.192	1,171	3,880
colonial	88	0.693	0.464	0	1
lprice	88	5.633	0.304	4.710	6.586
lassess	88	5.718	0.262	5.292	6.563
llotsize	88	8.905	0.544	6.908	11.437
lsqrft	88	7.573	0.259	7.066	8.264

Square feet vs hundreds of square feet

Table 3: House price regressions, 1990 Boston sample

	Dependent variable:	
	price	
	(1)	(2)
sqrft	140.211***	
	(11.817)	
sqrft100		14,021.100***
		(1,181.664)
Constant	11,204.150	11,204.150
	(24,742.610)	(24,742.610)
Observations	88	88
R^2	0.621	0.621
Note:	*p<0.1; **p<0.05; ***p<0.01	

The meaning of linear

When we use the term "linear" regression, we mean linear in the parameters.

Each of the following models is linear in this sense:

$$y_i = \beta_0 + \beta_1 x_i^2 + u_i$$
$$\log y_i = \beta_0 + \beta_1 x_i + u_i$$
$$\log y_i = \beta_0 + \beta_1 \log x_i + u_i.$$

This is not:

$$y_i = \beta_0 x_i^{\beta_1} + u_i.$$

Functional forms involving logs

A *log-level* model reflects an exponential relationship between y and x:

$$y_i = \exp(\beta_0 + \beta_1 x_i + u_i) \quad \Rightarrow \quad \log y_i = \beta_0 + \beta_1 x_i + u_i$$

In this case, β_1 is interpreted as a *rate of return*:

$$\beta_1 = \frac{d \log y_i}{dx_i} = \frac{dy_i / y_i}{dx_i}.$$

A *log-log* model is implied by the following specification for y:

$$y_i = \beta_0 \, x_i^{\beta_1} \exp(u_i) \quad \Rightarrow \quad \log y_i = \log \beta_0 + \beta_1 \log x_i + u_i$$

In this case, β_1 is interpreted as an *elasticity*:

$$\beta_1 = \frac{d \log y_i}{d \log x_i} = \frac{dy_i / y_i}{dx_i / x_i} = \frac{dy_i}{dx_i} \frac{x_i}{y_i}.$$

Regression of log house price on log square footage

Table 4: House price regressions, 1990 Boston sample

	Dependent variable:		
	price		lprice
	(1)	(2)	(3)
sqrft	140.211***		
	(11.817)		
sqrft100		14,021.100***	
		(1,181.664)	
lsqrft			0.873***
			(0.085)
Constant	11,204.150	11,204.150	−0.975
	(24,742.610)	(24,742.610)	(0.641)
Observations	88	88	88
R^2	0.621	0.621	0.553
Note:			*p<0.1; **p<0.05; ***p<0.01

Functional form interpretation summary

Table 2.3 **Summary of Functional Forms Involving Logarithms**

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Source: Wooldridge

Rules of thumb for specifying variables in log form

Consider using the log form when the variable is a positive monetary value (wages, salaries, sales, market value, etc) or typically a large integer value (population, employment, enrollment, etc).

Variables measured in years (education, experience, tenure, etc) are generally used in their level form.

Variables that are measured in shares, fractions and percentages (enrollment rate, arrest rates, unemployment rate, etc) can appear in either form, but note that marginal effects in the levels involve *percentage point* changes, but *percentage* changes in the logs.

Log transformations should not be used for variables that can take on values less than or equal to zero.

Estimator = Parameter + Sampling Error

Lemma. $\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \overline{x}) u_i}{\sum_i (x_i - \overline{x})^2}$.

To show this, note that the numerator of $\hat{\beta}_1$ can be written as

$$\begin{aligned}
\sum_i (x_i - \overline{x})(y_i - \overline{y}) &= \sum_i (x_i - \overline{x}) y_i - \overline{y} \sum_i (x_i - \overline{x}) \\
&= \sum_i (x_i - \overline{x}) y_i \\
&= \sum_i (x_i - \overline{x})(\beta_0 + \beta_1 x_i + u_i) \\
&= \beta_0 \sum_i (x_i - \overline{x}) + \beta_1 \sum_i (x_i - \overline{x}) x_i + \sum_i (x_i - \overline{x}) u_i \\
&= \beta_1 \sum_i (x_i - \overline{x})^2 + \sum_i (x_i - \overline{x}) u_i .
\end{aligned}$$

Gauss-Markov assumptions

1. The dependent variable y_i is linearly related to x_i as described in the population regression function.
2. $\{y_i, x_i\}$ represents a random sample of size N on the variables in the population model.
3. The values of the x_i exhibit variation.
4. $E(u \mid x) = 0$. (CMA)
5. $\text{var}(u \mid x) = E(u^2 \mid x) = \sigma^2$. (homoscedasticity)

Under the GM assumptions, OLS is the *best linear unbiased estimator (BLUE)* of the linear regression model.

Usual standard error of $\hat{\beta}_1$

Under the GM assumptions,

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2}$$

This results follows directly from evaluating the expected value of the square of $\hat{\beta}_1$'s sampling error, $E\left\{ \left[\frac{\sum_i (x_i - \overline{x}) u_i}{\sum_i (x_i - \overline{x})^2} \right]^2 \right\}$, conditional on x_i.

To estimate $\sqrt{\text{var}(\beta_1)}$, we use

$$\hat{\sigma}^2 = \frac{1}{N-2} \sum_i \hat{u}_i^2,$$

which gives

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_i (x_i - \overline{x})^2}}.$$

OLS asymptotics

For causal inference, we really only need OLS to have two properties: *consistency* and *asymptotic normality*.

For these properties, we don't need homoscedasticity, which is a dubious assumption anyway.

Dropping homoscedasticity complicates the estimator variance and standard-error expressions and presents a challenge to estimating them. At the same time, relying on the usual OLS standard errors in the presence of heteroscedasticity is a mistake because they are biased, which invalidates the usual test statistics.

The good news is a well-established estimator exists which is *robust* to heteroscedasticity. By "robust", we mean that asymptotically-valid inference is possible whether or not heteroscedasticity exists.

Before we get to it, we should state the consistency and asymptotic normality results.

Consistency

Under assumptions 1-4, the OLS estimators are *consistent*.

This is easy to demonstrate. Consider $\hat{\beta}_1$ and use the lemma to write

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{N}\sum_i (x_i - \overline{x})u_i}{\frac{1}{N}\sum_i (x_i - \overline{x})^2}.$$

Then, by the LLN, $\text{plim}\frac{1}{N}\sum_i (x_i - \overline{x})u_i = \text{cov}(x, u)$ and $\text{plim}\frac{1}{N}\sum_i (x_i - \overline{x})^2 = \text{var}(x)$, so that

$$\text{plim}\,\hat{\beta}_1 = \beta_1 + \frac{\text{cov}(x, u)}{\text{var}(x)} = \beta_1.$$

By similar arguments, $\hat{\sigma}^2 = \frac{\sum_i \hat{u}_i^2}{N-2}$ is a consistent estimator of σ^2.

Asymptotic normality

Under assumptions 1-4, the OLS estimators are asymptotically normally distributed.

By the CLT,

$$\sqrt{N}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N\Big(0, \text{avar}(\hat{\beta}_1)\Big),$$

where $\text{avar}(\hat{\beta}_1)$ is the asymptotic variance of $\hat{\beta}_1$.

Thus, we are able to treat $\hat{\beta}_1$ as asymptotically or approximately normal:

$$\hat{\beta}_1 \overset{A}{\sim} N\left(\beta_1, \frac{1}{N} \cdot \text{avar}(\hat{\beta}_1)\right)$$

Likewise, we can treat the standardized version of $\hat{\beta}_1$, on which the usual t statistic is based, is also asymptotically normal:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \xrightarrow{d} N(0, 1).$$

Heteroscedasticity-robust standard errors

Heteroscedasticity means that the variance of u depends on x:

$$\text{var}(u_i \mid x_i) = \sigma_i^2.$$

Using the lemma and standard asymptotic arguments, we can show

$$\text{avar}(\hat{\beta}_1) = \frac{E[(x_i - \overline{x})u_i]^2}{[\text{var}(x)]^2}.$$

As shown by White (1980), a consistent estimator of $\text{avar}(\hat{\beta}_1)$ is

$$\widehat{\text{avar}}(\hat{\beta}_1) = \frac{\sum_i (x_i - \overline{x})^2 \hat{u}_i^2}{[\sum_i (x_i - \overline{x})^2]^2} \Rightarrow \text{robust se}(\hat{\beta}_1) = \sqrt{\frac{\sum_i (x_i - \overline{x})^2 \hat{u}_i^2}{[\sum_i (x_i - \overline{x})^2]^2}}$$

where $\hat{u}_i$ is based on consistent estimators of β_0 and β_1.

The function `vcovHC` in the `sandwich` package automates the computation of heteroscedasticity-robust standard errors. See the `sandwich` vignette for details.

Asymptotically valid CIs

Given the asymptotic distribution of $\hat{\beta}_1$ and a consistent estimator of avar$(\hat{\beta}_1)$, we can state

$$P\left(-z_{(1-\alpha/2)} < \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} < z_{(1-\alpha/2)}\right) = 1 - \alpha$$

where z is a critical value from the standard normal distribution. Thus, $(1-\alpha)$ CI for β_1 is

$$\hat{\beta}_1 \pm z_{(1-\alpha/2)}\,\text{se}(\hat{\beta}_1).$$

Asymptotically valid test statistics

Because the standardized version of the OLS estimator is asymptotically standard normal, the usual t tests are asymptotically valid:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{se}(\hat{\beta}_1)} \xrightarrow{d} N(0,1).$$

The simplest version of the t test in a regression context is the *test of significance*:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0.$$

The t statistic for this null just the ratio of the estimator to its standard error:

$$t = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}.$$

Economic versus statistical significance

As standard practice would have it, we reject the null at significance level α if the value of the test statistic is greater than z_α and we do the same of the p value is less than α. Just as larger t-statistic values provide greater evidence against the null, so do smaller p values.

While a large test statistic value speaks to statistical significance, economic significance is tied up in the magnitude of $\hat{\beta}_1$. The corresponding t statistic can be large either because the value of $\hat{\beta}_1$ is large or its estimated standard error is small. Thus, it will be wise to take both in consideration when interpreting the results of statistical inference. Depending on sample size, a p value larger than .05 may not lead to accepting the null.

Return to the Card (1995) data

Compute and report robust standard errors

```
library(sandwich)

card_simple <- lm(lwage ~ educ, card)
card_tab2   <- lm(lwage ~ educ+ exper + expersq + black + south
                        + smsa + reg661 + reg662 + reg663
                        + reg664 + reg665 + reg666 + reg667
                        + reg668 + smsa66, card)

vcov_tab2   <- vcovHC(card_tab2, type="HC1")
se_tab2_robust <- sqrt(diag(vcov_tab2))

stargazer(card_simple, card_tab2, card_tab2,
          title="Card Table 2, Column (2) replication",
          dep.var.caption  = "",
          dep.var.labels.include = FALSE,
          omit = c("exper","expersq","black","south","smsa",
                   "reg661","reg662","reg663","reg664",
                   "reg665","reg666","reg667","reg668","smsa66"),
          add.lines = list(c("Controls", "No", "Yes","Yes"),
                           c("Robust se", "No", "No","Yes")),
          se = list(NULL, NULL, se_tab2_robust),
          header = FALSE,
          font.size = "footnotesize",
          omit.stat = c("adj.rsq","ser","f"))
```

Card (1995) Table 2, Column (2)

Table 5: Card Table 2, Column (2) replication

	(1)	(2)	(3)
educ	0.052***	0.075***	0.075***
	(0.003)	(0.003)	(0.004)
Constant	5.571***	4.739***	4.739***
	(0.039)	(0.072)	(0.075)
Controls	No	Yes	Yes
Robust se	No	No	Yes
Observations	3,010	3,010	3,010
R^2	0.099	0.300	0.300
Note:		*p<0.1; **p<0.05; ***p<0.01	

Section 2

MULTIPLE REGRESSION

What changes when we add xs?

Substantively very little.

Everything we've established has a straightforward mapping to the multiple regression model, so what follows is mostly an exercise in translation and extension.

There are some important new concepts to introduce though – most notably the *Frisch-Waugh-Lovell theorem* and *omitted variable bias*.

First, let's restate the linear model and the problem that defines the OLS estimators of its parameters.

New CEF and population regression model

New linear approximation to the CEF with added xs:

$$E(y|x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + u = \beta_0 + \sum_{k=1}^{K} x_k \beta_k + u.$$

Little k indexes the explanatory variables and big K is their number.

New population regression model based on the new CEF:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K + u.$$

Reconciling the two leads to a new CMA:

$$E(u \,|\, x_1, \ldots, x_K) = 0.$$

Is the new CMA more plausible because it includes more xs? This seems like an important *research design* question.

Interpretation

The mechanical interpretation of β_1 is now the *partial effect* of a unit change in x_1 holding $x_2, \ldots, x_K$ and u constant:

$$\Delta y = \beta_1 \Delta x_1, \quad \Delta x_2 = \ldots = \Delta x_K = \Delta u = 0$$

Of course, this interpretation extends to any β_k.

If y is the log wage, then β_1 measures the average rate of return to another year of schooling, controlling for experience, race etc.

The CMA – the claim of *mean independence* between u and the xs – is necessary to characterize the relationship as *causal*.

Ordinary least squares redux

The OLS of the β_k are the values that minimize

$$Q = \sum_i^N (y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_{i1} - \cdots - \tilde{\beta}_K x_{iK})^2.$$

The $K + 1$ first-order conditions for this problem are:

$$\begin{aligned}
\frac{\partial Q}{\partial \hat{\beta}_0} &= \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}) = \sum_i \hat{u}_i = 0 \\
\frac{\partial Q}{\partial \hat{\beta}_1} &= \sum_i x_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}) = \sum_i x_{i1}\hat{u}_i = 0 \\
&\vdots \\
\frac{\partial Q}{\partial \hat{\beta}_K} &= \sum_i x_{iK}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_K x_{iK}) = \sum_i x_{iK}\hat{u}_i = 0.
\end{aligned}$$

Method of moments redux

The CMA implies the following $K+1$ population moment conditions:

$$E(u) = 0 \quad \text{and} \quad E(x_k u) = 0, \quad k = 1, \ldots, K.$$

The sample counterparts are the OLS first-order conditions:

$$\frac{1}{N}\sum_i \hat{u}_i = 0 \quad \text{and} \quad \frac{1}{N}\sum_i x_{ik}\hat{u}_i = 0, \quad k = 1, \ldots, K.$$

Again, the OLS and MM estimators of the β_k are the same.

Frisch-Waugh-Lovell theorem

The *Frisch-Waugh-Lovell (FWL) theorem* clarifies what it means to "hold constant" or "control for" other factors.

It says that the OLS estimator of any β_k can be obtained by a simple regression of y on a "residualized" x_k. You can think of this as a two-step process:

1. Regress x_{ik} on all of the other xs and save the residuals, which we'll call $\hat{r}_{ik}$. The $\hat{r}_{ik}$ are the residualized x_{ik}, which have the effects of the other xs "partialed out".
2. Regress y_i on $\hat{r}_{ik}$. The estimated coefficient of $\hat{r}_{ik}$ will be the OLS estimator of β_k in the multiple regression model.

Turning multiple regression into simple regression

Consider the multiple regression model and focus on the estimation of β_1:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + u_i.$$

Apply OLS to the "artificial" regression of x_1 on the other xs and compute the residuals:

$$x_{i1} = \alpha_0 + \alpha_1 x_{i2} + \cdots + \alpha_{K-1} x_{iK} + r_{i1}$$
$$\hat{r}_{i1} = x_{i1} - \hat{x}_{i1} = x_{i1} - \hat{\alpha}_0 - \hat{\alpha}_1 x_{i2} - \cdots - \hat{\alpha}_K x_{iK}.$$

By construction, the $\hat{r}_{i1}$ capture the part of x_{i1} uncorrelated with the other xs:

$$\sum_i x_{ik} \hat{r}_{i1} = 0, \ k = 2, \ldots, K.$$

The simple regression of y_i on $\hat{r}_{i1}$ produces the OLS estimator of β_1:

$$\hat{\beta}_1 = \frac{\sum_i \hat{r}_{i1} y_i}{\sum_i \hat{r}_{i1}^2}.$$

Residualizing *educ* using the Card (1995) data

How to partial out the effects of the controls on education

```
educ_partial <- lm(educ ~ exper + expersq + black + south
                        + smsa + reg661 + reg662 + reg663
                        + reg664 + reg665 + reg666 + reg667
                        + reg668 + smsa66, card)
r_educ <- resid(educ_partial)
card_tab2_partial <- lm(lwage ~ r_educ, card)
```

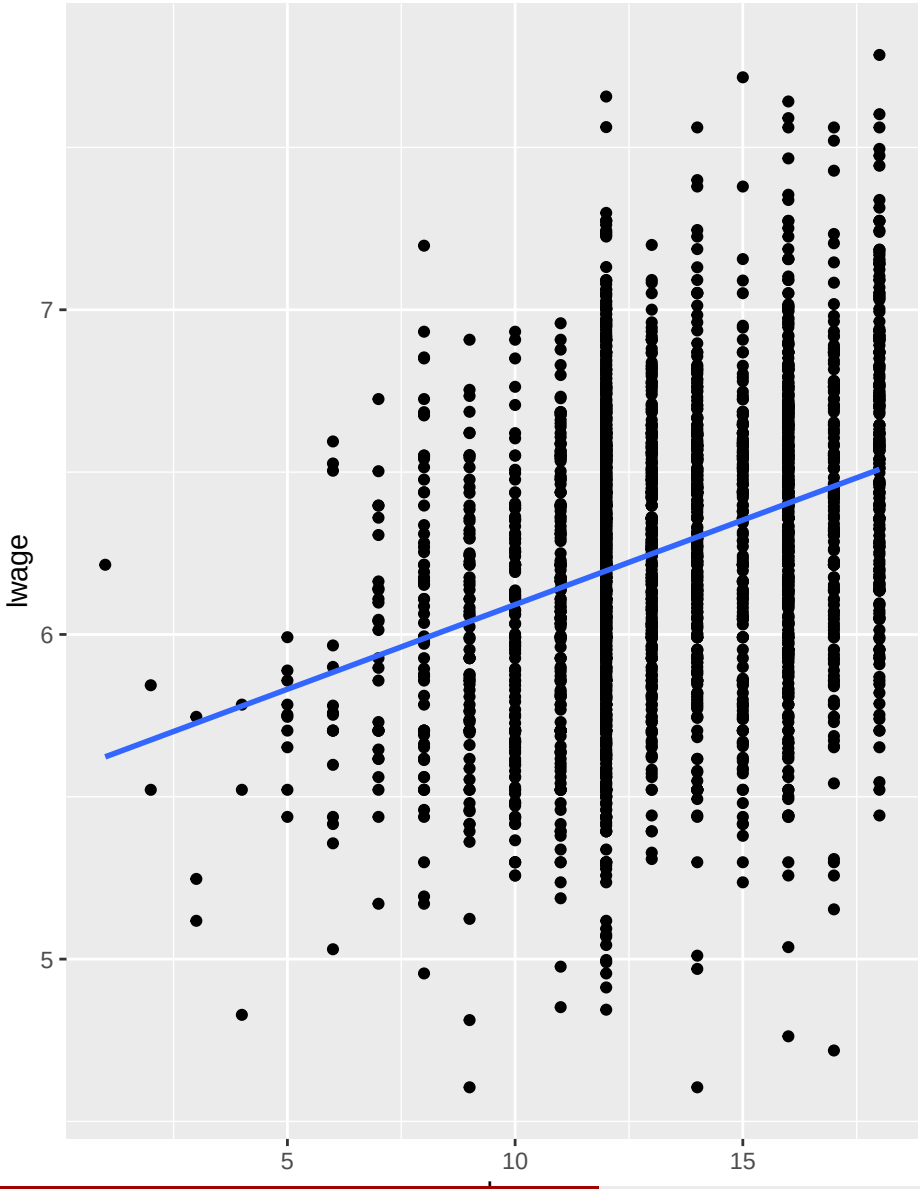
Replicating Table 2, Column (2) two ways

Table 6: Card Table 2, Column (2) replication

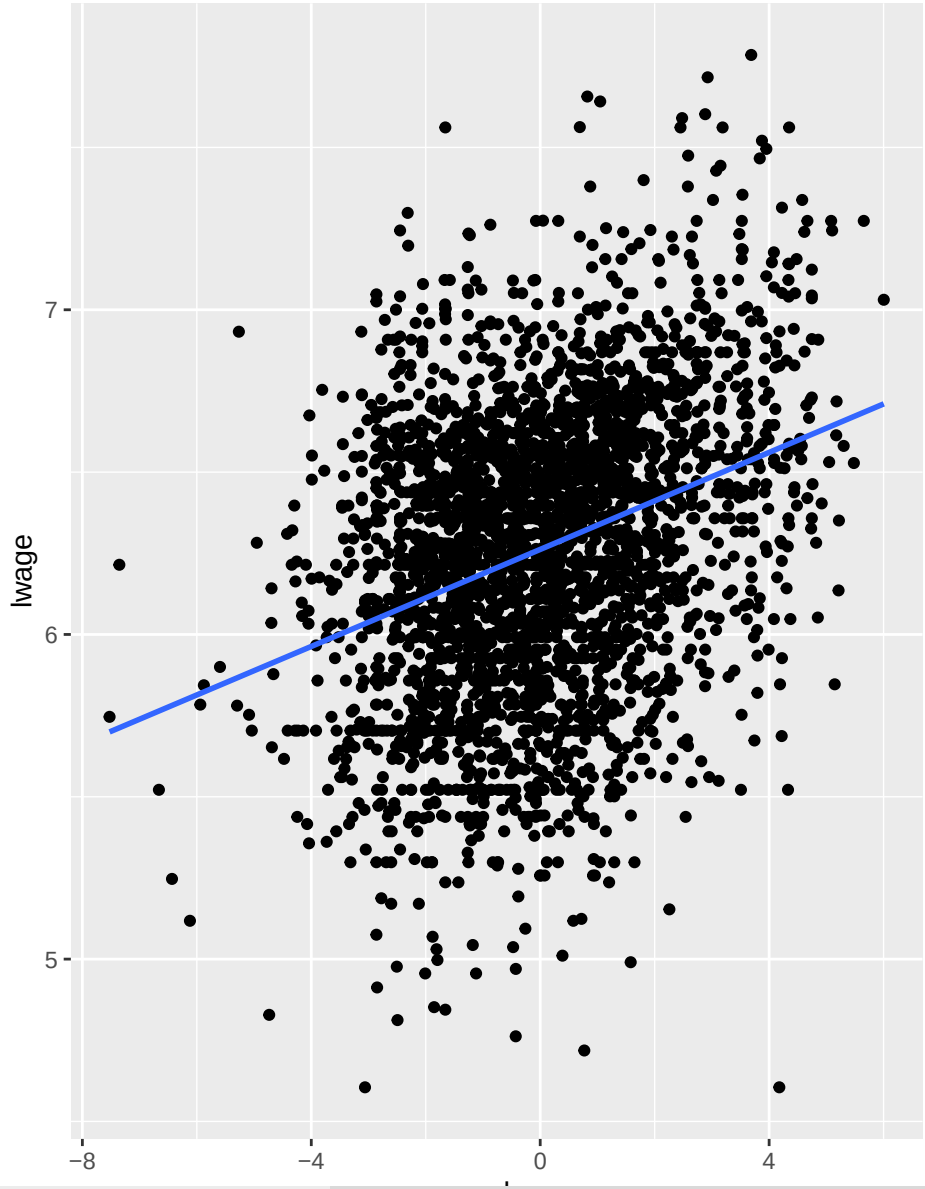
	(1)	(2)
educ	0.075***	
	(0.003)	
r_educ		0.075***
		(0.004)
Constant	4.739***	6.262***
	(0.072)	(0.008)
Controls	Yes	Yes
Observations	3,010	3,010
R^2	0.300	0.107
Note:	*p<0.1; **p<0.05; ***p<0.01	

Two simple regressions visualized

Log wages and education
regression slope =.052



Log wages and residualized education
regression slope =.075



Effect of omitting experience

Table 7: Log wage regressions, without and with experience

	(1)	(2)
educ	0.052***	0.093***
	(0.003)	(0.004)
exper		0.041***
		(0.002)
Constant	5.571***	4.666***
	(0.039)	(0.064)
Observations	3,010	3,010
R^2	0.099	0.181
Note:	*p<0.1; **p<0.05; ***p<0.01	

Omitted variable bias (OVB) formula

Consider a model with $K = 2$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i, \quad \text{with } E(u_i | x_{i1}, x_{i2}) = 0.$$

If x_{i2} is omitted in estimation, OLS yields

$$\begin{aligned} \tilde{\beta}_1 &= \frac{\sum_i (x_{i1} - \overline{x}_1) y_i}{\sum_i (x_{i1} - \overline{x}_1)^2} \\ &= \beta_1 + \beta_2 \frac{\sum_i (x_{i1} - \overline{x}_1) x_{i2}}{\sum_i (x_{i1} - \overline{x}_1)^2} + \frac{\sum_i (x_{i1} - \overline{x}_1) u_i}{\sum_i (x_{i1} - \overline{x}_1)^2}. \end{aligned}$$

Taking expectations, conditional on x_{i1} and x_{i2}, we obtain

$$E(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{\sum_i (x_{i1} - \overline{x}_1) x_{i2}}{\sum_i (x_{i1} - \overline{x}_1)^2} = \beta_1 + \beta_2 \tilde{\delta}.$$

where $\tilde{\delta}$ is the estimated coefficient of x_{i1} in a regression of x_{i2} on x_{i1}.

Signing the bias

Using the OVB formula, $\tilde{\beta}_1$ is biased unless

1. $\beta_2 = 0$ (x_2 does not belong in the model).
2. $\tilde{\delta} = 0$ (x_1 and x_2 are uncorrelated).

It is difficult to sign the bias in models with more explanatory variables because multiple correlations are involved. In general, OVB will affect all of the coefficient estimates, even those not directly correlated with the omitted variable.

The simple OVB formula is nevertheless helpful in sorting out the consequences of unobservables on regression estimates.

Variance decomposition and regression fit (again)

The variance decomposition for the multiple regression model is essentially the same:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_K x_{iK}$$

$$SST = \sum_i (y_i - \bar{y})^2$$

$$SSE = \sum_i (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{u}_i^2$$

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

And everything we said about R^2 in the simple model carries over.

Assumptions 1-4 (again)

Here is the multiple regression version of the 4 basic assumptions required for consistency and asymptotic normality:

1. The dependent variable y_i is linearly related to the x_{ik} as described in the PRF.
2. $\{y_i, x_{i1}, \ldots, x_{iK}\}$ represents a random sample of size N on the variables in the population model.
3. No perfect *multicollinearity* among the x_{ik}.
4. $E(u \mid x_1, \ldots x_K) = 0$. (CMA)

Estimator = Parameter + Sampling Error (again)

Lemma. $\hat{\beta}_k = \frac{\sum_i \hat{r}_{ik} y_i}{\sum_i \hat{r}_{ik}^2} = \beta_k + \frac{\sum_i \hat{r}_{ik} u_i}{\sum_i \hat{r}_{ik}^2}$, where the $\hat{r}_{ik}$ are residuals from the regression of x_{ik} on all of the other xs.

To show this, note that the FOCs for the artificial regression are:

$$\sum_i \hat{r}_{ik} = 0 \quad \text{and} \quad \sum_i x_{ij} \hat{r}_{ik} = 0, \; j \neq k.$$

Then,

$$\sum_i x_{ik} \hat{r}_{ik} = \sum_i (\hat{x}_{ik} + \hat{r}_{ik}) \hat{r}_{ik} = \sum_i \hat{r}_{ik}^2.$$

Asymptotic distribution of the OLS estimators (again)

Under assumptions 1-4, the OLS estimators are consistent and asymptotically normal:

$$\sqrt{N}(\hat{\beta}_k - \beta_k) \xrightarrow{d} N\Big(0, \text{avar}(\hat{\beta}_k)\Big).$$

A robust estimator of $\text{avar}(\hat{\beta}_k)$ is

$$\widehat{\text{avar}}(\hat{\beta}_k) = \frac{\sum_i \hat{r}_{ik}^2 \hat{u}_i^2}{\sum_i \hat{r}_{ik}^2}.$$

Note that the denominator of $\text{var}(\hat{\beta}_k)$ can be written as

$$\sum_i \hat{r}_{ik}^2 = SSR_k = SST_k - SSE_k = SST_k(1 - R_k^2).$$

What would it mean if SST_k is small or R_k^2 is large?

Simple to multiple regression translation done

We now have pretty much everything we need – statistical results wise – to use regression in pursuit of causal inference.

Don't forget though that for regression to deliver on causal claims something like the CMA will have to hold.

Before we return to the PO framing, there are a few other regression issues to address.

Testing single linear restrictions

We use t tests for hypotheses involving *single linear restrictions*, like the usual test of significance, $H_0 : \beta_k = 0$.

Single linear restrictions can also involve multiple parameters, as the test for constant returns to scale, $H_0 : \beta_1 + \beta_2 = 1$ in the production function,

$$logy_i = \beta_0 + \beta_1 logK_i + \beta_2 logL_i + u_i.$$

In this case the test statistic is

$$t = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 1}{\sqrt{\text{var}(\hat{\beta}_1 + \hat{\beta}_2)}}.$$

Testing several linear restrictions

How do you test more than one restriction on the β_k?

Consider a model with $K = 4$,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i,$$

and the hypothesis,

$$H_0 : \beta_3 = \beta_4 = 0.$$

We construct the test statistic from the *unrestricted* (long) and *restricted* (short) regressions. Define the *SSR* from the unrestricted and restricted models:

- $SSR_u = SSR$ from OLS of y on $(const, x_1, \ldots, x_4)$.
- $SSR_r = SSR$ from OLS of y on $(const, x_1, x_2)$.

The F test

Using the SSRs from the long and short regressions, construct the test statistic as

$$F = \frac{(SSR_r - SSR_u)/2}{SSR_u/(N-5)}.$$

Under the assumption of normality, this statistic is distributed $F_{2,N-5}$. Hence the name. Asymptotically, the F statistic converges in distribution to a random variable with χ^2_q distribution divided by its df q.

In general, if H_0 imposes q restrictions,

$$F = \frac{(SSR_r - SSR_u)/q}{SSR_u/(N-K-1)} \sim F_{q,N-K-1}.$$

When $q = 1$, there is a direct relationship between the t and F statistics: $F = t^2$, which implies $t = \sqrt{F}$.

Multicollinearity

Assumption 3 rules out "perfect" multicollinearity, but says nothing about highly correlated regressors, which is what the term multicollinearity is usually invoked to describe.

Recall

$$\widehat{\text{avar}}(\hat{\beta}_k) = \frac{\sum_i \hat{r}_{ik}^2 \hat{u}_i^2}{\sum_i \hat{r}_{ik}^2}$$

where $\sum_i \hat{r}_{ik}^2 = SST_k(1 - R_k^2)$. If R_k^2 is close to 1, $\text{se}(\hat{\beta}_k)$ will be large.

Because multicollinearity (not the perfect version) doesn't violate any assumptions, it does not affect OLS estimator properties. Responses like dropping variables will trade off bias for precision.

More to the point, high correlation between a subset of controls may be irrelevant to the inference you want to conduct on a single causal factor.

Quadratics, interactions and average partial effects

Consider a regression model with linear and quadratic terms:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

The effect of a unit change in x is

$$\frac{dy}{dx} = \beta_1 + 2\beta_2 x.$$

Interactions are evaluated similarly. Consider

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u.$$

The partial effect of x_1 is

$$\frac{dy}{dx} = \beta_1 + \beta_3 x_2$$

In both cases, it is common to report the *average partial effect (APE)*. For example, in the interaction case:

$$APE_{x_1} = \hat{\beta}_1 + \hat{\beta}_3 \overline{x}_2.$$

Evaluating interaction effects

Direct APE calculation

```
pricereg_int1 <- lm(price ~ sqrft + lotsize + sqrft:lotsize, hprice1)
bhat <- coef(pricereg_int1)
lotsizebar <- mean(hprice1$lotsize)
bhat["sqrft"] + bhat["sqrft:lotsize"]*lotsizebar
```

sqrft 104.9717

```
dpricedsqrft_test <- linearHypothesis(pricereg_int1,
                                      c("sqrft + 9020*sqrft:lotsize"))
```

Rescale so that main effect is the APE

```
mean_center <- function(x) {
  scale(x, scale = FALSE)
}
mc_hprice1 <- mean_center(hprice1)
hprice1_mc <- as.data.frame(mc_hprice1)
pricereg_int2 <- lm(price ~ sqrft + lotsize + sqrft:lotsize, hprice1_mc)
```

Estimated APEs

Table 8: Significance test of partial effect of square feet

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
85	462936193563	NA	NA	NA	NA
84	212994500600	1	249941692963	98.5711	0

Table 9: House price regressions with interaction term

	(1)	(2)
sqrft	47.178***	104.972***
	(16.922)	(10.573)
lotsize	−9.829***	3.073***
	(2.013)	(0.562)
sqrft:lotsize	0.006***	0.006***
	(0.001)	(0.001)
Constant	163,982.100***	−6,838.777
	(32,308.640)	(5,481.518)
Observations	88	88
R^2	0.768	0.768
Note:	*p<0.1; **p<0.05; ***p<0.01	

Categorical *x*s

Area	Examples
Individual characteristics	race, gender, education
Firm characteristics	industry, ownership
Program evaluation	policy indicator, program participant
Panel data	"fixed effects"

Interpretating the coefficient of a single dummy

The single dummy variable has the effect of an *intercept shifter*.

To see this, consider the log wage regression

$$lwage = \beta_0 + \beta_1 \, exper + \delta_0 \, educ + u,$$

where $educ = 1$ if an individual has at least a BA, and 0 otherwise. We interpret δ_0 as the difference in conditional means:

$$\delta_0 = E(lwage \,|\, exper, educ = 1) - E(lwage \,|\, exper, educ = 0)$$

Because wages are expressed in logs, δ_0 is the approximate average percentage difference in wages. The exact percentage difference is obtained by exponentiating δ_0:

$$\frac{wage(1) - wage(0)}{wage(0)} = \exp(\delta_0) - 1.$$

Multiple categories

What if there are more than 2 education categories? Let

$$educ = \begin{cases} 2, \text{if BA} \\ 1, \text{if HSG} \\ 0, \text{if less than HSG} \end{cases}$$

How would you interpret δ_0 is this case? What's the problem here?

$$\begin{aligned} E(lwage \,|\, educ = 2) &= \beta_0 + \beta_1 \, exper + 2\delta_0 \\ E(lwage \,|\, educ = 1) &= \beta_0 + \beta_1 \, exper + \delta_0 \\ E(lwage \,|\, educ = 0) &= \beta_0 + \beta_1 \, exper \end{aligned}$$

Always allow for separate category effects (intercepts):

$$wage = \beta_0 + \beta_1 \, exper + \delta_0 \, BA + \delta_1 \, HSG + u.$$

And, don't fall into the *dummy variable trap*.

Interactions between dummies

Suppose we want to allow for gender differences in the return to a college degree. Then, we interact the college degree dummy with a female indicator:

$$lwage = \beta_0 + \beta_1 \, exper + \delta_0 \, BA + \delta_1 \, female + \delta_2 \, BA \cdot female + u.$$

The gender difference in wages is

$$E(lwage|educ = female = 1) - E(wage|educ = 1, female = 0) = \delta_1 + \delta_2$$

What is the gender difference in the return to a college degree?

Interactions between continuous variables and dummies

Let *educ* = years of schooling and consider a model that interacts *educ* and *female*:

$$wage = \beta_0 + \beta_1 \, educ + \delta_0 \, female + \delta_1 \, educ \cdot female + u.$$

Now the *heterogeneity* in the returns to education is reflected in different slopes for women and men:

$$E(wage \mid educ, female = 1) = (\beta_0 + \delta_0) + (\beta_1 + \delta_1) educ$$

What is another way to interpret *fully interacted* or *saturated* model like this?

Biddle and Hamermesh (1990) sleep data

Table 11: Summary statistics, male

Statistic	N	Mean	St. Dev.	Min	Max
sleep	400	3,252.407	435.200	1,485	4,575
totwrk	400	2,434.670	848.317	0	6,415
educ	400	12.867	2.767	5	17
age	400	39.132	11.199	23	65
yngkid	400	0.160	0.367	0	1

Table 12: Summary statistics, female

Statistic	N	Mean	St. Dev.	Min	Max
sleep	306	3,284.588	456.250	755	4,695
totwrk	306	1,715.405	916.245	0	4,065
educ	306	12.667	2.809	1	17
age	306	38.402	11.533	23	64
yngkid	306	0.088	0.284	0	1

Effects of gender on sleep

Table 13: Sleep and gender, BH (1990) data

	(1)	(2)	(3)
totwrk	−0.182***	−0.140***	−0.140***
	(0.024)	(0.028)	(0.026)
educ	−13.052*	−10.205	−10.205
	(7.414)	(9.589)	(9.164)
age	7.157	−30.357	−30.357*
	(14.320)	(18.531)	(17.710)
agesq	−0.045	0.368	0.368*
	(0.168)	(0.223)	(0.213)
yngkid	60.380	−118.283	−118.283
	(59.023)	(93.188)	(89.062)
male			−590.521
			(488.792)
totwrk:male			−0.042
			(0.037)
educ:male			−2.847
			(11.968)
age:male			37.513
			(23.123)
agesq:male			−0.413
			(0.276)
yngkid:male			178.663*
			(108.105)
Constant	3,648.208***	4,238.729***	4,238.729***
	(310.039)	(384.892)	(367.852)
Observations	400	306	706
R^2	0.156	0.098	0.131
Note:			*p<0.1; **p<0.05; ***p<0.01

Testing the null of no gender effect

Table 14: Chow test of gender effects on sleep

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
700	123267451	NA	NA	NA	NA
694	121052555	6	2214896	2.116351	0.0494922

Binary outcomes and linear probability models

Suppose y is binary. Does that change anything? The short answer for what we care about is no.

As long as the CMA holds, the CEF is still

$$E(y|x_1, \ldots, x_K) \equiv E(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K,$$

and regression provides the best linear approximation to the CEF. In addition, it is always the case that the *response probability* is the same as the conditional expectation:

$$P(y = 1|\mathbf{x}) = E(y|\mathbf{x}) \equiv p(\mathbf{x}).$$

Under assumptions 1-4, OLS will consistently estimate the partial effects of the xs on the response probability:

$$\frac{\partial \hat{P}(y = 1|\mathbf{x})}{\partial x_k} = \hat{\beta}_k.$$

Issues with LPMs

However, there are two well-known issues with LPMS:

1. LPMs are heteroscedastic by construction.
   - Because y is Bernoulli, $\text{var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})]$.
   - So always report robust se's.
2. Predicted response probabilities may fall outside the unit interval.
   - You can still report the *percentage correctly predicted*, defining

$$\tilde{y}_i = \begin{cases} 1, \text{if } \hat{y}_i \geq .5 \\ 0, \text{if } \hat{y}_i < .5 \end{cases}$$

   - Or, estimate a true probability model.

If you are primarily interested in prediction, the second issue should probably move you toward a logit or probit model.

Back to potential outcomes

Recall that if we have set of covariates, $\mathbf{x}_i$, such that

$$\{y_{0i}, y_{1i}\} \perp\!\!\!\perp D_i \,|\, \mathbf{x}_i,$$

we can compute the $ATE(\mathbf{x}_i) = E(y_{1i} - y_{0i|}|\mathbf{x}_i)$ as

$$E(y_i|\mathbf{x}_i, D_i = 1) - E(y_i|\mathbf{x}_i, D_i = 0).$$

Assuming the ATE is constant, we can express this difference as the coefficient of the treatment variable in a regression specification of the CEF:

$$E(y_i|\mathbf{x}_i, D_i) = \beta_0 + \delta D_i + \mathbf{x}_i\beta.$$

Selection on observables

Under the CIA, *OLS* applied to

$$y_i = \beta_0 + \delta D_i + \mathbf{x}_i\beta + u_i$$

will consistently estimate δ because

$$E(u_i|\mathbf{x}_i, D_i) = 0.$$

If the CIA holds, we can treat D_i as good as random conditional on the covariates. This is sometimes called *selection on observables*.

Good controls

You might think that the more controls, the better. Not so fast.

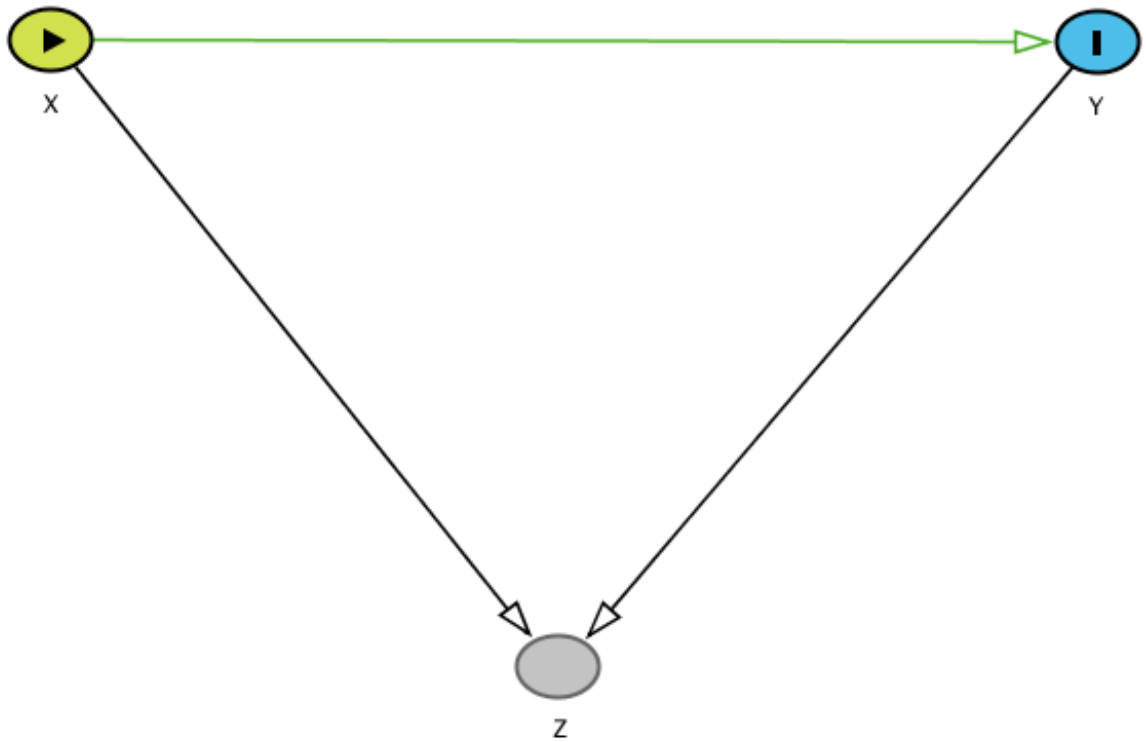
Beware of the *bad control* – a variable whose values are affected by the treatment. Such variables are outcomes in their own right and including them introduces selection bias. In estimating the returns to education, it's why we don't control for occupation.

Bad controls are also referred to as *colliders* and the selection bias including them causes is called *collider bias*.

Variables measured prior to the assignment of treatment status are generally *good controls*. Timing matters.

Where timing is uncertain, you will need to offer clear causal reasoning explaining why the controls are not affected by the treatment.

Colliding



ECON 7710

Panel data

Chris Cornwell

Terry College of Business

Fall 2021

Cross-section and repeated cross-section data

With *cross-section data*, we assume the random variables comprising the regression model are iid draws from some population.

Random sampling over multiple time periods combines *different units* from each time period to create *repeated cross-section* data. In this case, the iid assumption no longer makes sense because population distributions change over time. Thus, we assume the random variables comprising the model are independently, but not identically, distributed (inid).

This is easily accommodated by including time-period dummies in repeated cross-section models. All cross-section estimation methods carry over.

Panel data

In *panel data*, the *same units* are observed over time. We typically assume the units are iid in the cross-section dimension, but such an assumption is generally not plausible in the time-series dimension, because unit-level outcomes are often *serially correlated*.

Time-period dummies are also routinely included in panel-data models. The serial correlation problem is another matter.

In typical panels, the cross-section dimension N is larger than the time-series dimension T, so asymptotic arguments are usually made under the condition that $N \to \infty$ while T remains fixed.

Panels are usually *balanced* in the sense that each unit is observed an equal number of times, but they need not be. What problems do you think might arise when using an unbalanced panel?

NC crime data

Structure of the county-level panel

```
crime.xt <- pdata.frame(crime4, index=c("county","year"))
str(index(crime.xt))
```

```
## Classes 'pindex' and 'data.frame':   630 obs. of  2 variables:
##  $ county: Factor w/ 90 levels "1","3","5","7",..: 1 1 1 1 1 1 1 2 2 2 ...
##  $ year  : Factor w/ 7 levels "81","82","83",..: 1 2 3 4 5 6 7 1 2 3 ...
```

```
str(crime.xt$crmrte)
```

```
##  'pseries' Named num [1:630] 0.0399 0.0383 0.0303 0.0347 0.0366 ...
##  - attr(*, "names")= chr [1:630] "1-81" "1-82" "1-83" "1-84" ...
##  - attr(*, "index")=Classes 'pindex' and 'data.frame':   630 obs. of  2 variables:
##   ..$ county: Factor w/ 90 levels "1","3","5","7",..: 1 1 1 1 1 1 1 2 2 2 ...
##   ..$ year  : Factor w/ 7 levels "81","82","83",..: 1 2 3 4 5 6 7 1 2 3 ...
```

```
pdim(crime.xt)
```

```
## Balanced Panel: n = 90, T = 7, N = 630
```

Key variables

	county	year	crmrte	prbarr	prbconv	prbpris	avgsen	polpc
1-81	1	81	0.03988490	0.289696	0.402062	0.472222	5.61	0.00178678
1-82	1	82	0.03834490	0.338111	0.433005	0.506993	5.59	0.00176659
1-83	1	83	0.03030480	0.330449	0.525703	0.479705	5.80	0.00183577
1-84	1	84	0.03472590	0.362525	0.604706	0.520104	6.89	0.00188588
1-85	1	85	0.03657300	0.325395	0.578723	0.497059	6.55	0.00192436
1-86	1	86	0.03475240	0.326062	0.512324	0.439863	6.90	0.00189522
1-87	1	87	0.03560360	0.298270	0.527596	0.436170	6.71	0.00182786
3-81	3	81	0.01639210	0.202899	0.869048	0.465753	8.45	0.00059392
3-82	3	82	0.01906510	0.162218	0.772152	0.377049	5.71	0.00070467
3-83	3	83	0.01514920	0.181586	1.028170	0.438356	8.69	0.00065866
3-84	3	84	0.01366210	0.194986	0.885714	0.500000	8.01	0.00060890
3-85	3	85	0.01203460	0.206897	0.909091	0.366667	8.59	0.00064134
3-86	3	86	0.01299820	0.156069	1.037040	0.392857	6.03	0.00067621
3-87	3	87	0.01525320	0.132029	1.481480	0.450000	6.35	0.00074588
5-81	5	81	0.00933716	0.406593	0.270270	0.500000	5.53	0.00082085

Panel-data regression model

Extending the cross-section regression model to panel data just requires another subscript:

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_K x_{itK} + a_i + u_{it}, \quad t = 1, \ldots, T,$$

where t denotes time periods and a_i is a fixed (time-invariant) *unobserved effect* that may be correlated with the xs. You can refer to them as *fixed effects* for short.

The primary advantage of panel data is the ability they afford to control for the fixed effects and potentially solve a kind of omitted variables problem.

To see how this works, we should recall the *FWL theorem*. The FWL theorem says that the OLS estimator of any β_k can be obtained by a simple regression of y on a "residualized" x_k. Remember, residualizing means *partialing out* the effects of the other xs.

Partialing out the effects

We can think of the a_i as coefficients of unit-specific dummy variables and regress y on the xs and a dummy for each unit, effectively estimating an intercept for each unit. For typical large-N panels this is impractical, so we partial out the effects first.

"Partialing out" means *time-demeaning* in this case. Here is how it works. Consider a simple panel regression with one explanatory variable,

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}.$$

Now, "collapse" it to the individual level by computing the time average of the model variables:

$$\overline{y}_i = \beta_0 + \beta_1 \overline{x}_i + a_i + \overline{u}_i,$$

where e.g. $\overline{y}_i = \frac{1}{T}\sum_t y_{it}$. Subtracting the time-averaged version from the original model eliminates the a_i:

$$y_{it} - \overline{y}_i = \beta_1 (x_{it} - \overline{x}_i) + u_{it} - \overline{u}_i,$$

or simply

$$\tilde{y}_{it} = \beta_1 \tilde{x}_{it} + \tilde{u}_{it}.$$

Fixed-effects estimation

Estimation proceeds by simply applying OLS to the time-demeaned model, the general version of which is

$$\tilde{y}_{it} = \beta_1 \tilde{x}_{it1} + \cdots + \beta_K \tilde{x}_{itK} + \tilde{u}_{it}.$$

This procedure is commonly called the *fixed-effects (FE) estimator*. It is also sometimes referred to as the *within* estimator because it makes use of within-unit variation over time.

The obvious next question is what are its properties? Or, under what conditions does FE have the properties we want – namely consistency and asymptotic normality?

FE assumptions

Under assumptions 1-4, the FE estimator is consistent and asymptotically normal as $N \to \infty$:

1. The dependent variable y_{it} is linearly related to the x_{itk} as described in the model,

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_K x_{itK} + a_i + u_{it}, \quad t = 1, \ldots, T,$$

2. $\{y_{it}, x_{it1}, \ldots, x_{itK}\}$ represents a random sample from the cross section.
3. No perfect multicollinearity among the $\tilde{x}_{itk}$.
4. $E(u_{it} \mid \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, a_i) = 0$. (CMA)

Note that assumption 4 requires that the u_{it} are uncorrelated with the x_{itk} at every time period. This is called *strict exogeneity*.

Asymptotically-valid inference

Asymptotically valid inference is possible with a consistent estimator of the FE standard errors.

It's complicated because we should allow for both *heteroscedasticity* (the variance of the us depends on the xs) and *serial correlation* (u_{it} and u_{is} are correlated.

The robust approach to standard-error estimation is to *cluster* at the unit level. Consider the case where $T = 2$. The clusters we have in mind are the 2×2 matrices,

$$E(u_i u_i') = \begin{bmatrix} E(u_{i1}^2) & E(u_{i1}u_{i2}) \\ E(u_{i2}u_{i1}) & E(u_{i2}^2) \end{bmatrix}, \quad i = 1, \ldots, N.$$

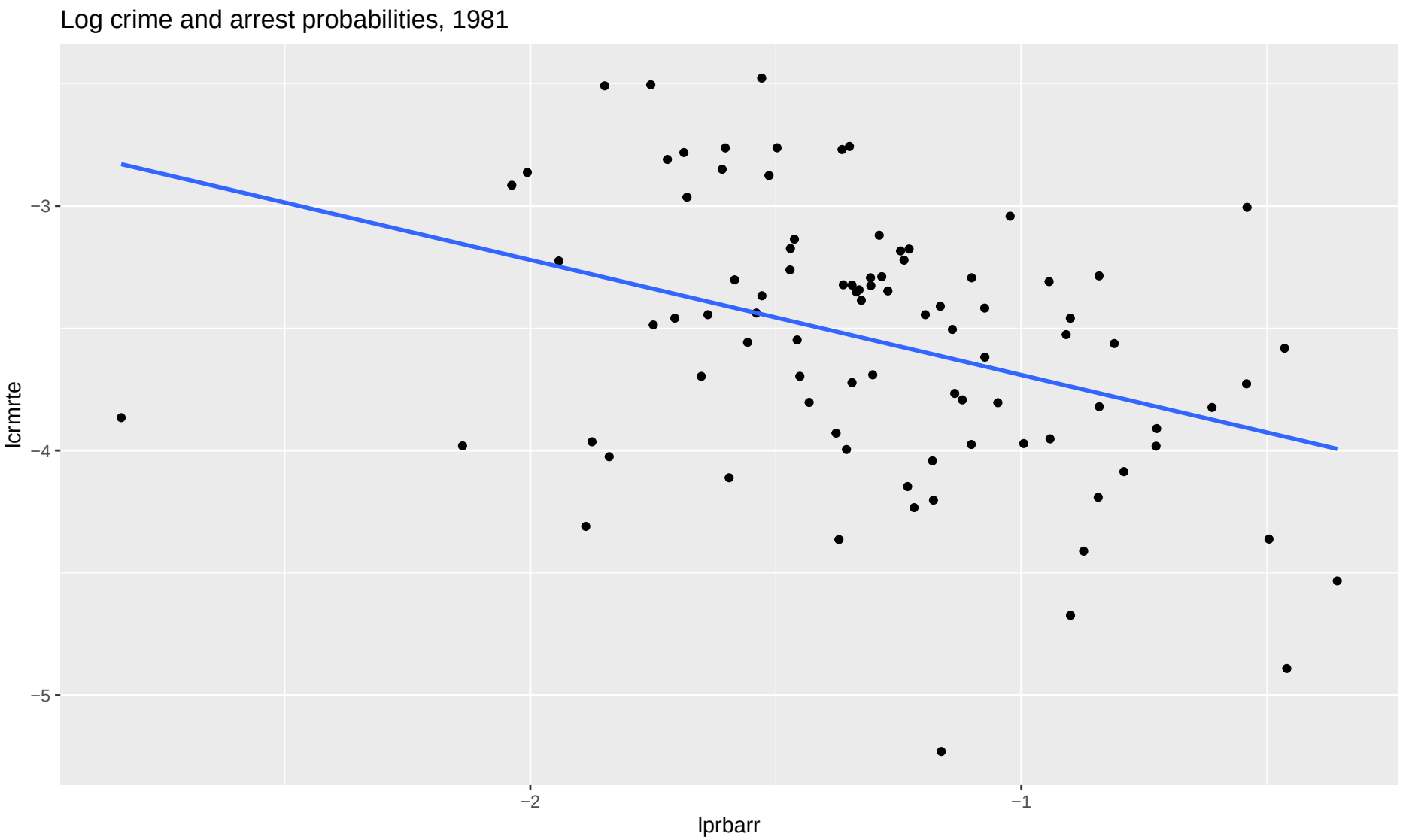
Roughly speaking, "clustering" involves plugging in the sample counterparts of these expected values using FE residuals.

NC crime data summary statistics

Table 1: Summary statistics of main variables, NC crime data

Statistic	N	Mean	St. Dev.	Min	Max
crmrte	630	0.032	0.018	0.002	0.164
prbarr	630	0.307	0.171	0.059	2.750
prbconv	630	0.689	1.690	0.068	37.000
prbpris	630	0.426	0.087	0.149	0.679
avgsen	630	8.955	2.658	4.220	25.830
polpc	630	0.002	0.003	0.0005	0.036

County crime rates and arrest probabilities



OLS and FE estimation of a panel crime model

FE estimation with robust standard errors

```
pols <- plm(lcrmrte ~ lprbarr+lprbconv+lprbpris+lavgsen+lpolpc+year,
            data=crime.xt, model="pooling")
fe   <- plm(lcrmrte ~ lprbarr+lprbconv+lprbpris+lavgsen+lpolpc+year,
            data=crime.xt, model="within")
# COMPUTE CLUSTER-ROBUST STANDARD ERRORS
fe_cov_clustered <- vcovHC(fe, method = "arellano")
fe_se_clustered <- sqrt(diag(fe_cov_clustered))
```

OLS and FE results

Table 2: Crime regressions, NC panel

	POLS	FE	FE
	(1)	(2)	(3)
lprbarr	-0.720^{***}	-0.360^{***}	-0.360^{***}
	(0.037)	(0.032)	(0.059)
lprbconv	-0.546^{***}	-0.286^{***}	-0.286^{***}
	(0.026)	(0.021)	(0.051)
lprbpris	0.248^{***}	-0.183^{***}	-0.183^{***}
	(0.067)	(0.032)	(0.045)
lavgsen	-0.087	-0.004	-0.004
	(0.058)	(0.026)	(0.033)
lpolpc	0.366^{***}	0.424^{***}	0.424^{***}
	(0.030)	(0.026)	(0.084)
Constant	-2.082^{***}		
	(0.252)		
Observations	630	630	630
R^2	0.570	0.434	0.434
Note:		*p<0.1; **p<0.05; ***p<0.01	

ECON 7710

Difference in Differences

Chris Cornwell

Terry College of Business

Fall 2021

Do Castle Doctrine expansions increase violent crime?

What is the *duty to retreat principle* and the *castle doctrine*?

Castle doctrine expansions:

- Remove duty to retreat in places outside the home.
- Remove duty to retreat anywhere one has the right to be.
- Add presumption of reasonable fear.
- Remove civil liability.

Incentives:

- Reduce expected cost of legal lethal force ($\Rightarrow$ increased homicides).
- Increase expected cost of violent crime ($\Rightarrow$ decreased violent crime).

Castle doctrine expansions

Table 1
States that Expanded Castle Doctrine Between 2000 and 2010

State	Effective Date	Removes duty to retreat somewhere outside home	Removes duty to retreat in any place one has a legal right to be	Presumption of reasonable fear	Removes civil liability
Alabama	6/1/06	Yes	Yes	No	Yes
Alaska	9/13/06	Yes	No	Yes	Yes
Arizona	4/24/06	Yes	Yes	Yes	Yes
Florida	10/1/05	Yes	Yes	Yes	Yes
Georgia	7/1/06	Yes	Yes	No	Yes
Indiana	7/1/06	Yes	Yes	No	Yes
Kansas	5/25/06	Yes	Yes	No	Yes
Kentucky	7/12/06	Yes	Yes	Yes	Yes
Louisiana	8/15/06	Yes	Yes	Yes	Yes
Michigan	10/1/06	Yes	Yes	No	Yes
Mississippi	7/1/06	Yes	Yes	Yes	Yes
Missouri	8/28/07	Yes	No	No	Yes
Montana	4/27/09	Yes	Yes	Yes	No
North Dakota	8/1/07	Yes	No	Yes	Yes
Ohio	9/9/08	Yes	No	Yes	Yes
Oklahoma	11/1/06	Yes	Yes	Yes	Yes
South Carolina	6/9/06	Yes	Yes	Yes	Yes
South Dakota	7/1/06	Yes	Yes	No	No
Tennessee	5/22/07	Yes	Yes	Yes	Yes
Texas	9/1/07	Yes	Yes	Yes	Yes
West Virginia	2/28/08	Yes	Yes	No	No

TWFE empirical model (panel regression)

Cheng and Hoekstra estimate a TWFE model of the form:

$$\log outcome_{it} = \beta_0 CDL_{it} + \beta_1 X_{it} + c_i + u_t + \varepsilon_{it}.$$

Variable definitions:

- Outcomes are crimes of a particular type per 100,000 people for state i in year t.
- CDL is a castle-doctrine-law indicator (weighted by the proportion of the year it is effective).
- c_i and u_t are state and year fixed effects.

The model is estimated by OLS and standard errors are clustered at the state level.

Data

Crime data provided by the FBI UCR for the 2000-2010 period

- Burglary, robbery, aggravated assault (deterrence)
- Murder + non-negligent manslaughter (homicide)
- Larceny, auto theft (falsification)

Controls

- Police (UCR), incarceration rate (BJS)
- Median family income, poverty rate (ACS)
- Racial demographics (ACS)
- Public assistance spending (Census)

Florida log homicide DD plot

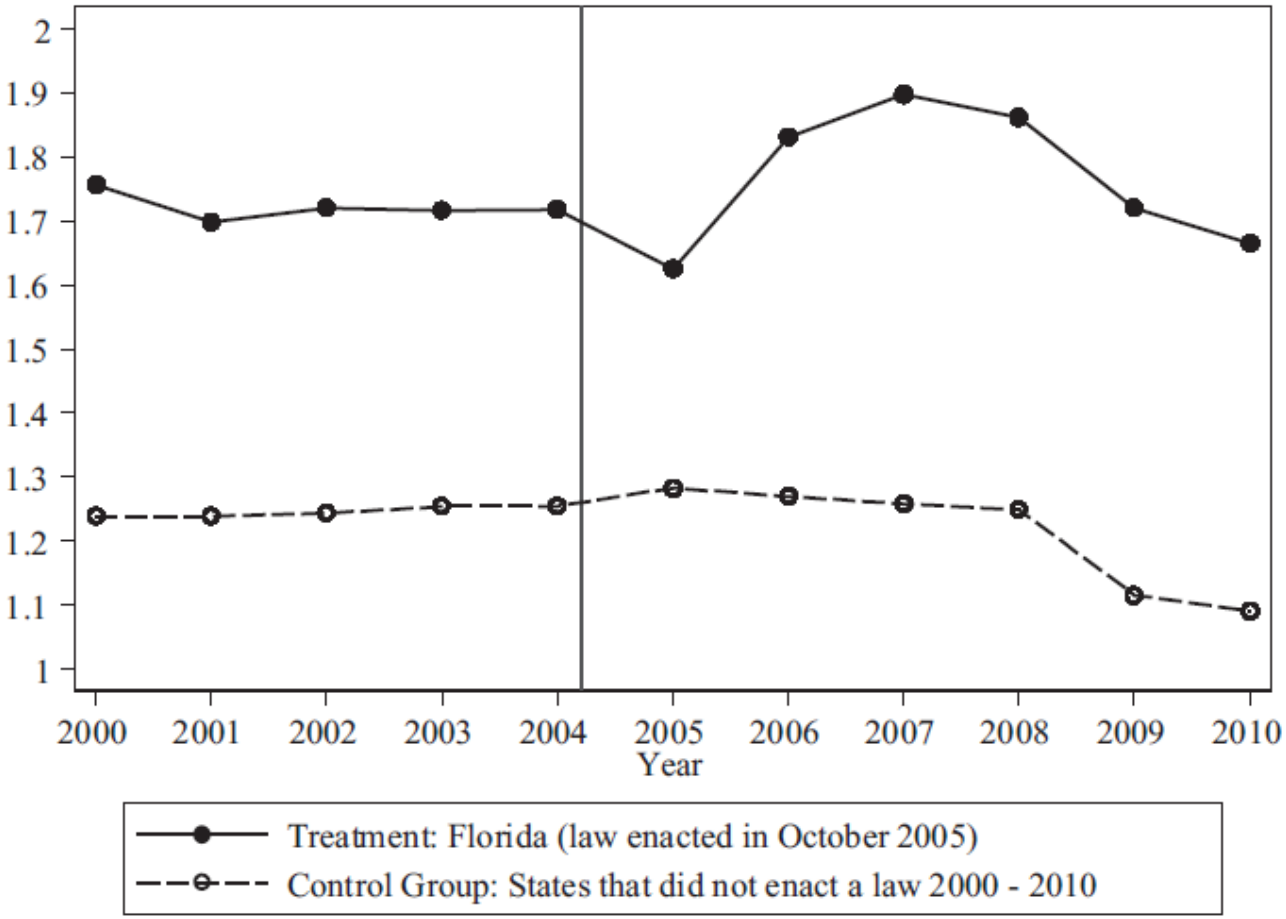


Figure 1a: State Adopting in 2005 (Florida)

Falsification/placebo exercises

Table 3
Falsification Tests: The Effect of Castle Doctrine Laws on Larceny and Motor Vehicle Theft

	OLS—Weighted by State Population					
	1	2	3	4	5	6
Panel A: Larceny	Log (Larceny Rate)					
Castle Doctrine Law	0.00300	–0.00660	–0.00910	–0.00858	–0.00401	–0.00284
	(0.0161)	(0.0147)	(0.0139)	(0.0165)	(0.0128)	(0.0180)
0 to 2 years before adoption of castle doctrine law				0.00112		
				(0.0105)		
Observation	550	550	550	550	550	550
Panel B: Motor Vehicle Theft	Log (Motor Vehicle Theft Rate)					
Castle Doctrine Law	0.0517	–0.0389	–0.0252	–0.0294	–0.0165	–0.00708
	(0.0563)	(0.0448)	(0.0396)	(0.0469)	(0.0354)	(0.0372)
0 to 2 years before adoption of castle doctrine law				–0.00896		
				(0.0216)		
Observation	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Controls for Larceny or Motor Theft					Yes	
State-Specific Linear Time Trends						Yes

Overview of difference in differences

A DD design is appealing when the CIA (CMA) is not tenable and you have "pre-treatment" data.

The canonical design with one treated and one control group compares the difference in mean outcomes for the treated observations before and after treatment with the difference in mean outcomes for the controls before and after treatment.

The key identifying assumption is that the treatment and control-group outcomes would follow *parallel trends* in absence of the treatment. If it holds, DD will identify an ATT.

When there are multiple treatment groups and staggered treatment periods, the story is more complicated.

Canonical 2×2 DD

Suppose there are two groups, treated and control, and two periods, before and after.

Wrong counterfactuals

- Before and after comparisons: miss time trends in the outcome.
- Treated and untreated comparisons: miss potential selection into treatment.

The canonical DD design compares before and after outcomes for the treated group with those for the control group:

Canonical DD

	Before	After
Treated	$\overline{y}_{treat,before}$	$\overline{y}_{treat,after}$
Control	$\overline{y}_{control,before}$	$\overline{y}_{control,after}$

$$DD = \overline{y}_{treat,after} - \overline{y}_{treat,before} - (\overline{y}_{control,after} - \overline{y}_{control,before})$$

DD estimand

A DD research design focuses on the *average treatment effect on the treated (ATT)*:

$$ATT = E(y_1 - y_0|D = 1, t = 1),$$

where $t = 1$ (0) indicates after (before).

The cell means of the 2×2 case estimate:

		Estimands	
		Before ($t = 0$)	After ($t = 1$)
Treated	$D = 1$	$E(y_0\|D = 1, t = 0)$	$E(y_1\|D = 1, t = 1)$
Untreated	$D = 0$	$E(y_0\|D = 0, t = 0)$	$E(y_0\|D = 0, t = 1)$

The problem is that *we do not observe* $E(y_0|D = 1, t = 1)$.

Parallel trends assumption

Wrong counterfactuals revisited:

- Before and after comparison assumes

$$E(y_0|D=1,t=1) = E(y_0|D=1,t=0).$$

- Treated and untreated comparisons assumes

$$E(y_0|D=1,t=1) = E(y_0|D=0,t=1).$$

Assume treatment and control groups follow a *parallel trend*:

		Estimands	
		Before ($t=0$)	After ($t=1$)
Treated	$D=1$	$\mu+\gamma$	$\mu+\gamma+\eta+\delta$
Untreated	$D=0$	μ	$\mu+\eta$

Implication of parallel trends

Under the parallel trends assumption:

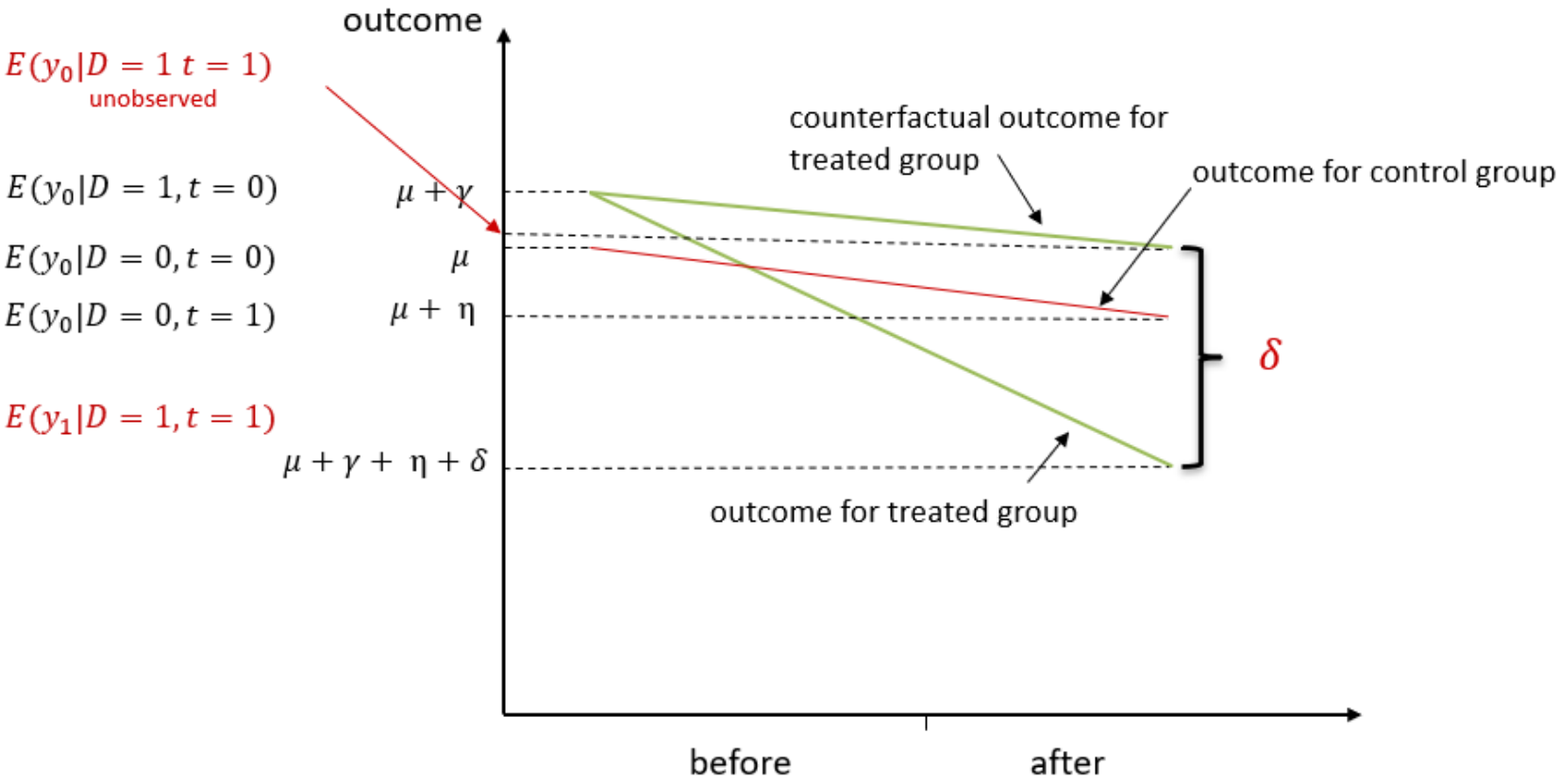
$$E(y_0|D=1,t=1) - E(y_0|D=1,t=0) = \\ E(y_0|D=0,t=1) - E(y_0|D=0,t=0)$$

so that the DD in cell means estimates

$$(\eta + \delta) - \eta = \delta = ATT.$$

The parallel trends assumption not only requires the outcomes of the treatment and control group to follow a common trend, but also γ (which you can interpret as a *selection bias* term) does not change over time.

Parallel trends illustrated



Simple DD regression

Under the parallel trends assumption,

		Estimands	
		Before ($t = 0$)	After ($t = 1$)
Treated	$D = 1$	$\mu + \gamma$	$\mu + \gamma + \eta + \delta$
Untreated	$D = 0$	μ	$\mu + \eta$

we can estimate the ATT by applying OLS to

$$y = \mu + \gamma\, treat + \eta\, after + \delta\, treat \cdot after + u.$$

The regression formulation is appealing because it readily

- accommodates covariates (why include them?)
- generalizes for multiple time periods and treatment groups
- facilitates standard error estimation

Worker's compensation and injury duration (Meyer 1995)

What is the effect of increased *worker's compensation* and the *time out of work*? Meyer addresses this question using "natural experiments" in KY and MI, whereby the max benefit was raised by 65-70 percent in the early 1980s.

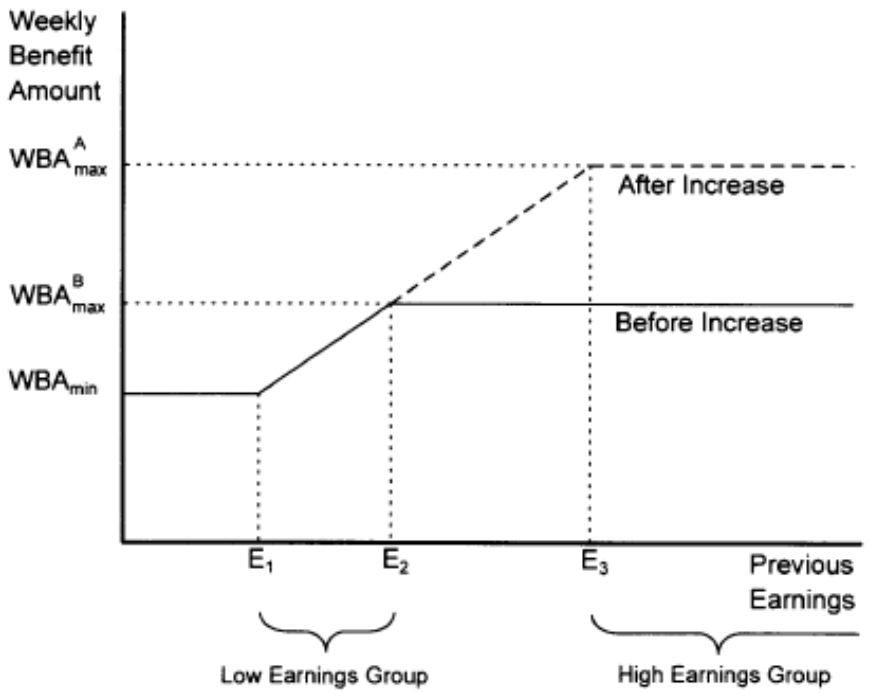


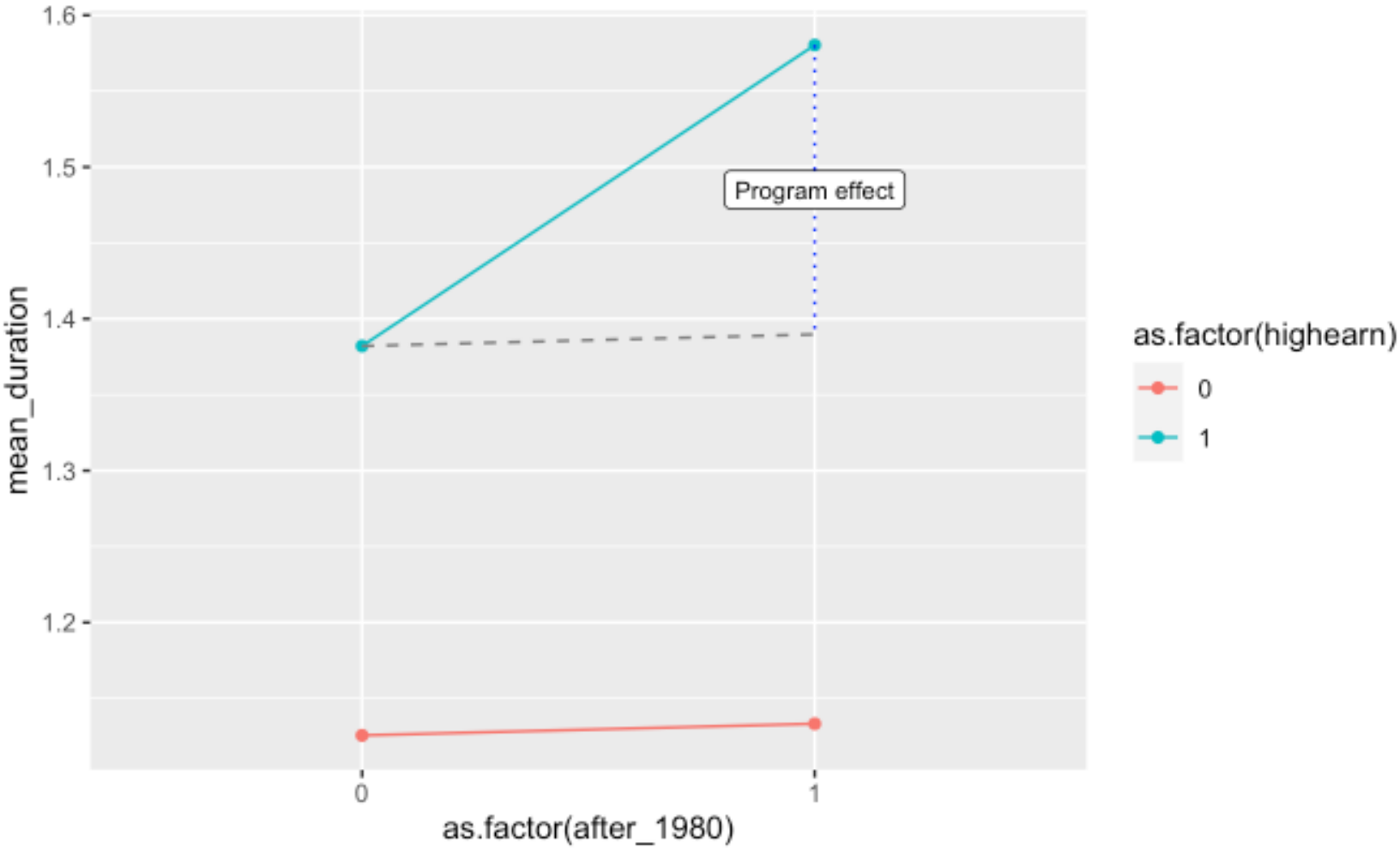
FIGURE 1. TEMPORARY TOTAL BENEFIT SCHEDULE BEFORE AND AFTER AN INCREASE IN THE MAXIMUM WEEKLY BENEFIT

Meyer (1995) basic finding

TABLE 4—KENTUCKY AND MICHIGAN: DURATION AND MEDICAL COSTS OF TEMPORARY TOTAL DISABILITIES DURING THE YEARS BEFORE AND AFTER BENEFIT INCREASES

	High earnings		Low earnings		Differences		Difference in differences
Variable	Before increase (1)	After increase (2)	Before increase (3)	After increase (4)	[(2)−(1)] (5)	[(4)−(3)] (6)	[(5)−(6)] (7)
Mean duration (weeks)							
Kentucky	11.16	12.89	6.25	7.01	1.72	0.76	0.96
	(0.83)	(0.83)	(0.30)	(0.41)	(1.17)	(0.51)	(1.28)
Michigan	14.76	19.42	10.94	13.64	4.66	2.70	1.96
	(2.25)	(2.67)	(1.09)	(1.56)	(3.49)	(1.90)	(3.97)
Median duration (weeks)							
Kentucky	4.00	5.00	3.00	3.00	1.00	0.00	1.00
	(0.14)	(0.20)	(0.11)	(0.12)	(0.25)	(0.16)	(0.29)
Michigan	5.00	7.00	4.00	4.00	2.00	0.00	2.00
	(0.45)	(0.67)	(0.22)	(0.28)	(0.81)	(0.35)	(0.89)
75th percentile, duration (weeks)							
Kentucky	8.00	10.00	7.00	7.00	2.00	0.00	2.00
	(0.28)	(0.45)	(0.21)	(0.24)	(0.53)	(0.32)	(0.62)
Michigan	10.00	14.00	8.50	9.00	4.00	0.50	3.50
	(0.74)	(1.88)	(0.54)	(0.57)	(2.03)	(0.79)	(2.17)
Mean of log duration							
Kentucky	1.38	1.58	1.13	1.13	0.20	0.01	0.19
	(0.04)	(0.04)	(0.03)	(0.03)	(0.05)	(0.04)	(0.07)
Michigan	1.58	1.87	1.41	1.51	0.29	0.10	0.19
	(0.09)	(0.10)	(0.06)	(0.06)	(0.13)	(0.08)	(0.16)

Identification in Meyer (1995)



Examining the parallel trends assumption

Strategies for examining the reasonableness of the parallel trends assumption:

- DD plots
- Falsification or placebo tests
  - Alternative control groups
  - Alternative outcomes
  - Testing leads
- Including unit-specific trends

All of these strategies require more than two periods.

Cheng and Hoekstra (2013) employ most of these.

Generalizing DD regression to multiple periods and treatments

Consider a model of outcome y for individual i in group g at time t (ignoring covariates):

$$y_{igt} = \gamma_g + \eta_t + \delta D_{gt} + u_{igt},$$

where D_{gt} is a treatment indicator for group g in period t, γ_g is a group effect and η_t is a period effect.

More commonly, analysis is carried out on group-level aggregates:

$$y_{gt} = \gamma_g + \eta_t + \delta D_{gt} + u_{gt}, \quad g = 1, \dots G; \; t = 1, \dots T.$$

The group-level model is a *two-way fixed-effects (TWFE)* panel regression. Parallel trends implies

$$E(y_{0,gt}|g, t) = \gamma_g + \eta_t.$$

Back to Cheng and Hoekstra (2013)

TWFE model:

$$\log outcome_{it} = \beta_0 CDL_{it} + \beta_1 X_{it} + c_i + u_t + \varepsilon_{it},$$

where

- Outcomes are crimes of a particular type per 100,000 people for state i in year t.
- CDL is a castle-doctrine-law indicator (weighted by the proportion of the year it is effective).
- c_i and u_t are state and year fixed effects.

The model is estimated by OLS and standard errors are clustered at the state level.

Falsification/placebo exercises again

Table 3
Falsification Tests: The Effect of Castle Doctrine Laws on Larceny and Motor Vehicle Theft

	OLS—Weighted by State Population					
	1	2	3	4	5	6
Panel A: Larceny	Log (Larceny Rate)					
Castle Doctrine Law	0.00300	–0.00660	–0.00910	–0.00858	–0.00401	–0.00284
	(0.0161)	(0.0147)	(0.0139)	(0.0165)	(0.0128)	(0.0180)
0 to 2 years before adoption of castle doctrine law				0.00112		
				(0.0105)		
Observation	550	550	550	550	550	550
Panel B: Motor Vehicle Theft	Log (Motor Vehicle Theft Rate)					
Castle Doctrine Law	0.0517	–0.0389	–0.0252	–0.0294	–0.0165	–0.00708
	(0.0563)	(0.0448)	(0.0396)	(0.0469)	(0.0354)	(0.0372)
0 to 2 years before adoption of castle doctrine law				–0.00896		
				(0.0216)		
Observation	550	550	550	550	550	550
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes
Time-Varying Controls			Yes	Yes	Yes	Yes
Controls for Larceny or Motor Theft					Yes	
State-Specific Linear Time Trends						Yes

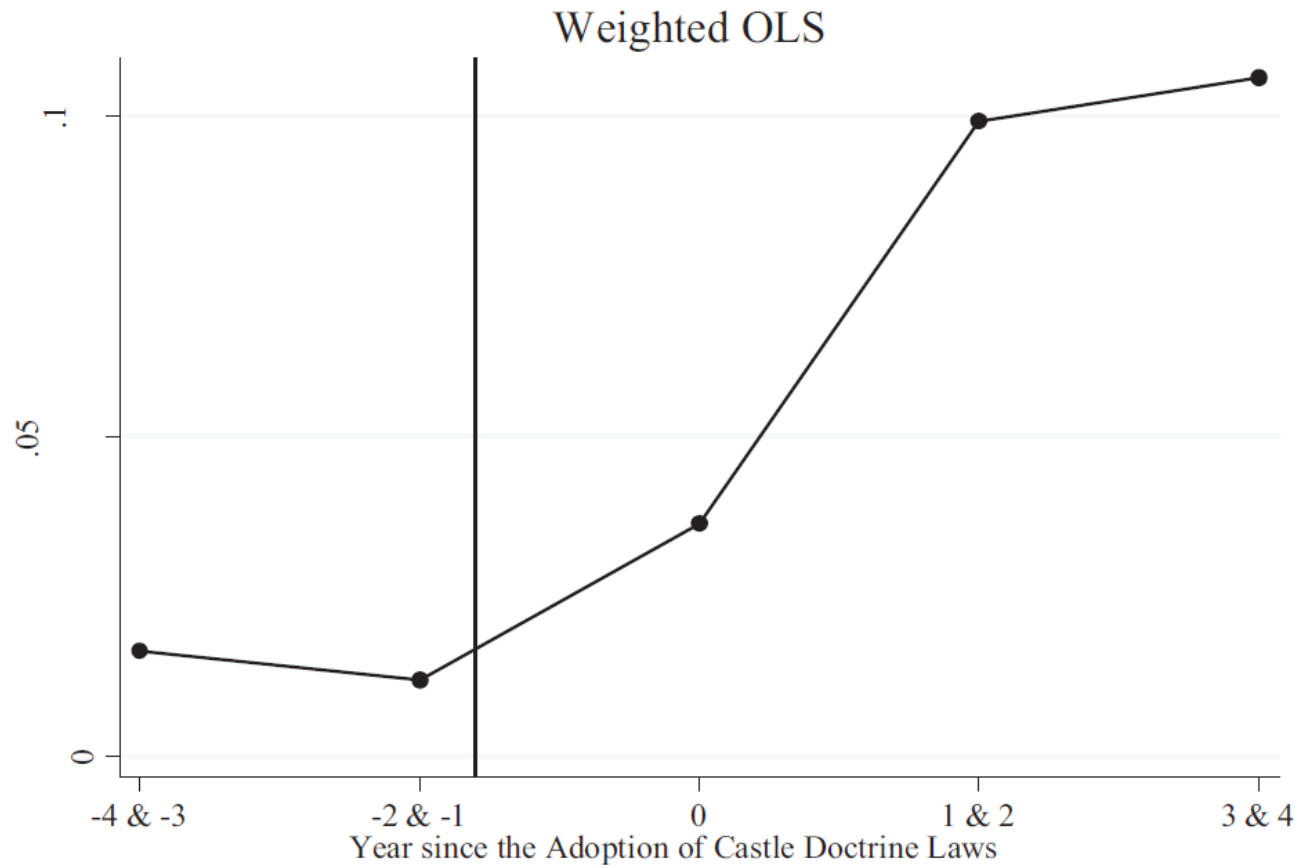
Deterrence (burglary and robbery) results

Table 4

The Deterrence Effects of Castle Doctrine Laws: Burglary, Robbery, and Aggravated Assau

	OLS—Weighted by State Population					
	1	2	3	4	5	6
Panel A: Burglary	Log (Burglary Rate)					
Castle Doctrine Law	0.0780***	0.0290	0.0223	0.0181	0.0327*	0.0237
	(0.0255)	(0.0236)	(0.0223)	(0.0265)	(0.0165)	(0.0207)
0 to 2 years before adoption of castle doctrine law				–0.00906		
				(0.0133)		
Panel B: Robbery	Log (Robbery Rate)					
Castle Doctrine Law	0.0408	0.0344	0.0262	0.0197	0.0376**	0.0515*
	(0.0254)	(0.0224)	(0.0229)	(0.0257)	(0.0181)	(0.0274)
0 to 2 years before adoption of castle doctrine law				–0.0138		
				(0.0153)		

Divergence in log homicide rates after law change



Homicide results

Table 5
The Effect of Castle Doctrine Laws on Homicide

Panel A: Log Homicide Rate (OLS—Weighted)	1	2	3	4	5	6
Castle Doctrine Law	0.0801**	0.0946***	0.0937***	0.0955**	0.0985***	0.100**
	(0.0342)	(0.0279)	(0.0290)	(0.0367)	(0.0299)	(0.0388)
0 to 2 years before adoption of castle doctrine law				0.00398		
				(0.0222)		
Observations	550	550	550	550	550	550

TWFE estimation

TWFE regression (ignoring the covariates) is equivalent to OLS of $\tilde{y}_{gt}$ on $\tilde{D}_{gt}$, where

$$\tilde{y}_{gt} = y_{gt} - \overline{y}_g - (\overline{y}_t - \overline{\overline{y}}) \quad \text{and} \quad \tilde{D}_{gt} = D_{gt} - \overline{D}_g - (\overline{D}_t - \overline{\overline{D}})$$

This data transformation reflects the partialing out of the group and time effects.

If the treatment effect is *homogenous*, TWFE will identify the ATT.

What role do covariates play?

How should inference be conducted?

What if the treatment effect is heterogenous?

It's complicated.

With multiple time periods and variation of treatment timing, we compare

1. newly treated units to never treated units.
2. newly treated units to not yet treated units.
3. newly treated units to already treated units.

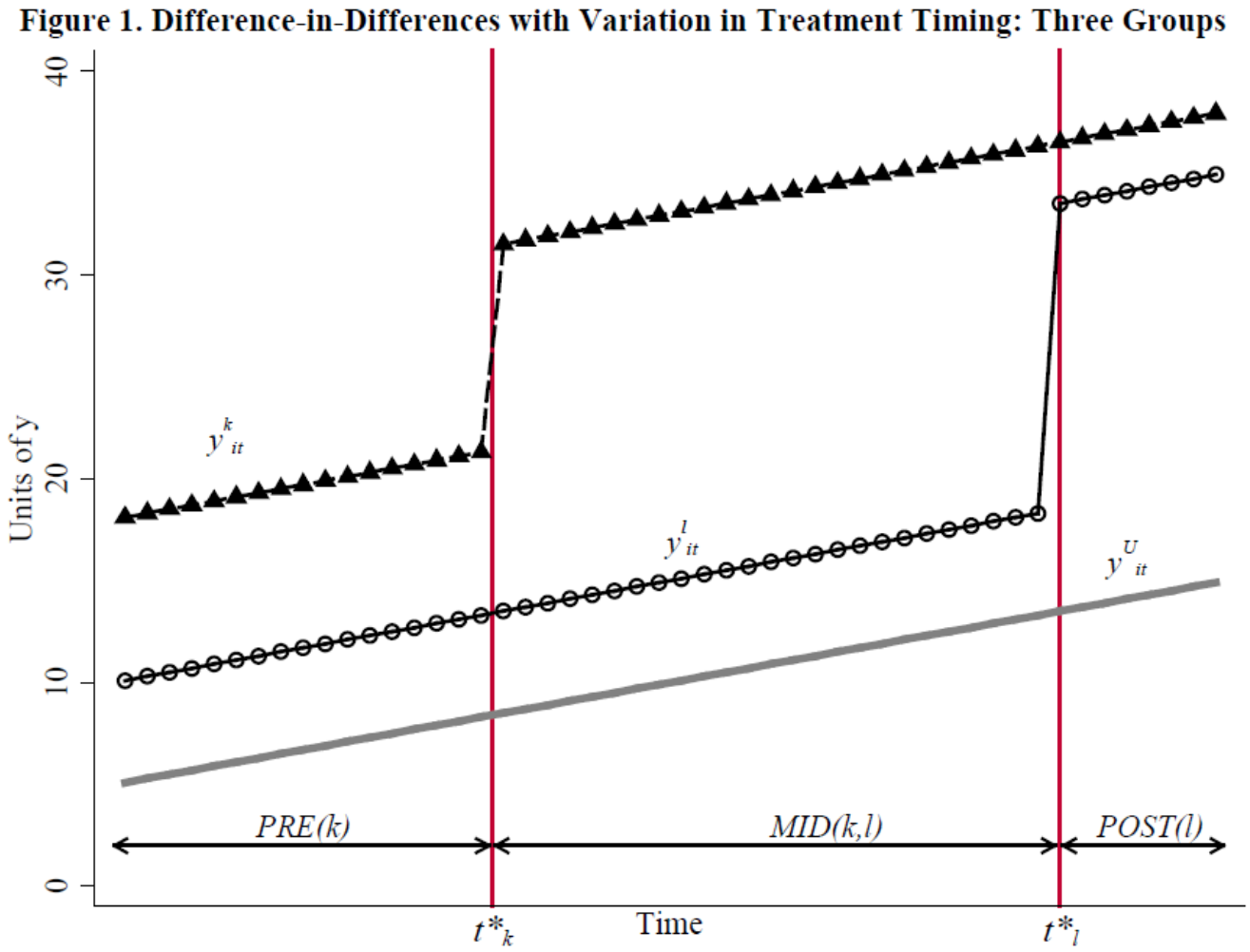
Comparisons 1 and 2 make sense from the standard DD perspective.

Comparison 3 does not because the already treated groups do not represent the path of untreated potential outcomes

Comparison 3 reflects *dynamic* effects of the treatment, which can lead to bias in the estimation of the ATT, even under parallel trends.

TWFE estimates a weighted average of all 2×2 effects.

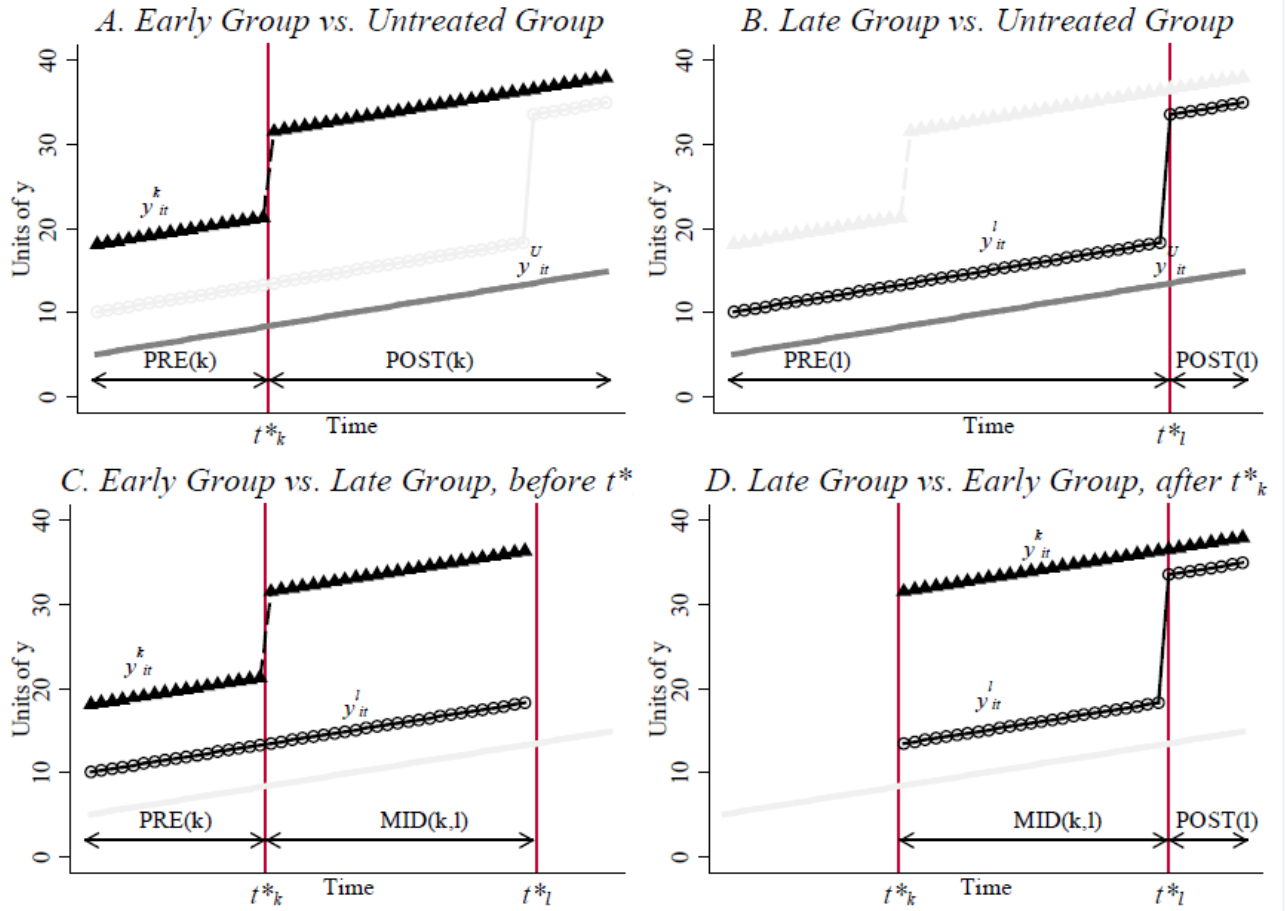
Staggered treatment groups



{Source: Goodman-Bacon (2019)}

2×2 comparisons

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



{Source: Goodman-Bacon (2019)}

Heterogeneous ATT estimation

This is a vibrant area of econometric research.

A popular approach is that proposed by Callaway and Sant'anna (2021). They focus on estimating separate *group-time ATTs*:

$$ATT(g,t) = E[Y_t(g) - Y_t(0)|G = g],$$

where $Y_t(g)$ is the potential outcome in period t if they become treated in period g, $Y_t(0)$ is the untreated potential outcome and G is the time period when treatment starts.

Depending on the number of groups and time periods, the CS estimator will produce many ATTs. This is consistent with the perspective of heterogeneity, but CS also provide guidance for constructing meaningful aggregates.

The CS estimator can be implemented with Callaway's `did` package. You will find a nice explainer in Cunningham's substack.

ECON 7710

Regression Discontinuity

Chris Cornwell

Terry College of Business

Fall 2021

Does drinking increase mortality?

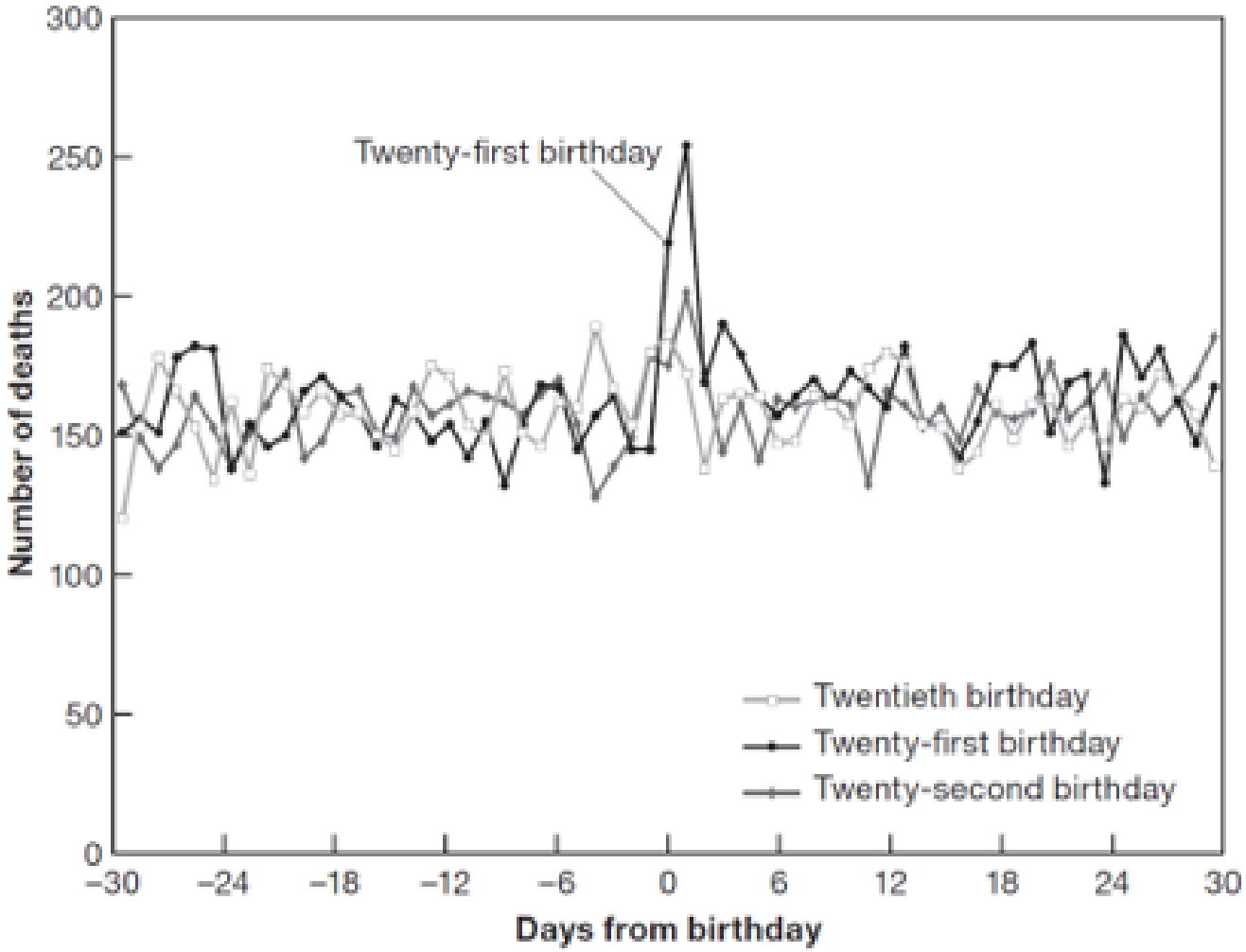
Hard question because drinkers differ from non-drinkers on average in ways that affect health outcomes. It's hard to imagine a rich enough set of controls to simply compare mortality between the two groups.

Carpenter and Dobkin (2009) address the question using a *regression discontinuity* design.

The essence of the design is that minimum legal drinking age (MLDA) laws exogenously change drinking opportunities at age 21, while the non-age-related determinants of drinking are likely unchanging just before and after age 21.

Birthdays and funerals

FIGURE 4.1
Birthdays and funerals



Source: Angrist and Pischke (2015)

Data

Alcohol consumption (National Health Interview Survey)

- Stratified random sample covering the period 1997-2005.
- Drinking during the previous year and lifetime.

Mortality (National Center for Health Statistics)

- Universe of all deaths during the period 1997-2004.
- Focus on individuals aged 19-22.
- Distinguish external (potentially alcohol-related) from internal causes.

Empirical model

Carpenter and Dobkin specify empirical models for each outcome y as

$$y_{ai} = X_{ai}\beta^y + g^y(a) + D_{ai}\pi^y + v^y_{ai}.$$

Variable definitions:

- Outcomes are alcohol consumption of different intensities and mortality of different types for individual i at age a.
- X_{ai} is a vector of covariates.
- $g^y(a)$ is a smooth function of age specified as a low-order polynomial.
- D_{ai} is a 21-year-old indicator.

Drinking participation RD plots

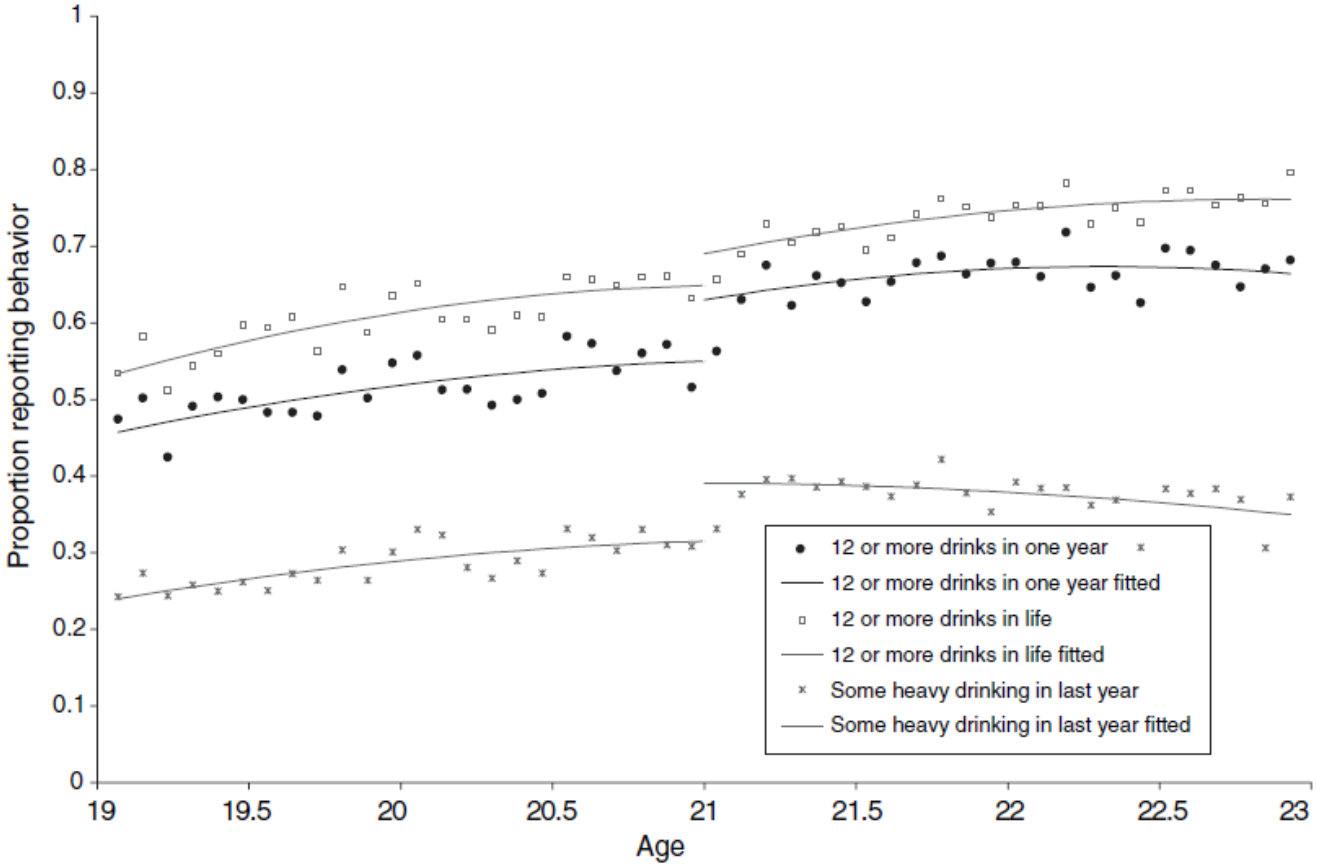


FIGURE 1. AGE PROFILE OF DRINKING PARTICIPATION

Source: Carpenter and Dobkin (2009).

RD overview

In standard regression analysis...

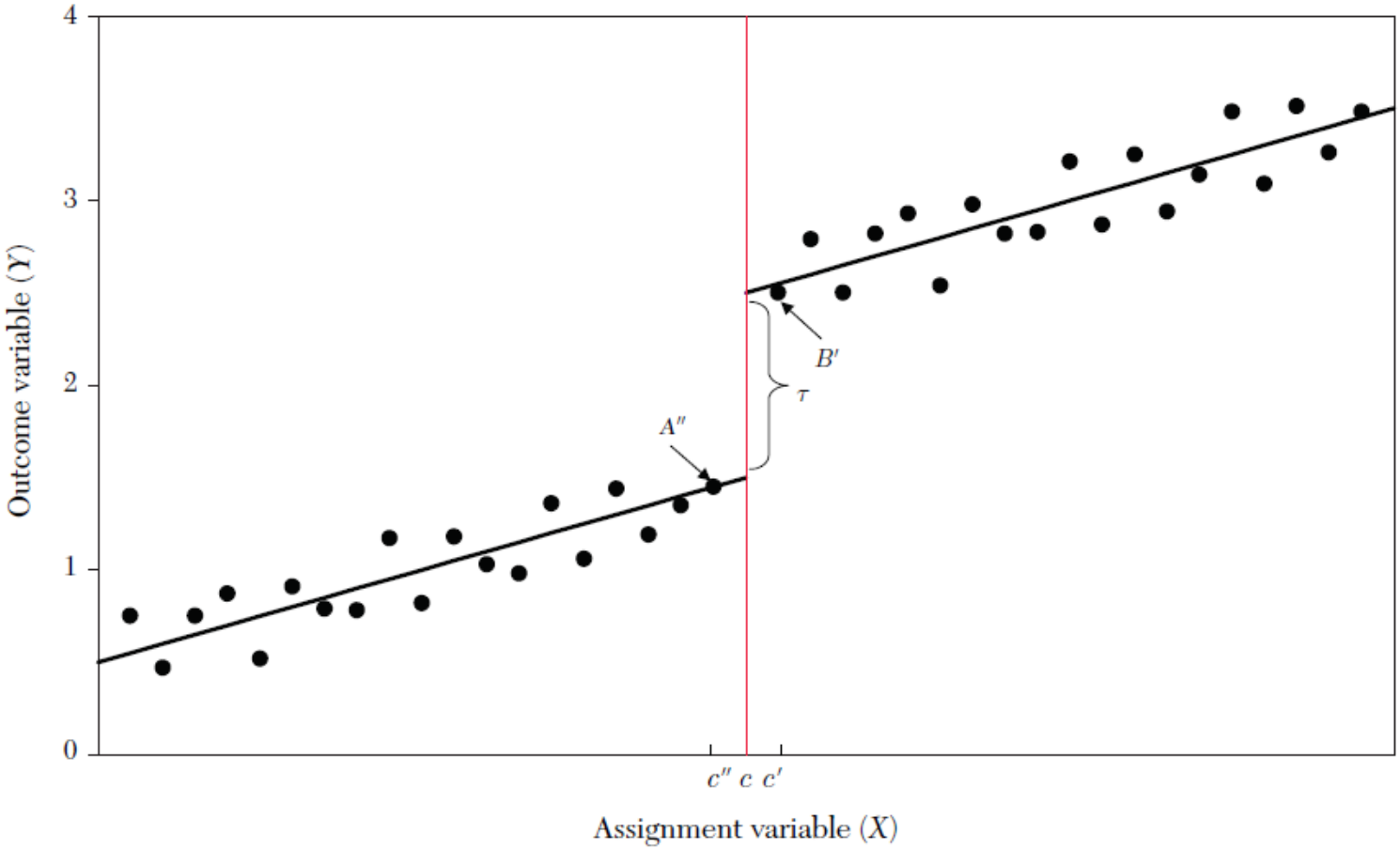
- The treatment effect is identified if the CIA holds.
- Individuals with different values of D and the same values of the controls are "matched" and their conditional means compared.

In (sharp) RD analysis...

- The CIA holds trivially holds because treatment is a deterministic function of a particular covariate, called the *running variable*.
- However, individuals with different values of D have different values of the covariate by construction, so there is no overlap.

So, what does an RD design identify?

Simple RD illustrated



Source: Lee and Lemieux (2010)

Basic sharp RD setup

Suppose treatment assignment is determined by the running variable x:

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

Thus, the CIA holds trivially in this situation:

$$E(y_{gi}|x_i, D_i) = E(y_{gi}|x_i), \quad g = 0, 1.$$

But no overlap means that the usual ATE cannot be estimated without heroic extrapolations. Instead, RD designs focus on a *local ATE* or LATE, defined as the average treatment effect at the cutoff, c:

$$E(y_{1i} - y_{0i}|x_i = c) = \mu_1(c) - \mu_0(c).$$

The key identifying assumption is that the conditional means, the $\mu_g(\cdot)$, are *continuous* at c.

Simple regression formulation

The $\mu_g(\cdot)$ can be easily fit as regressions of observed outcomes on x and D. Recall that we observe:

$$y_i = y_{0i} + (y_{1i} - y_{0i})D_i.$$

In the simplest case, we could specify

$$\begin{aligned} y_{0i} &= \beta_0 + \beta_1 x_i + u_i \\ y_{1i} &= y_{0i} + \delta, \end{aligned}$$

so that

$$y_i = \beta_0 + \beta_1 x_i + \delta D_i + u_i,$$

and

$$\mu_1(c) - \mu_0(c) = \delta.$$

More flexible regression formulations

A common linear specification for both $\mu_1(x_i)$ and $\mu_0(x_i)$ will be too restrictive in most instances.

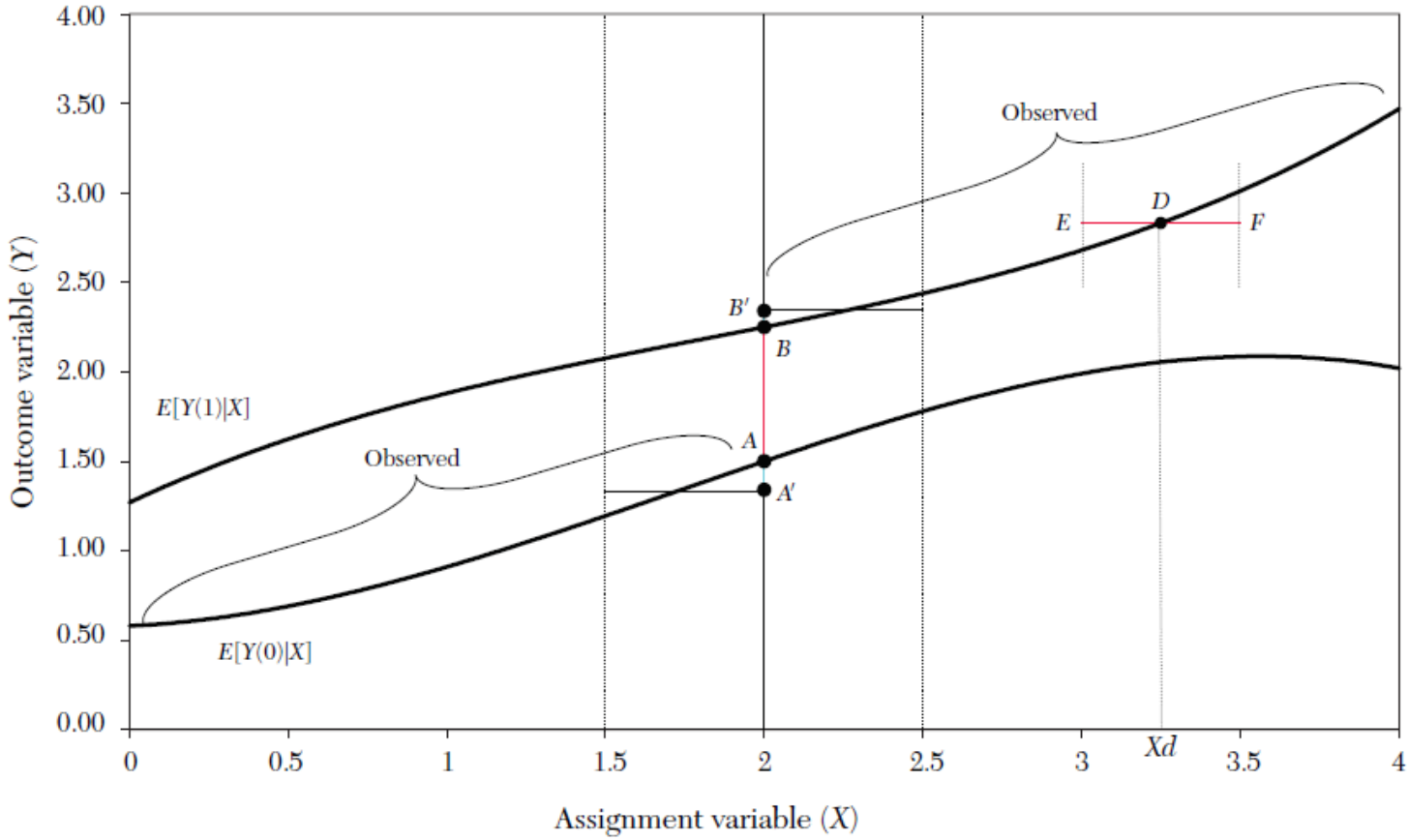
Thus, RD designs typically allow the functional forms of $\mu_1(x_i)$ and $\mu_0(x_i)$ to vary on either side of the cutoff.

This is accomplished by interacting x and D:

$$y_i = \beta_0 + \delta D_i + \beta_1(x_i - c) + \beta_2 D_i(x_i - c) + u_i,$$

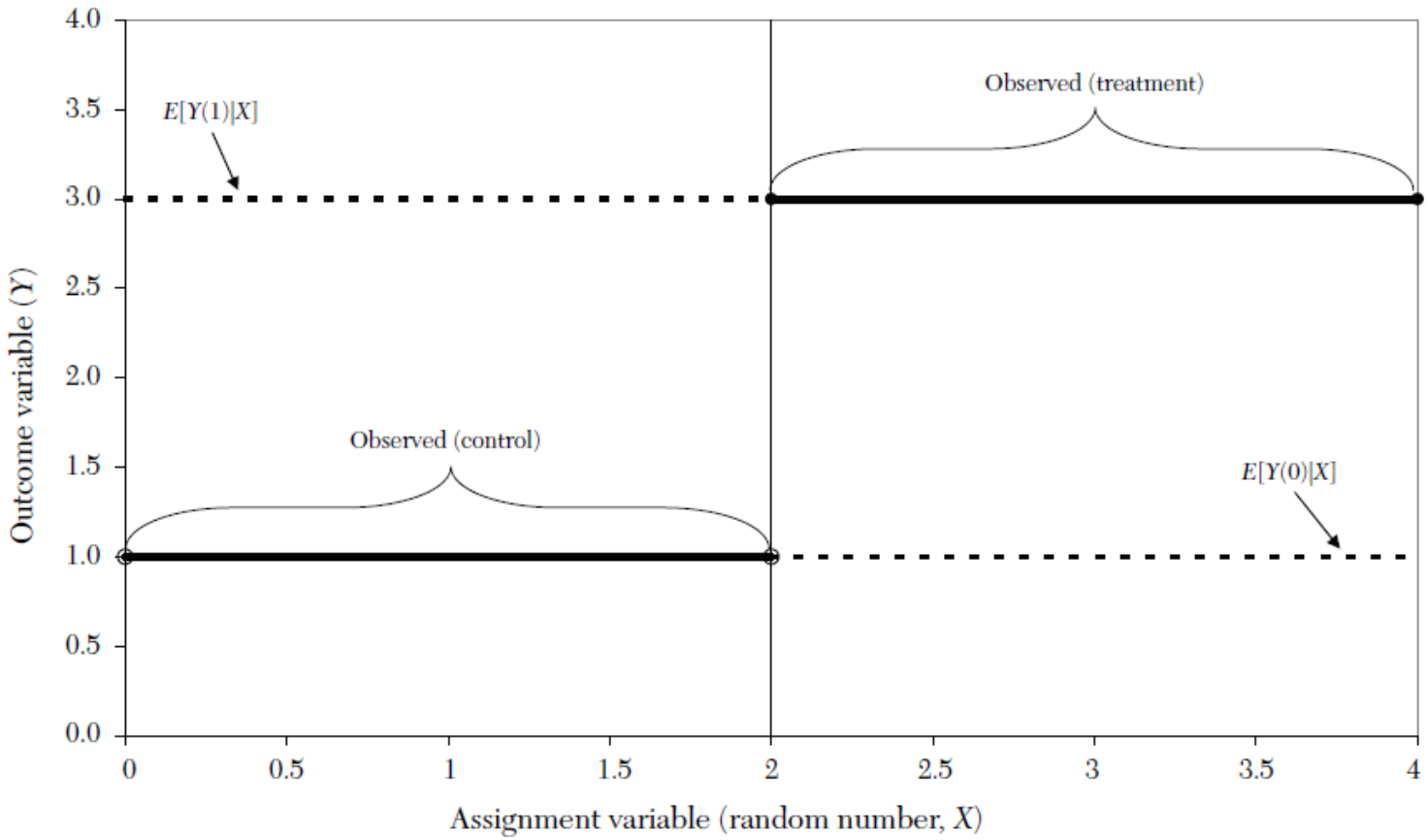
Centering the values of x_i at c guarantees that δ will still have a LATE interpretation. (To see this, just evaluate x_i at c in this formulation.)

Continuity and potential outcomes



Because there is no value of x for which we observe treated and control-group individuals, RD exploits the continuity assumption to extrapolate at the discontinuity in the realized outcome. For observations close to c RD essentially produces a random experiment. Source: Lee and Lemeieux (2010).

A random experiment as an RD design



Compare RD with a random experiment. Here x values are drawn from a uniform(0,4) distribution and treatment is assigned if $x > 2$. Source: Lee and Lemieux (2010).

Getting the functional form right

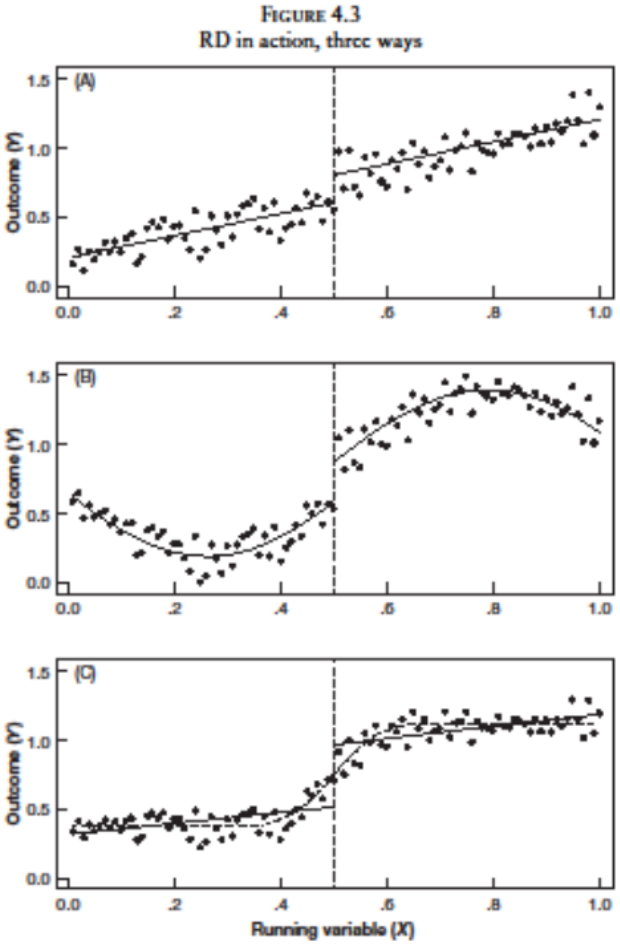


FIGURE 4.3
RD in action, three ways

Notes: Panel A shows RD with a linear model for $E[Y_i|X_i]$; panel B adds some curvature. Panel C shows nonlinearity mistaken for a discontinuity. The vertical dashed line indicates a hypothetical RD cutoff.

Linear model for $f(x)$:

$$y = \alpha + \beta x + \delta D + \epsilon$$

Quadratic model for $f(x)$ with different coefficients on the left and right of c:

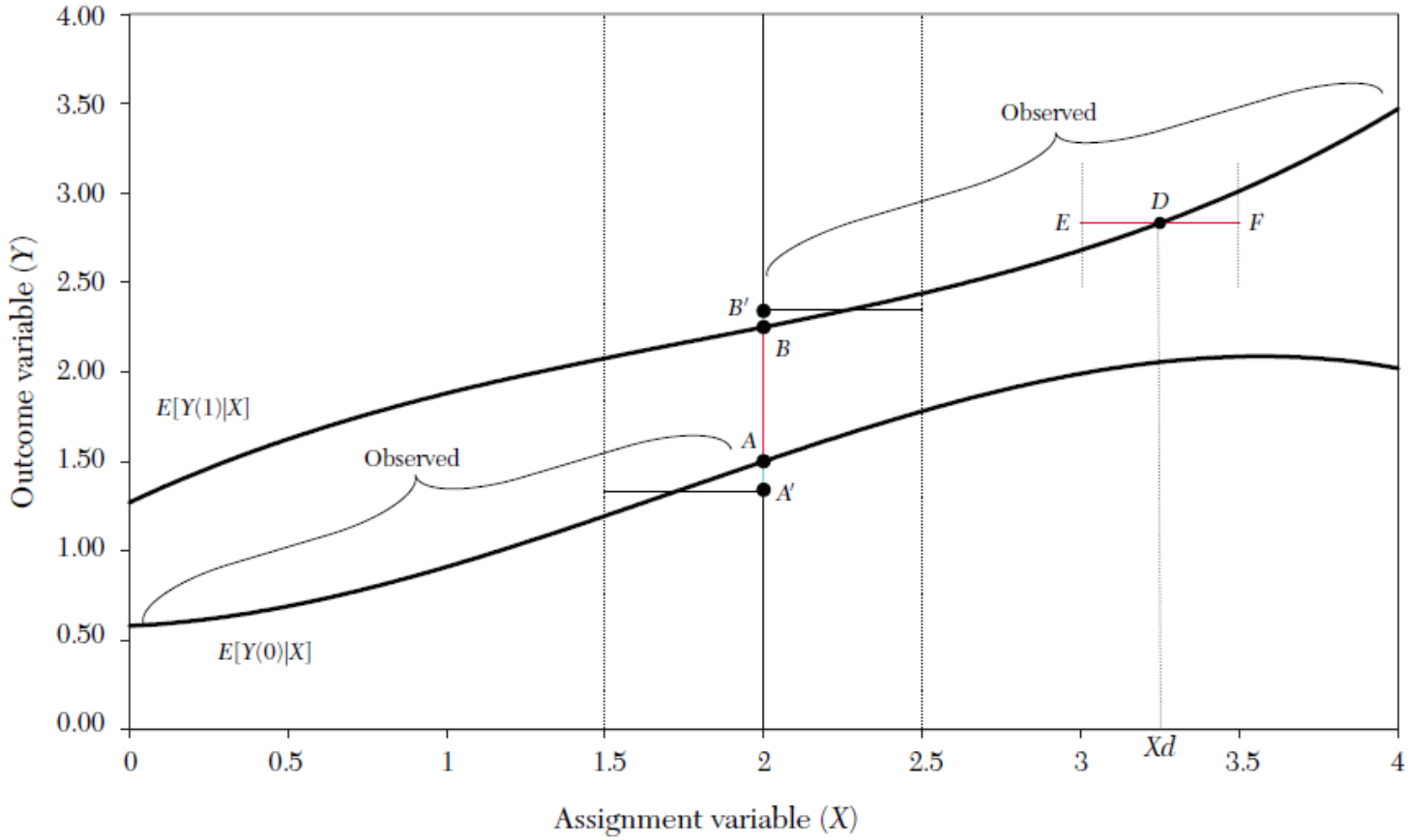
$$y = \alpha + \beta_1(x-c) + \beta_2(x-c)^2 + \beta_3[(x-c)D] + \beta_4[(x-c)^2 D] + \delta D + \epsilon$$

Center x at c to ensure that δ reflects the jump at c.

Nonlinearity mistaken for a discontinuity because $f(x)$ is incorrectly specified as linear.

Source: Angrist and Pischke (2015).

Local averaging and bin size



Estimate $y = \alpha + \beta x + \delta D + u$ in a narrow bin around c ($c - b \leq c \leq c + b$). What are the tradeoffs with smaller and larger bin sizes? Source: Lee and Lemeieux (2010).

Defending an RD design

- Show the distribution of the running variable. (Is it smooth?)
- Present the main RD graph using binned local averages. (Is it there?)
- Graph a benchmark specification. (Check conditional mean specification.)
- Explore the sensitivity of the results to a range of bandwidths and functional forms. (Are the results robust?)
- Conduct a parallel RD analysis on the baseline covariates. (There should be no discontinuities.)
- Explore the sensitivity of the results to the inclusion of baseline covariates. (Does the estimated treatment effect stand up?)

Packages like `rddtools` and `rdrobust` automate some of these defensive actions.

Back to Carpenter and Dobkin

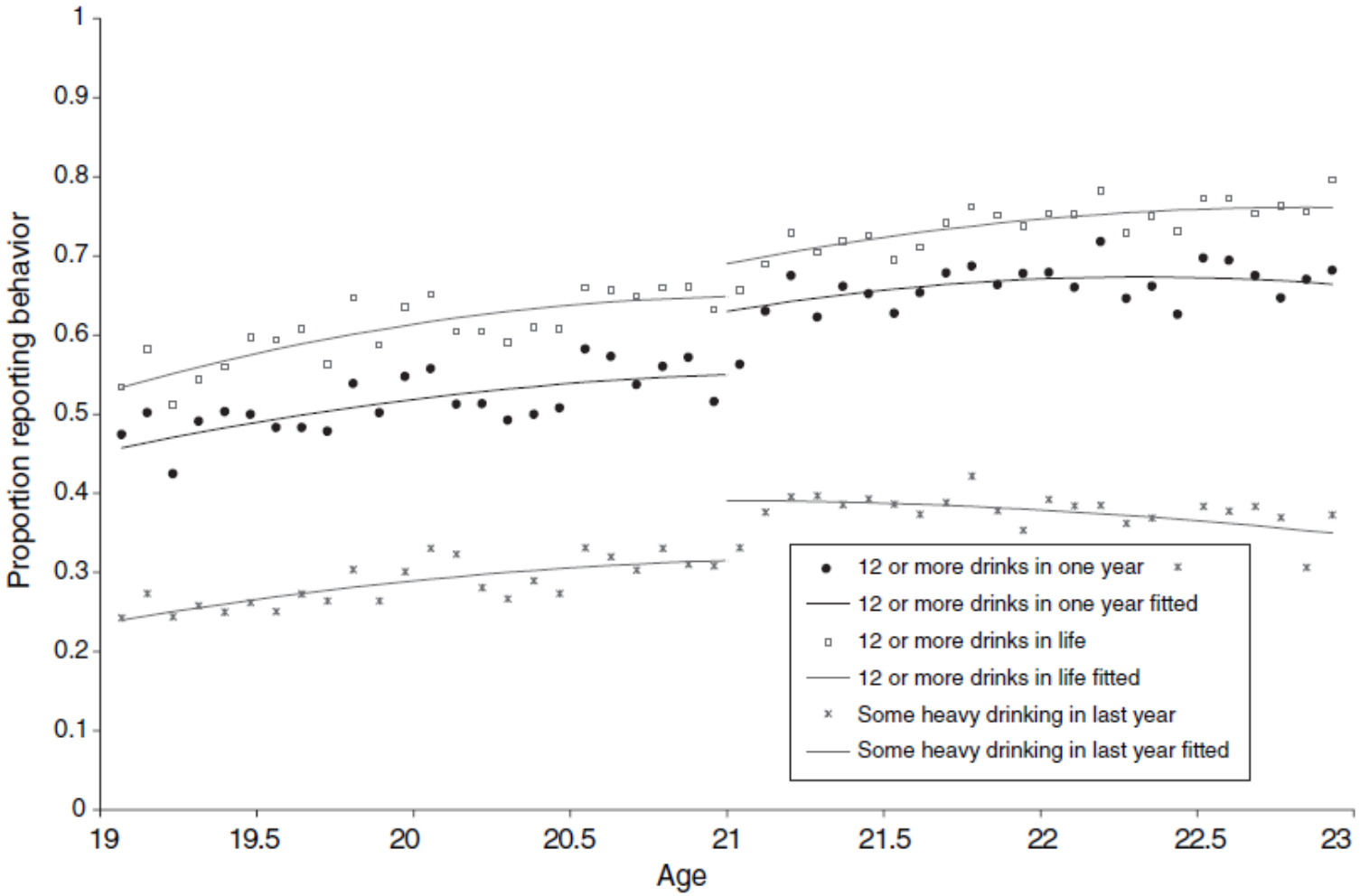


FIGURE 1. AGE PROFILE OF DRINKING PARTICIPATION

Source: Carpenter and Dobkin (2009).

Drinking participation RD estimates

TABLE 1—ALCOHOL CONSUMPTION: PARTICIPATION

	(1)	(2)	(3)	(4)	(5)
12 or more drinks in lifetime					
Over 21	0.0418	0.0316	0.0268	0.0198	0.0199
	(0.0242)	(0.0301)	(0.0292)	(0.0423)	(0.0179)
Observations	16,107	16,107	16,107	16,107	
R^2	0.02	0.03	0.10	0.10	
Prob > Chi-Squared			0.00	0.61	
12 or more drinks in one year					
Over 21	0.0796	0.0657	0.0611	0.0603	0.0461
	(0.0254)	(0.0313)	(0.0301)	(0.0438)	(0.0218)
Observations	16,107	16,107	16,107	16,107	
R^2	0.02	0.03	0.11	0.11	
Prob > Chi-Squared			0.00	0.56	
Any heavy drinking in last year					
Over 21	0.0761	0.0527	0.0492	0.0262	0.0398
	(0.0248)	(0.0304)	(0.0291)	(0.0430)	(0.0201)
Observations	16,107	16,107	16,107	16,107	
R^2	0.01	0.01	0.10	0.10	
Prob > Chi-Squared			0.00	0.67	
Covariates	N	N	Y	Y	N
Weights	N	Y	Y	Y	N
Quadratic terms	Y	Y	Y	Y	N
Cubic terms	N	N	N	Y	N
LLR	N	N	N	N	Y

Source: Carpenter and Dobkin (2009).

Drinking intensity RD plots

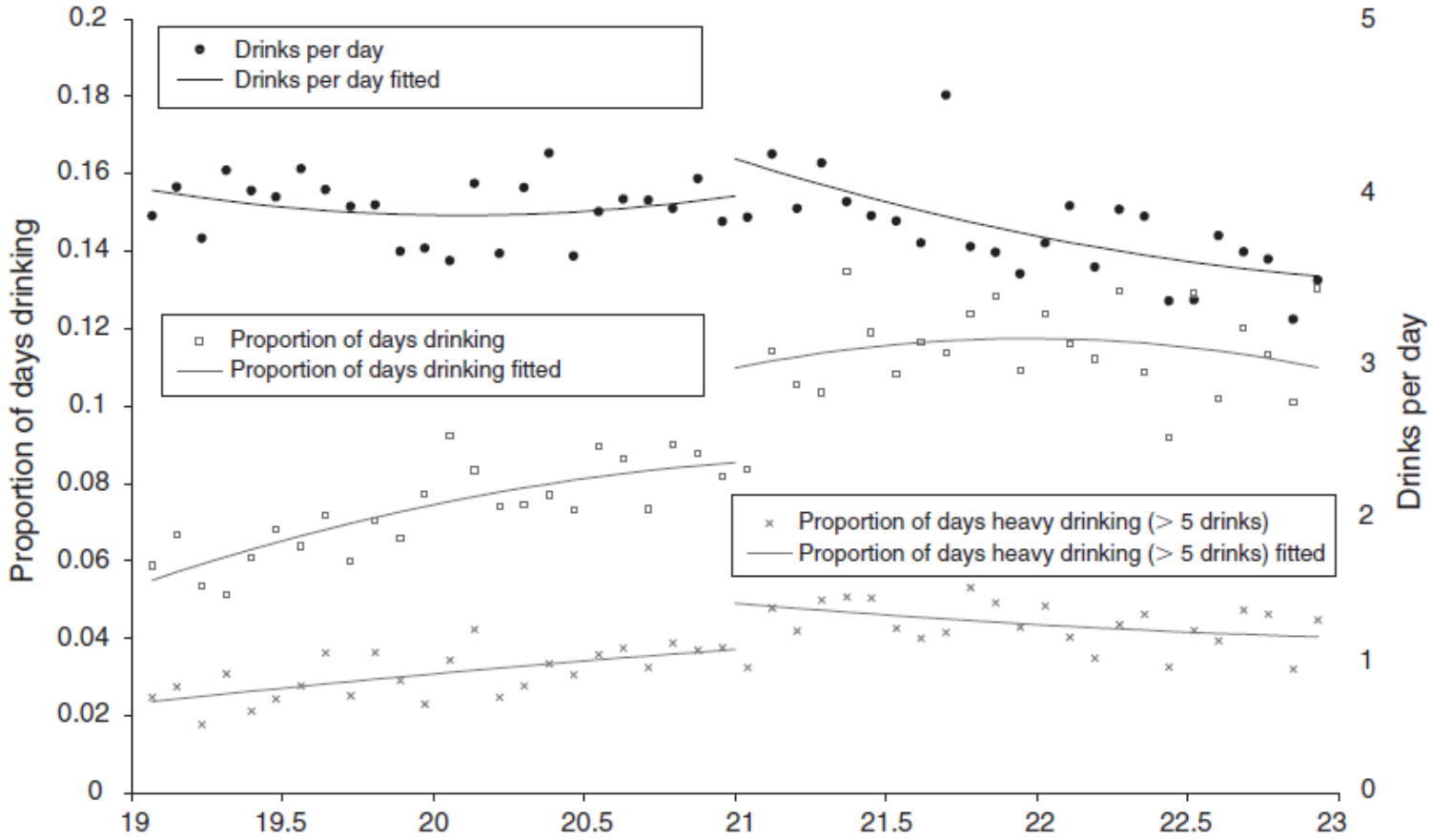


FIGURE 2. AGE PROFILE OF DRINKING INTENSITY

Source: Carpenter and Dobkin (2009).

Drinking intensity RD estimates

TABLE 2—ALCOHOL CONSUMPTION: INTENSITY

	(1)	(2)	(3)	(4)	(5)
Proportion of days drinking					
Over 21	0.0245	0.0180	0.0182	0.0119	0.0107
	(0.0086)	(0.0097)	(0.0095)	(0.0135)	(0.0072)
Observations	16,107	16,107	16,107	16,107	
R^2	0.02	0.02	0.07	0.07	
Prob > Chi-Squared			0.00	0.56	
Proportion of days heavy drinking					
Over 21	0.0120	0.0075	0.0075	0.0021	0.0026
	(0.0061)	(0.0063)	(0.0062)	(0.0091)	(0.0048)
Observations	15,825	15,825	15,825	15,825	
R^2	0.00	0.01	0.05	0.05	
Prob > Chi-Squared			0.00	0.72	
Drinks per day on days drinking					
Over 21	0.2387	0.2068	0.2465	0.2806	0.1886
	(0.2810)	(0.3403)	(0.3291)	(0.4782)	(0.2024)
Observations	9,906	9,906	9,906	9,906	
R^2	0.00	0.00	0.07	0.07	
Prob > Chi-Squared			0.00	0.92	
Covariates	N	N	Y	Y	N
Weights	N	Y	Y	Y	N
Quadratic terms	Y	Y	Y	Y	N
Cubic terms	N	N	N	Y	N
LLR	N	N	N	N	Y

Source: Carpenter and Dobkin (2009).

Testing for covariate discontinuities

TABLE 3—AGE PROFILE OF DEMOGRAPHIC CHARACTERISTICS FROM THE NHIS

	Male (1)	White (2)	Black (3)	Hispanic (4)	No HS Diploma (5)	HS Diploma (6)	Employed (7)	Looking for work (8)	No health insurance (9)
Over 21	0.0156	0.0172	−0.0250	0.0095	−0.0108	0.0040	0.0019	−0.0061	0.0043
	(0.0260)	(0.0293)	(0.0211)	(0.0203)	(0.0230)	(0.0291)	(0.0302)	(0.0171)	(0.0293)
Constant	0.4405	0.6440	0.1415	0.1638	0.1810	0.2941	0.6453	0.0849	0.2970
	(0.0168)	(0.0195)	(0.0137)	(0.0129)	(0.0154)	(0.0194)	(0.0203)	(0.0114)	(0.0191)
Observations	16,107	16,107	16,107	16,107	16,107	16,107	16,107	16,107	16,107
R^2	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.00	0.00

Notes: Sample from NHIS Sample Adult File (1997–2005). Standard errors in parentheses. The regressions do not include covariates other than a second order polynomial in age interacted with the over 21 dummy. Since Age = persons age − 21, the constant is the predicted value for people about to turn 21.

Source: Carpenter and Dobkin (2009).

Mortality RD plots, external vs internal

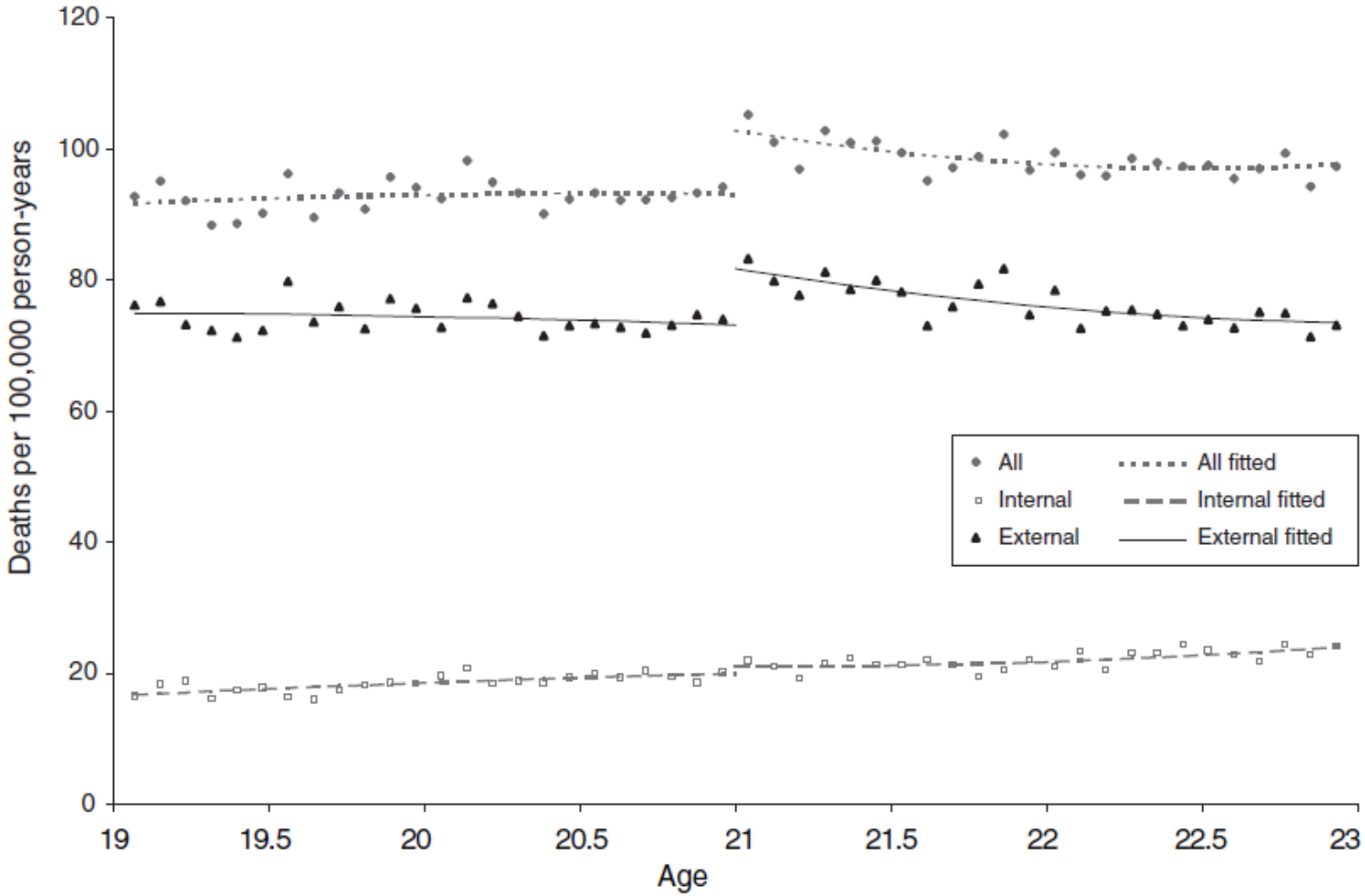


FIGURE 3. AGE PROFILE FOR DEATH RATES

Source: Carpenter and Dobkin (2009).

Mortality RD estimates, external vs internal

TABLE 4—DISCONTINUITY IN LOG DEATHS AT AGE 21

	(1)	(2)	(3)	(4)
Deaths due to all causes				
Over 21	0.096	0.087	0.091	0.074
	(0.018)	(0.017)	(0.023)	(0.016)
Observations	1,460	1,460	1,460	1,458
R^2	0.04	0.05	0.05	
Prob > Chi-Squared		0.000	0.735	
Deaths due to external causes				
Over 21	0.110	0.100	0.096	0.082
	(0.022)	(0.021)	(0.028)	(0.021)
Observations	1,460	1,460	1,460	1,458
R^2	0.06	0.08	0.08	
Prob > Chi-Squared		0.000	0.788	
Deaths due to internal causes				
Over 21	0.063	0.054	0.094	0.066
	(0.040)	(0.040)	(0.053)	(0.031)
Observations	1,460	1,460	1,460	1,458
R^2	0.10	0.10	0.10	
Prob > Chi-Squared		0.000	0.525	
Covariates	N	Y	Y	N
Quadratic terms	Y	Y	Y	N
Cubic terms	N	N	Y	N
LLR	N	N	N	Y

Notes: See Notes from Table 1. The dependent variable is the log of the number of deaths that occurred x days from the person's twenty-first birthday. External deaths include all deaths with mention of an injury, alcohol use, or drug use. The Internal Death category includes all deaths Nt coded as external. Please see Web Appendix C for the ICD codes for each of the categories above. The first three columns give the estimates from polyNmial regressions on age interacted with a dummy for being over 21.

Source: Carpenter and Dobkin (2009).

Mortality RD plots, external

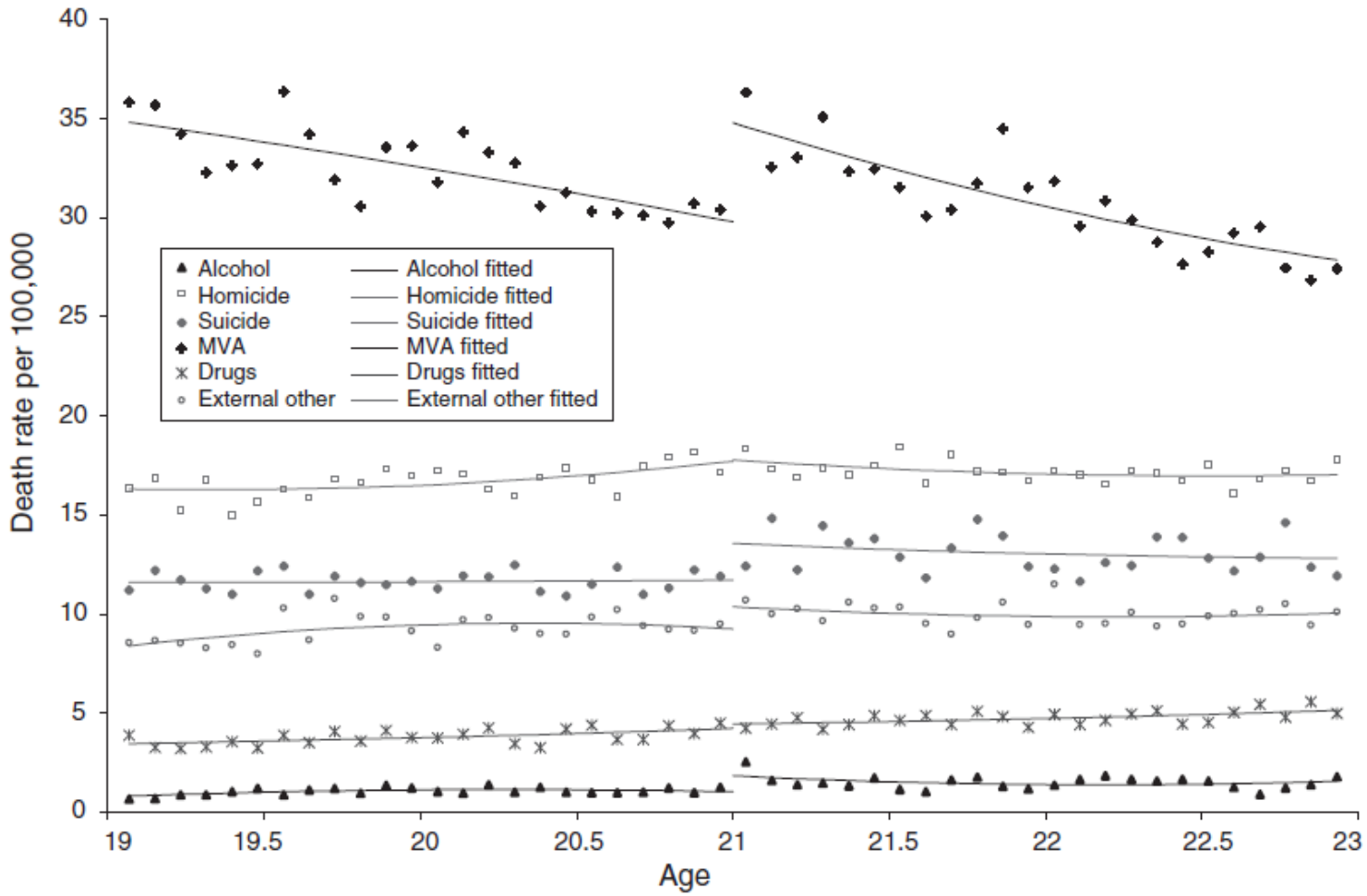


FIGURE 4. AGE PROFILES FOR DEATH RATES BY EXTERNAL CAUSE

Source: Carpenter and Dobkin (2009).

Mortality RD estimates, external

TABLE 5—DISCONTINUITY IN LOG DEATHS AT AGE 21 DUE TO EXTERNAL CAUSES

	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
	Alcohol				Homicide			
Over 21	0.388	0.346	0.406	0.441	0.009	0.002	−0.003	−0.014
	(0.119)	(0.116)	(0.156)	(0.117)	(0.045)	(0.045)	(0.061)	(0.041)
Observations	1,460	1,460	1,460	1,458	1,460	1,460	1,460	1,458
R^2	0.03	0.04	0.04			0.01	0.01	0.01
Prob > Chi-Squared		0.000	0.228			0.000	0.495	
	Suicide				Motor vehicle accidents			
Over 21	0.160	0.154	0.135	0.105	0.158	0.143	0.145	0.139
	(0.059)	(0.059)	(0.086)	(0.045)	(0.033)	(0.032)	(0.044)	(0.032)
Observations	1,460	1,460	1,460	1,458	1,460	1,460	1,460	1,458
R^2	0.02	0.02	0.02	0.15	0.16	0.16		
Prob > Chi-Squared		0.000	0.892	0.000	0.666			
	Drugs				Other external causes			
Over 21	0.070	0.067	0.004	−0.016	0.087	0.098	0.098	0.074
	(0.081)	(0.082)	(0.107)	(0.078)	(0.060)	(0.059)	(0.075)	(0.043)
Observations	1,460	1,460	1,460	1,458	1,460	1,460	1,460	1,458
R^2	0.04	0.04	0.04	0.01	0.01	0.01		
Prob > Chi-Squared		0.000	0.643			0.000	0.877	
Covariates	N	Y	Y	N	N	Y	Y	N
Quadratic terms	Y	Y	Y	N	Y	Y	Y	N
Cubic terms	N	N	Y	N	N	N	Y	N
LLR	N	N	N	Y	N	N	N	Y

Source: Carpenter and Dobkin (2009).

Implication of findings

"Our estimates suggest that reducing the drinking age nationally to age 20 would result in approximately 408 additional deaths among 20-year-olds. Given a value of a statistical life of $8.4 million (in 2007 US dollars) this gives a total cost of about $3.4 billion per year."

ECON 7710

Instrumental Variables

Chris Cornwell

Terry College of Business

Fall 2021

Throwback to regression

Consider the simple regression model,

$$y_i = \beta x_i + u_i, \quad i = 1, \ldots, N, \tag{1}$$

where $E(u_i|x_i) = 0$. In this case we know that

$$\begin{aligned} E(x_i u_i) = E[x_i(y_i - \beta x_i)] &= 0 \\ \Rightarrow E(x_i y_i) &= \beta E(x_i^2) \\ \Rightarrow \beta &= \frac{E(x_i y_i)}{E(x_i^2)}, \end{aligned}$$

which can be consistently estimated as

$$\hat{\beta}_{OLS} = \frac{\frac{1}{N}\sum x_i y_i}{\frac{1}{N}\sum x_i^2} = \frac{\sum x_i y_i}{\sum x_i^2}.$$

What if

What if $E(u_i|x_i) \neq 0$, but we had access to another variable, z, such that $E(u_i|z_i) = 0$ and $E(z_i x_i) \neq 0$?

Then, we could identify β in (1) as

$$\begin{aligned} E(z_i u_i) = E[z_i(y_i - \beta x_i)] &= 0 \\ \Rightarrow E(z_i y_i) &= \beta E(z_i x_i) \\ \Rightarrow \beta &= \frac{E(z_i y_i)}{E(z_i x_i)}, \end{aligned}$$

which can be consistently estimated as

$$\hat{\beta}_{IV} = \frac{\frac{1}{N}\sum z_i y_i}{\frac{1}{N}\sum z_i x_i} = \frac{\sum z_i y_i}{\sum z_i x_i}.$$

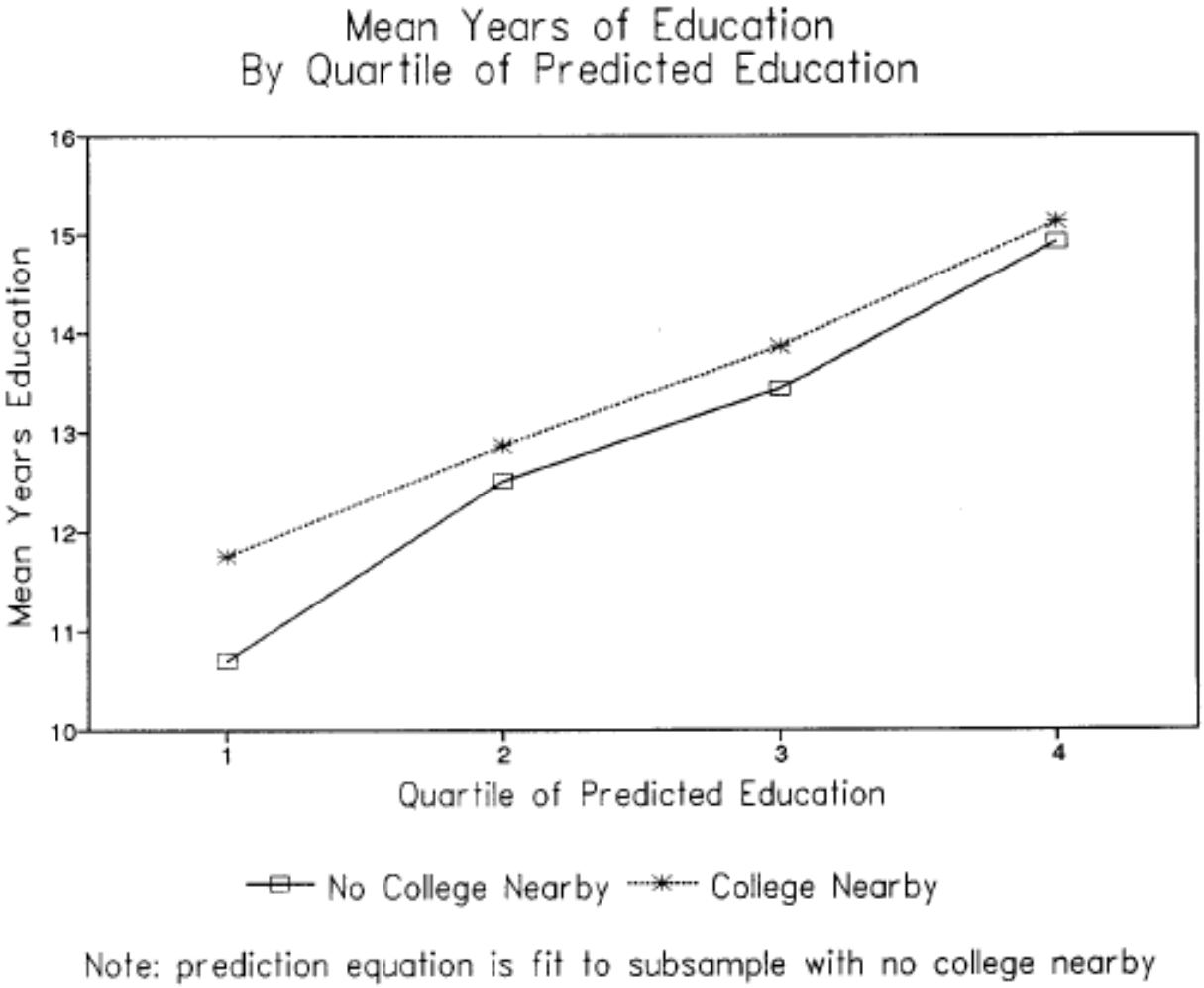
What is the returns to schooling for young men?

In principle, with a random sample of young men and a rich set of controls, we could credibly invoke ignorability of treatment (the CIA) and overlap and estimate the returns to schooling using OLS.

Unfortunately, that won't sell because there are always important unobservable characteristics that would call a selection-on-observables story into question.

But what if we had access to a source of variation in education that is plausibly *exogenous*, allowing us regard treatment assignment (to higher levels of education) to be as good as random. This is the approach Card (1995) takes with college proximity, which he uses as an *instrument* for education.

Predicting education by college proximity



Source: Card (1995). Prediction equation includes region and urban indicators (as of 1966), age and race indicators and family background variables.

Data

Recall that Card's sample was constructed from the National Longitudinal Survey of Young Men (NLSYM), which began with 5525 men between 14-24 years old in 1966. His analysis is based on the 1976 wave, so the men in his sample are 24-34 years old.

Table 1: Summary statistics of main variables

Statistic	N	Mean	St. Dev.	Min	Max
wage	3,010	577.282	262.958	100	2,404
educ	3,010	13.263	2.677	1	18
exper	3,010	8.856	4.142	0	23
black	3,010	0.234	0.423	0	1
south	3,010	0.404	0.491	0	1
smsa	3,010	0.713	0.452	0	1
nearc4	3,010	0.682	0.466	0	1
nearc2	3,010	0.441	0.497	0	1

Empirical model

Card's empirical model involves two equations: one that is *structural* in the sense it models the causal effect of interest, and one called the *first stage* that captures the relationship between the instrument(s) and treatment assignment:

$$y_i = X_{1i}\alpha_1 + S_i\beta + u_i \quad \text{(structural equation)}$$
$$S_i = X_{1i}\gamma_1 + C_i\delta_0 + v_i \quad \text{(first stage)},$$

where y_i is the log wage for person i, S_i is years of schooling, X_{1i} is a set of controls and C_i is an indicator for growing up with a college nearby.

Because S is likely correlated with u, OLS applied to the structural equation will not identify the causal effect of schooling on log wages.

IV overview

Consider the simple *structural model*,

$$y = \alpha + \delta D + u. \tag{2}$$

You can think of IV as a chain reaction:

$$\text{instrument } (z) \rightarrow \text{treatment assigment } (D) \rightarrow \text{outcome } (y)$$

In the Card study, the chain reaction runs from college proximity (C) to educational attainment (S) to log wages (y).

What's important is that we can regard z as good as random and z affects y only through D.

Here, "as good as random" means that z is independent of the potential outcomes (y_0, y_1) and potential treatment assignments (D_0, D_1)

IV chain reaction

The chain reaction starts with the *first stage*:

$$D = \pi_0 + \pi_1 z + v_D. \tag{3}$$

The path to the outcomes runs through the *reduced form*, which expresses the outcome in terms of treatment:

$$\begin{aligned} y &= \alpha + \delta D + u \quad \text{(structural equation)} \\ &= \alpha + \delta(\pi_0 + \pi_1 z + v_D) + u \\ &= (\alpha + \delta\pi_0) + \delta\pi_1 z + (\delta v_D + u) \\ &= \gamma_0 + \gamma_1 z + v_y, \end{aligned} \tag{4}$$

which implies the effect of D on y is

$$\delta\pi_1 = \gamma_1 \Rightarrow \delta = \frac{\gamma_1}{\pi_1} \tag{5}$$

IV identifies a LATE

It is straightforward to see that the effect of D on y is

$$\delta\pi_1 = \gamma_1 \Rightarrow \delta = \frac{\gamma_1}{\pi_1} = \frac{E(y|z=1) - E(y|z=0)}{E(D|z=1) - E(D|z=0)}, \tag{6}$$

which is a *Local Average Treatment Effect (LATE)*.

Return to the context of the Card paper, where z is college proximity and D is categorical (like high-school graduation) and y is (log) wages.

Then IV identifies the effect of D on y as the difference in mean log wages between those who grew up near a college and those who did not divided by the difference in HSG graduate rates between those two groups.

The IV estimator of δ in equation (4) can be obtained by simply replacing the conditional expectations with the corresponding sample averages. This is sometimes called the *Wald* or *grouping estimator*.

How would you interpret δ if treatment is *continuous*?

A LATE for whom

There are four sub-populations defined by treatment assignment. Let's consider them in the Card context:

- *Never-takers* ($D_1 = D_0 = 0$). Individuals who would never complete high school, whether or not they grew up near a college.
- *Always-takers* ($D_1 = D_0 = 1$). Individuals who would always complete high school, whether or not they grew up near a college.
- *Compliers* ($D_1 - D_0 = 1$). Individuals who complete high school because they grew up near a college.
- *Defiers* ($D_1 - D_0 = -1$). Individuals who would NOT complete high school because they grew up near a college. (Rule out.)

IV identifies

$$\delta = \frac{\gamma_1}{\pi_1} = \frac{E(y|z=1) - E(y|z=0)}{E(D|z=1) - E(D|z=0)} = E(y_1 - y_0|D_1 > D_0), \tag{7}$$

or the effect of the treatment on the sub-population of compliers.

LATE theorem

More formally, if

1. z is as good as random;
2. z is excludable from the structural equation;
3. the first stage exists; and
4. there are no defiers,

$$\delta = \frac{\gamma_1}{\pi_1} = \frac{E(y|z=1) - E(y|z=0)}{E(D|z=1) - E(D|z=0)} = E(y_1 - y_0|D_1 > D_0).$$

Comments on complier sub-populations

Different instruments identify different LATEs. Thus, differences in complier sub-populations may help you sort out variation in estimated LATEs across instruments.

To the extent complier sub-populations are similar to other populations of interest, *external validity* is enhanced.

To the extent complier sub-populations differ but the estimated LATEs don't, a claim of *treatment effect homogeneity* is more plausible.

Compliers cannot be individually identified, but you can obtain the size of the complier population from the first stage and describe the distribution of their characteristics.

2SLS: IV in practice

In practice, IV is implemented in contexts where there are controls and multiple instruments, which leads us to *two-stage least squares (2SLS)*.

Consider a general formulation of the structural equation and first stage:

$$\begin{aligned} y_i &= \beta_0 + \delta D_i + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + u_i \\ D_i &= \pi_0 + \pi_1 x_{i1} + \ldots + \pi_K x_{iK} \\ &\quad + \pi_{K+1} z_{i1} + \ldots + \pi_{K+J} z_{iJ} + v_i. \end{aligned} \tag{8}$$

2SLS involves:

1. OLS on the first stage to construct $\hat{D}_i$.
2. OLS of y_i on $\hat{D}_i$ and the x_{ik}.

The coefficient of $\hat{D}_i$ in the second stage is the 2SLS estimate of δ.

2SLS as IV

Recall from the LATE theorem, a simple IV design identifies δ as the ratio of reduced-form and first-stage coefficients:

$$\delta = \frac{\gamma_1}{\pi_1} = \frac{\text{cov}(y, z)/\text{var}(z)}{\text{cov}(D, z)/\text{var}(z)}. \tag{9}$$

This perspective carries over to 2SLS with a residualized z:

$$\delta = \frac{\gamma_1}{\pi_1} = \frac{\text{cov}(y, \tilde{z})/\text{var}(\tilde{z})}{\text{cov}(D, \tilde{z})/\text{var}(\tilde{z})}, \tag{10}$$

where $\tilde{z}$ is a residual from a regression of z on the xs.

2SLS inference

2SLS standard errors are computed from residuals constructed from 2SLS coefficient estimates:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\delta} D_i + \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_K x_{iK},$$

NOT the residuals from the second-stage regression:

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\delta} \widehat{\mathbf{D}}_\mathbf{i} + \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_K x_{iK}.$$

This is one reason we don't do 2SLS manually in two stages. Instead, we rely on IV packages like `ivreg`.

Importance of first-stage fit

In principle, the more instruments you have the better will be the first-stage fit, but not if they are *weak*. A weak instrument contributes little to explaining D.

IV is generally *biased* and the bias is worse when you use many weak instruments.

Every application of IV should check the first-stage fit by reporting the F statistic for the joint significance of the zs. The value of the F statistic will be produced if you choose `diagnostics=TRUE` when applying the `summary` function to your IV estimation object.

As it turns out, the bias of IV relative to OLS is proportional to $\frac{1}{F}$. Thus, an F of 10 would imply the bias of IV is roughly 10 percent of the OLS bias.

Consequently, an F of 10 has become a popular rule of thumb in evaluating first-stage fit.

Testing the validity of extra instruments

Make no mistake, we cannot test instrument validity in the *exactly identified* case as spelled out in assumptions 1 and 2 of the LATE theorem.

However, it is possible to test the validity of extra instruments in the *overly identified* case.

Consider the general model described in equation (6). A popular procedure for testing the *over-identifying restrictions* is to regression 2SLS residuals on the xs and zs and conduct an F test of the joint significance of the zs.

This is often referred to as a *Sargan test*. The (asymptotic) version of Sargan test statistic is also reported if you choose `diagnostics=TRUE` when applying the `summary` function to your IV estimation object.

2SLS best practices

- Report the exactly identified case because it involves the weakest assumptions. Examine
  - first stage
  - reduced form
  - structural equation

  Does each component make sense?
- Report the F statistic for the joint significance of the zs. The bigger (at least 10), the better.
- Report Sargan tests for the over-identifying restrictions.

Back to Card (1995)

Recall Card's empirical model:

$$y_i = X_{1i}\alpha_1 + S_i\beta + u_i \quad \text{(structural equation)}$$
$$S_i = X_{1i}\gamma_1 + C_i\delta_0 + v_i \quad \text{(first stage)},$$

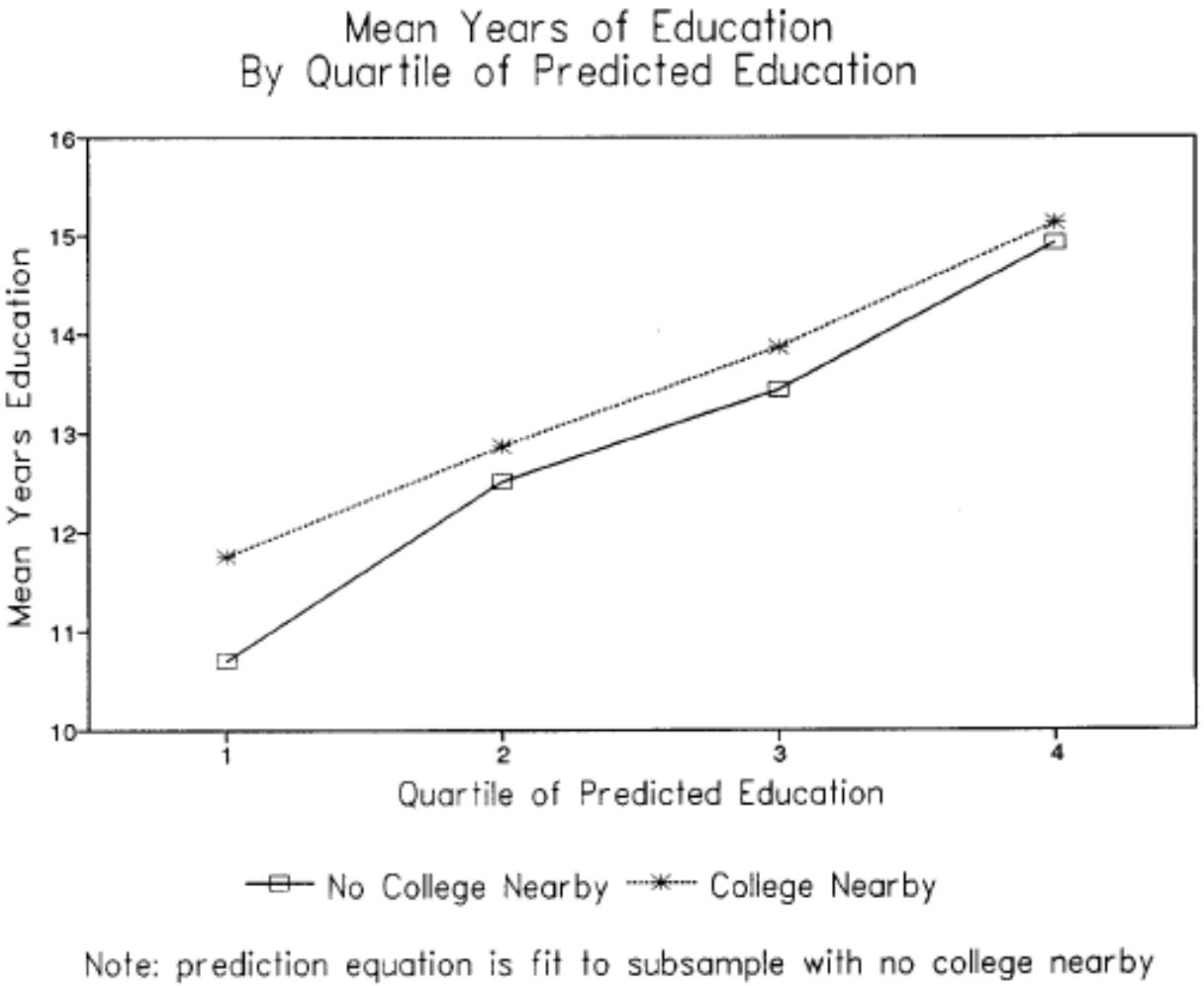
Again, the fundamental problem is that S is likely correlated with u, so OLS applied to the structural equation will not identify the causal effect of schooling on log wages.

OLS results

Table 2: Estimated Regression Models for Log Hourly Earnings

		(1)	(2)	(3)	(4)	(5)
1.	Education	0.074	0.075	0.073	0.074	0.073
		(0.004)	(0.003)	(0.004)	(0.004)	(0.004)
2.	Experience	0.084	0.085	0.085	0.085	0.085
		(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
3.	Experience-Squared /100	-0.224	-0.229	-0.230	-0.226	-0.229
		(0.032)	(0.032)	(0.032)	(0.032)	(0.032)
4.	Black Indicator	-0.190	-0.199	-0.194	-0.194	-0.189
		(0.017)	(0.018)	(0.019)	(0.019)	(0.019)
5.	Live in South	-0.125	-0.148	-0.146	-0.145	-0.146
		(0.015)	(0.026)	(0.026)	(0.026)	(0.026)
6.	Live in SMSA	0.161	0.136	0.136	0.137	0.138
		(0.015)	(0.020)	(0.020)	(0.020)	(0.020)
7.	Region in 1966 (8 indicators)	no	yes	yes	yes	yes
8.	Live in SMSA in 1966	no	yes	yes	yes	yes

College proximity research design



Note: prediction equation is fit to subsample with no college nearby

Simple IV result

Table 3: Reduced Form and Structural Estimates of Education and Earnings Models

	Reduced Form Models: Education		Reduced Form Models: Earnings		Structural Models of Earnings	
	(1)	(2)	(3)	(4)	(5)	(6)
A: Treat Experience and Experience Squared as Exogenous						
1. Live Near College in 1966	0.320 (0.088)	0.322 (0.083)	0.042 (0.018)	0.045 (0.018)	--	--
2. Education	--	--	--	--	0.132 (0.055)	0.140 (0.055)
3. Family Background Variables[a]	no	yes	no	yes	no	yes

Overidentification and IV diagnostics in R

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.3396868  0.8945378   3.733 0.000192 ***
educ         0.1570594  0.0525782   2.987 0.002839 **
exper        0.1188149  0.0228061   5.210 2.02e-07 ***
expersq     -0.0023565  0.0003475  -6.781 1.43e-11 ***
black       -0.1232778  0.0521500  -2.364 0.018147 *
south       -0.1431945  0.0284448  -5.034 5.08e-07 ***
smsa         0.1007530  0.0315193   3.197 0.001405 **
reg661      -0.1029760  0.0434224  -2.371 0.017779 *
reg662      -0.0002286  0.0337943  -0.007 0.994602
reg663       0.0469556  0.0326490   1.438 0.150484
reg664      -0.0554084  0.0391828  -1.414 0.157437
reg665       0.0515041  0.0475678   1.083 0.279005
reg666       0.0699968  0.0533049   1.313 0.189237
reg667       0.0390596  0.0497499   0.785 0.432446
reg668      -0.1980371  0.0525350  -3.770 0.000167 ***
smsa66       0.0150626  0.0223360   0.674 0.500132

Diagnostic tests:
                 df1  df2 statistic  p-value
Weak instruments   2 2993     7.893 0.000381 ***
Wu-Hausman         1 2993     2.926 0.087286 .
Sargan             1   NA     1.248 0.263905
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```