

Applied Econometrics:

Causal Inference and Research Design,

Professor Scott Cunningham, Spring 2017

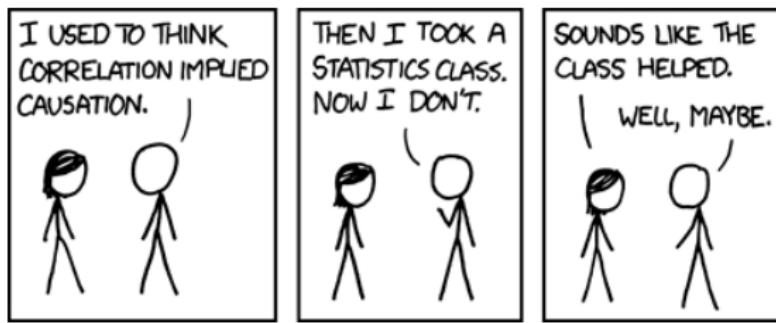


Figure: xkcd

Basic information

- Professor Scott Cunningham's contact information:
 - scunning@gmail.com or scott_cunningham@baylor.edu
 - phone: (254) 537-2239
- Office hours: MW: 11:00am to 12:15am in my office

Structure and Assessment

The fundamental theme linking all lectures will be the estimation of *causal effects*

- Part 1 (8 lectures) covers “the core” of applied econometrics
- Part 2 (10 lectures) covers “selection on observables” methodologies
- Part 3 (10 lectures) covers “selection on unobservables” methodologies

Assessment

Your grade will be based on three things:

- 4 problem sets (includes replications) *weight: 10 percent each.*
 - You will need to purchase a copy of STATA (grad plan) or use the school's licensed version on the network as the problem sets require programming.
- 1 presentation *weight: 20 percent*
- 2 exams *weight: 20 percent each.* Final is cumulative. No make up exams.

Required readings

- Textbooks
 - ① Angrist and Pischke (2009) *Mostly Harmless Econometrics* (MHE)
 - ② Morgan and Winship (2014) *Counterfactuals and Causal Inference* (MW)
- Readings:
 - We will also discuss a number of papers in each lecture, each of which you will need to learn inside and out.
 - Lecture slides and reading lists are available
 - The exam will cover material from all required readings.

Content of Part 1: The Core

- Part 1 includes 2 lectures on each of the following topics:
 - Statistics and Econometrics Review
 - Expectation operator, mean, covariance, variance
 - Ordinary Least Squares (derivation, properties, intuition, etc.)
 - Conditional expectation function and linear regression
 - Modern theories of causality
 - Rubin causal model
 - Pearl's directed acyclic graphical (DAG) model
 - Randomized experiments

Content of Part 2: Selection on observables

- Part 2 includes lectures on each of the following topics
 - 1 Regression
 - 2 Stratification
 - 3 Matching
 - 4 Regression discontinuity design

Content of Part 3: Selection on unobservables

- Part 3 includes lectures on the following topics
 - ① Panel methods, differences-in-differences and synthetic control
 - ② Instrumental variables

Introduction: OLS Review

- Derivation of the OLS estimator
- Algebraic properties of OLS
- Statistical Properties of OLS
- Variance of OLS and standard errors

Terminology

y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor
LHS	RHS

The terms “explained” and “explanatory” are probably best, as they are the most descriptive and widely applicable. But “dependent” and “independent” are used often. (The “independence” here is not really statistical independence.)

We said we must confront three issues:

- ① How do we allow factors other than x to affect y ?
- ② What is the functional relationship between y and x ?
- ③ How can we be sure we are capturing a *ceteris paribus* relationship between y and x ?

We will argue that the simple regression model

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

addresses each of them.

- The SLR model is a population model. When it comes to *estimating* β_1 (and β_0) using a random sample of data, we must restrict how u and x are related to each other.
- What we must do is restrict the way u and x relate to each other in the population.

First, we make a simplifying assumption (without loss of generality): the average, or expected, value of u is zero in the population:

$$E(u) = 0 \tag{2}$$

where $E(\cdot)$ is the expected value operator.

The presence of β_0 in

$$y = \beta_0 + \beta_1 x + u \quad (3)$$

allows us to assume $E(u) = 0$. If the average of u is different from zero, say α_0 , we just adjust the intercept, leaving the slope the same:

$$y = (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0) \quad (4)$$

where $\alpha_0 = E(u)$. The new error is $u - \alpha_0$ and the new intercept is $\beta_0 + \alpha_0$. The important point is that the slope, β_1 , has not changed.

An assumption that meshes well with our introductory treatment involves the mean of the error term for each “slice” of the population determined by values of x :

$$E(u|x) = E(u), \text{ all values } x \quad (5)$$

where $E(u|x)$ means “the expected value of u given x ”.
Then, we say u is **mean independent** of x .

- Suppose u is “ability” and x is years of education. We need, for example,

$$E(\text{ability}|x = 8) = E(\text{ability}|x = 12) = E(\text{ability}|x = 16)$$

so that the average ability is the same in the different portions of the population with an 8th grade education, a 12th grade education, and a four-year college education.

- Because people choose education levels partly based on ability, this assumption is almost certainly false.
- Suppose u is “land quality” and x is fertilizer amount. Then $E(u|x) = E(u)$ if fertilizer amounts are chosen independently of quality.

Combining $E(u|x) = E(u)$ (the substantive assumption) with $E(u) = 0$ (a normalization) gives

$$E(u|x) = 0, \text{ all values } x \quad (6)$$

Called the **zero conditional mean assumption**. Because the conditional expected value is a linear operator, $E(u|x) = 0$ implies

$$E(y|x) = \beta_0 + \beta_1 x \quad (7)$$

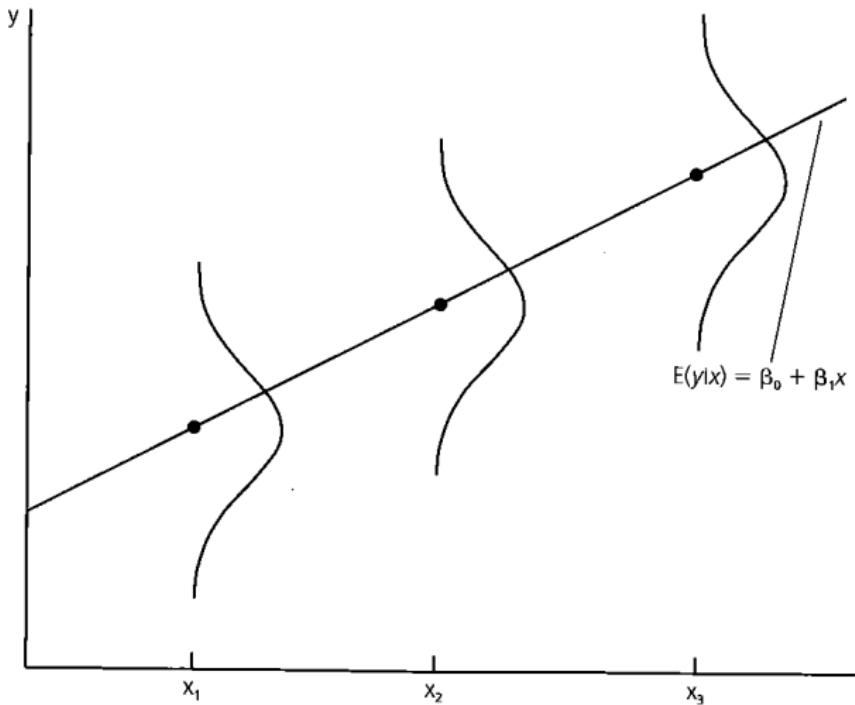
which shows the **population regression function** is a linear function of x .

- The straight line in the graph on the next page is what Wooldridge calls the **population regression function**, and what Angrist and Pischke call the **conditional expectation function**

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

- The conditional distribution of y at three different values of x are superimposed. for a given value of x , we see a range of y values: remember, $y = \beta_0 + \beta_1 x + u$, and u has a distribution in the population.

$E(y|x)$ as a linear function of x .



Deriving the Ordinary Least Squares Estimates

- Given data on x and y , how can we estimate the population parameters, β_0 and β_1 ?
- Let $\{(x_i, y_i) : i = 1, 2, \dots, n\}$ be a **random** sample of size n (the number of observations) from the population.
- Plug any observation into the population equation:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (8)$$

where the i subscript indicates a particular observation.

- We observe y_i and x_i , but not u_i (but we know it is there).

We use the two population restrictions:

$$E(u) = 0 \quad (9)$$

$$\text{Cov}(x, u) = 0 \quad (10)$$

to obtain estimating equations for β_0 and β_1 . We talked about the first condition. The second condition means that x and u are uncorrelated. Both conditions are implied by

$$E(u|x) = 0 \quad (11)$$

With $E(u) = 0$, $\text{Cov}(x, u) = 0$ is the same as $E(xu) = 0$. Next we plug in for u :

$$E(y - \beta_0 - \beta_1 x) = 0 \quad (12)$$

$$E[x(y - \beta_0 - \beta_1 x)] = 0 \quad (13)$$

These are the two conditions in the **population** that effectively determine β_0 and β_1 .

So we use their sample counterparts (which is a method of moments approach to estimation):

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (14)$$

$$n^{-1} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (15)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates from the data.

These are two linear equations in the two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$.

Pass the summation operator through the first equation:

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \quad (16)$$

$$= n^{-1} \sum_{i=1}^n y_i - n^{-1} \sum_{i=1}^n \hat{\beta}_0 - n^{-1} \sum_{i=1}^n \hat{\beta}_1 x_i \quad (17)$$

$$= n^{-1} \sum_{i=1}^n y_i - \hat{\beta}_0 - \hat{\beta}_1 \left(n^{-1} \sum_{i=1}^n x_i \right) \quad (18)$$

$$= \bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x} \quad (19)$$

We use the standard notation $\bar{y} = n^{-1} \sum_{i=1}^n y_i$ for the average of the n numbers $\{y_i : i = 1, 2, \dots, n\}$. For emphasis, we call \bar{y} a **sample average**.

We have shown that the first equation,

$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (20)$$

implies

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad (21)$$

Now, use this equation to write the intercept in terms of the slope:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (22)$$

Plug this into the second equation (but where we take away the division by n):

$$\sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (23)$$

so

$$\sum_{i=1}^n x_i[y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i] = 0 \quad (24)$$

Simple algebra gives

$$\sum_{i=1}^n x_i(y_i - \bar{y}) = \hat{\beta}_1 \left[\sum_{i=1}^n x_i(x_i - \bar{x}) \right] \quad (25)$$

So, the equation to solve is

$$\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right] \quad (26)$$

If $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$, we can write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x_i, y_i)}{\text{Sample Variance}(x_i)} \quad (27)$$

- The previous formula for $\hat{\beta}_1$ is important. It shows us how to take the data we have and compute the slope estimate.
- $\hat{\beta}_1$ is called the **ordinary least squares (OLS)** slope estimate.
- It can be computed whenever the sample variance of the x_i is not zero, which only rules out the case where each x_i has the same value.
- The intuition is that the variation in x is what permits us to identify its impact on y .

- Once we have $\hat{\beta}_1$, we compute $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. This is the OLS intercept estimate.
- These days, we let the computer do the calculations, which are tedious even if n is small.

- For any candidates $\hat{\beta}_0$ and $\hat{\beta}_1$, define a **fitted value** for each i as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (28)$$

We have n of these. This is the value we predict for y_i given that $x = x_i$.

- The “mistake” we make is the **residual**:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (29)$$

Suppose we measure the size of the mistake, for each i , by squaring it. Then we add them all up:

$$\sum_{i=1}^n \hat{u}_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (30)$$

This is called the **sum of squared residuals**. Finally, choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to *minimize* the sum of squared residuals. Using calculus (or other arguments), it can be shown that the solutions are the same we obtained before.

Algebraic Properties of OLS Statistics

Remember how we obtained $\hat{\beta}_0$ and $\hat{\beta}_1$. When an intercept is included, we will have:

$$\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (31)$$

The OLS residuals *always* add up to zero, by *construction*,

$$\sum_{i=1}^n \hat{u}_i = 0 \quad (32)$$

Because $y_i = \hat{y}_i + \hat{u}_i$ by definition,

$$n^{-1} \sum_{i=1}^n y_i = n^{-1} \sum_{i=1}^n \hat{y}_i + n^{-1} \sum_{i=1}^n \hat{u}_i \quad (33)$$

and so $\bar{y} = \bar{\hat{y}}$.

Similarly the way we obtained our estimates,

$$n^{-1} \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (34)$$

The sample covariance (and therefore the sample correlation) between the explanatory variables and the residuals is always zero:

$$n^{-1} \sum_{i=1}^n x_i \hat{u}_i = 0 \quad (35)$$

Because the \hat{y}_i are linear functions of the x_i , the fitted values and residuals are uncorrelated, too:

$$n^{-1} \sum_{i=1}^n \hat{y}_i \hat{u}_i = 0 \quad (36)$$

Both properties hold by construction. $\hat{\beta}_0$ and $\hat{\beta}_1$ were chosen to make them true.

A third property is that the point (\bar{x}, \bar{y}) is always on the OLS regression line. That is, if we plug in the average for x , we predict the sample average for y :

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad (37)$$

Again, we chose the estimates to make this true.

Expected Value of OLS

- Mathematical statistics: How do our estimators behave across different samples of data? On average, would we get the right answer if we could repeatedly sample?
- We need to find the expected value of the OLS estimators – in effect, the average outcome across all possible random samples – and determine if we are right on average.
- Leads to the notion of **unbiasedness**, which is a “desirable” characteristic for estimators.

$$E(\hat{\beta}) = \beta \tag{38}$$

- Remember, our objective is to estimate β_1 , the slope **population** parameter that describes the relationship between y and x .
- $\hat{\beta}_1$ is an **estimator** of that parameter obtained for a *specific* sample.
- Different samples will generate different estimates ($\hat{\beta}_1$) for the “true” β_1 , i.e. ($\hat{\beta}_1$) is a random variable.
- Unbiasedness is the idea that if we could take as many random samples on Y as we want from the population, and compute an estimate each time, the average of these estimates would be equal to β_1 .

Assumption SLR.1 (Linear in Parameters)

- The population model can be written as

$$y = \beta_0 + \beta_1 x + u \quad (39)$$

where β_0 and β_1 are the (unknown) population parameters.

- We view x and u as outcomes of random variables; thus, y is random.
- Stating this assumption formally shows that our goal is to estimate β_0 and β_1 .

Assumption SLR.2 (Random Sampling)

- We have a random sample of size n , $\{(x_i, y_i) : i = 1, \dots, n\}$, following the population model.
- We know how to use this data to estimate β_0 and β_1 by OLS.
- Because each i is a draw from the population, we can write, for each i ,

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (40)$$

- Notice that u_i here is the unobserved error for observation i . It is not the residual that we compute from the data!

Assumption SLR.3 (Sample Variation in the Explanatory Variable)

- The sample outcomes on x_i are not all the same value.
- This is the same as saying the sample variance of $\{x_i : i = 1, \dots, n\}$ is not zero.
- In practice, this is no assumption at all. If the x_i are all the same value, we cannot learn how x affects y in the population.

Assumption SLR.4 (Zero Conditional Mean)

- In the population, the error term has zero mean given any value of the explanatory variable:

$$E(u|x) = E(u) = 0. \quad (41)$$

- This is the key assumption for showing that OLS is unbiased, with the zero value not being important once we assume $E(u|x)$ does not change with x .
- Note that we can compute the OLS estimates whether or not this assumption holds, or even if there is an underlying population model.

Showing OLS is unbiased

How do we show $\hat{\beta}_1$ is unbiased for β_1 ? What we need to show is

$$E(\hat{\beta}_1) = \beta_1 \tag{42}$$

where the expected value means averaging across random samples.

Step 1: Write down a formula for $\hat{\beta}_1$. It is convenient to use

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \tag{43}$$

which is one of several equivalent forms.

It is convenient to define $SST_x = \sum_{i=1}^n (x_i - \bar{x})^2$, to total variation in the x_i , and write

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{SST_x} \quad (44)$$

Remember, SST_x is just some positive number. The existence of $\hat{\beta}_1$ is guaranteed by SLR.3.

Step 2: Replace each y_i with $y_i = \beta_0 + \beta_1 x_i + u_i$ (which uses SLR.1 and the fact that we have data from SLR.2).

The numerator becomes

$$\sum_{i=1}^n (x_i - \bar{x}) y_i = \sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i) \quad (45)$$

$$= \beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x}) x_i + \sum_{i=1}^n (x_i - \bar{x}) u_i \quad (46)$$

$$= 0 + \beta_1 \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (x_i - \bar{x}) u_i \quad (47)$$

$$= \beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x}) u_i \quad (48)$$

We used $\sum_{i=1}^n (x_i - \bar{x}) = 0$ and $\sum_{i=1}^n (x_i - \bar{x}) x_i = \sum_{i=1}^n (x_i - \bar{x})^2$.

We have shown

$$\hat{\beta}_1 = \frac{\beta_1 SST_x + \sum_{i=1}^n (x_i - \bar{x}) u_i}{SST_x} = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) u_i}{SST_x} \quad (49)$$

Note how the last piece is the slope coefficient from the OLS regression of u_i on x_i , $i = 1, \dots, n$. We cannot do this regression because the u_i are not observed.

Now define

$$w_i = \frac{(x_i - \bar{x})}{SST_x} \quad (50)$$

so we have

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i \quad (51)$$

- $\hat{\beta}_1$ is a linear function of the unobserved errors, u_i . The w_i are all functions of $\{x_1, x_2, \dots, x_n\}$.
- The (random) difference between $\hat{\beta}_1$ and β_1 is due to this linear function of the unobservables.

Step 3: Find $E(\hat{\beta}_1)$.

- Under Assumptions SLR.2 and SLR.4, $E(u_i|x_1, x_2, \dots, x_n) = 0$.
That means, *conditional* on $\{x_1, x_2, \dots, x_n\}$,

$$E(w_i u_i|x_1, x_2, \dots, x_n) = w_i E(u_i|x_1, x_2, \dots, x_n) = 0$$

because w_i is a function of $\{x_1, x_2, \dots, x_n\}$. (In the next slides I omit the conditioning in the expectations)

- This would not be true if, in the population, u and x are correlated.

Now we can complete the proof: conditional on $\{x_1, x_2, \dots, x_n\}$,

$$E(\hat{\beta}_1) = E\left(\beta_1 + \sum_{i=1}^n w_i u_i\right) \quad (52)$$

$$= \beta_1 + \sum_{i=1}^n E(w_i u_i) = \beta_1 + \sum_{i=1}^n w_i E(u_i) \quad (53)$$

$$= \beta_1 \quad (54)$$

Remember, β_1 is the fixed constant in the population. The estimator, $\hat{\beta}_1$, varies across samples and is the random outcome: before we collect our data, we do not know what $\hat{\beta}_1$ will be.

THEOREM (Unbiasedness of OLS)

Under Assumptions SLR.1 through SLR.4

$$E(\hat{\beta}_0) = \beta_0 \text{ and } E(\hat{\beta}_1) = \beta_1. \quad (55)$$

- Omit the proof for $\hat{\beta}_0$.

- Each sample leads to a different estimate, $\hat{\beta}_0$ and $\hat{\beta}_1$. Some will be very close to the true values $\beta_0 = 3$ and $\beta_1 = 2$. Nevertheless, some could be very far from those values.
- If we repeat the experiment again and again, and average the estimates, we would get very close to 2.
- The problem is, we do not know which kind of sample we have. We can never know whether we are close to the population value.
- We hope that our sample is "typical" and produces a slope estimate close to β_1 but we can never know.

Reminder

- **Errors** are the vertical distances between observations and the **unknown** Conditional Expectation Function. Therefore, they are unknown.
- **Residuals** are the vertical distances between observations and the **estimated** regression function. Therefore, they are known.

SE and the data

The correct SE estimation procedure is given by the underlying structure of the data

- It is very unlikely that all observations in a dataset are unrelated, but drawn from identical distributions (**homoskedasticity**)
- For instance, the variance of income is often greater in families belonging to top deciles than among poorer families (**heteroskedasticity**)
- Some phenomena do not affect observations individually, but they do affect groups of observations uniformly within each group (**clustered data**)

Variance of the OLS Estimators

- Under SLR.1 to SLR.4, the OLS estimators are unbiased. This tells us that, on average, the estimates will equal the population values.
- But we need a measure of dispersion (spread) in the sampling distribution of the estimators. We use the variance (and, ultimately, the standard deviation).
- We could characterize the variance of the OLS estimators under SLR.1 to SLR.4 (and we will later). For now, it is easiest to introduce an assumption that simplifies the calculations.

Assumption SLR.5 (Homoskedasticity, or Constant Variance)

The error has the same variance given any value of the explanatory variable x :

$$\text{Var}(u|x) = \sigma^2 > 0 \quad (56)$$

where σ^2 is (virtually always) unknown.

Because we assume SLR.4, that is, $E(u|x) = 0$ whenever we assume SLR.5, we can also write

$$E(u^2|x) = \sigma^2 = E(u^2) \quad (57)$$

Under the population Assumptions SLR.1 ($y = \beta_0 + \beta_1 x + u$),
SRL.4 ($E(u|x) = 0$) and SLR.5 ($Var(u|x) = \sigma^2$),

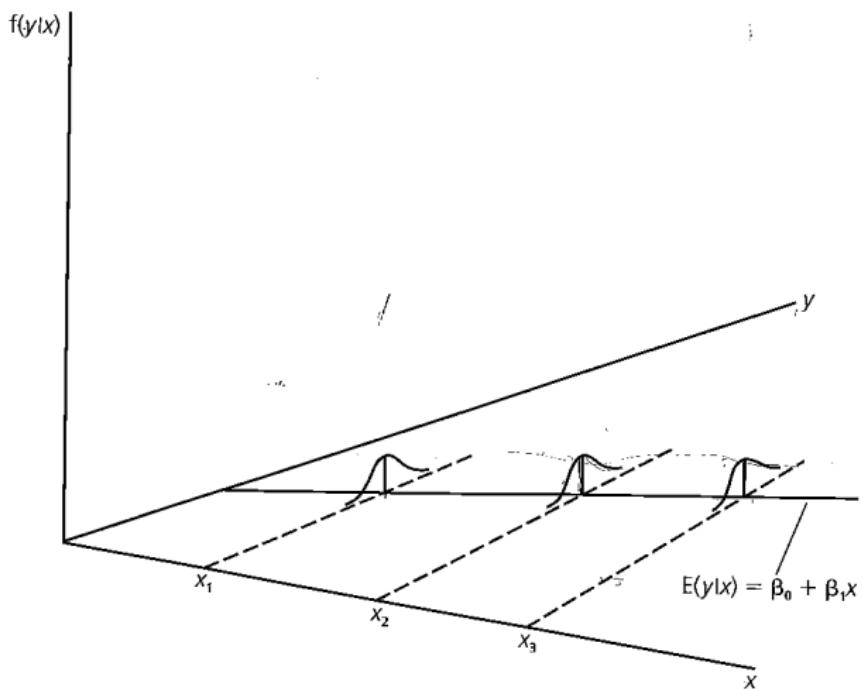
$$E(y|x) = \beta_0 + \beta_1 x \quad (58)$$

$$Var(y|x) = \sigma^2 \quad (59)$$

So the average or expected value of y is allowed to change with x – in fact, this is what interests us – but the variance does not change with x . (See Graphs on next two slides)

Figure 2.8

The simple regression model under homoskedasticity.



THEOREM (Sampling Variances of OLS)

Under Assumptions SLR.1 to SLR.2,

$$Var(\hat{\beta}_1|x) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x} \quad (60)$$

$$Var(\hat{\beta}_0|x) = \frac{\sigma^2 (n^{-1} \sum_{i=1}^n x_i^2)}{SST_x} \quad (61)$$

(conditional on the outcomes $\{x_1, x_2, \dots, x_n\}$).

To show this, write, as before,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n w_i u_i \quad (62)$$

where $w_i = (x_i - \bar{x})/SST_x$. We are treating this as nonrandom in the derivation. Because β_1 is a constant, it does not affect $Var(\hat{\beta}_1)$. Now, we need to use the fact that, for uncorrelated random variables, the variance of the sum is the sum of the variances.

The $\{u_i : i = 1, 2, \dots, n\}$ are actually independent across i , and so they are uncorrelated. So (remember that if we know x , we know w)

$$Var(\hat{\beta}_1|x) = Var\left(\sum_{i=1}^n w_i u_i|x\right) \quad (63)$$

$$= \sum_{i=1}^n Var(w_i u_i|x) = \sum_{i=1}^n w_i^2 Var(u_i|x) \quad (64)$$

$$= \sum_{i=1}^n w_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n w_i^2 \quad (65)$$

where the second-to-last equality uses Assumption SLR.5, so that the variance of u_i does not depend on x_i .

Now we have

$$\sum_{i=1}^n w_i^2 = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{(SST_x)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{(SST_x)^2} \quad (66)$$

$$= \frac{SST_x}{(SST_x)^2} = \frac{1}{SST_x} \quad (67)$$

We have shown

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \quad (68)$$

Usually we are interested in β_1 . We can easily study the two factors that affect its variance.

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \quad (69)$$

- ① As the error variance increases, i.e., as σ^2 increases, so does $Var(\hat{\beta}_1)$. The more “noise” in the relationship between y and x – that is, the larger variability in u – the harder it is to learn about β_1 .
- ② By contrast, more variation in $\{x_i\}$ is a *good* thing:

$$SST_x \uparrow \text{ implies } Var(\hat{\beta}_1) \downarrow \quad (70)$$

Notice that SST_x/n is the sample variance in x . We can think of this as getting close to the population variance of x , σ_x^2 , as n gets large. This means

$$SST_x \approx n\sigma_x^2 \tag{71}$$

which means, as n grows, $Var(\hat{\beta}_1)$ shrinks at the rate $1/n$. This is why more data is a good thing: it shrinks the sampling variance of our estimators.

The standard deviation of $\hat{\beta}_1$ is the square root of the variance. So

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}} \quad (72)$$

This turns out to be the measure of variation that appears in confidence intervals and test statistics.

Estimating the Error Variance

In the formula

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_x} \quad (73)$$

we can compute SST_x from $\{x_i : i = 1, \dots, n\}$. But we need to estimate σ^2 .

Recall that

$$\sigma^2 = E(u^2). \quad (74)$$

Therefore, if we could observe a sample on the errors, $\{u_i : i = 1, 2, \dots, n\}$, an unbiased estimator of σ^2 would be the sample average

$$n^{-1} \sum_{i=1}^n u_i^2 \quad (75)$$

But this is not an estimator because we cannot compute it from the data we observe, since u_i are unobserved.

How about replacing each u_i with its “estimate”, the OLS residual \hat{u}_i ?

$$u_i = y_i - \beta_0 - \beta_1 x_i \quad (76)$$

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (77)$$

\hat{u}_i can be computed from the data because it depends on the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$. Except by fluke,

$$\hat{u}_i \neq u_i \quad (78)$$

for any i .

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (79)$$

$$= u_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1) x_i \quad (80)$$

$E(\hat{\beta}_0) = \beta_0$ and $E(\hat{\beta}_1) = \beta_1$, but the estimators almost always differ from the population values in a sample.

Now, what about this as an estimator of σ^2 ?

$$n^{-1} \sum_{i=1}^n \hat{u}_i^2 = SSR/n \quad (81)$$

It is a true estimator and easily computed from the data after OLS. As it turns out, this estimator is slightly biased: its expected value is a little less than σ^2 .

The estimator does not account for the two restrictions on the residuals, used to obtain $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\sum_{i=1}^n \hat{u}_i = 0 \tag{82}$$

$$\sum_{i=1}^n x_i \hat{u}_i = 0 \tag{83}$$

There is no such restriction on the unobserved errors.

The unbiased estimator of σ^2 uses a **degrees-of-freedom** adjustment. The residuals have only $n - 2$ degrees-of-freedom, not n .

$$\hat{\sigma}^2 = \frac{SSR}{(n - 2)} \quad (84)$$

THEOREM: Unbiased Estimator of σ^2

Under Assumptions SLR.1 to SLR.5,

$$E(\hat{\sigma}^2) = \sigma^2 \quad (85)$$

In regression output, it is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{SSR}{(n-2)}} \quad (86)$$

that is usually reported. This is an estimator of $sd(u)$, the standard deviation of the population error. And $SSR = \sum_{i=1}^n \hat{u}^2$.

- $\hat{\sigma}$ is called the **standard error of the regression**, which means it is an estimate of the standard deviation of the error in the regression. Stata calls it the **root mean squared error**.
- Given $\hat{\sigma}$, we can now estimate $sd(\hat{\beta}_1)$ and $sd(\hat{\beta}_0)$. The estimates of these are called the **standard errors** of the $\hat{\beta}_j$.

- We just plug $\hat{\sigma}$ in for σ :

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{SST_x}} \quad (87)$$

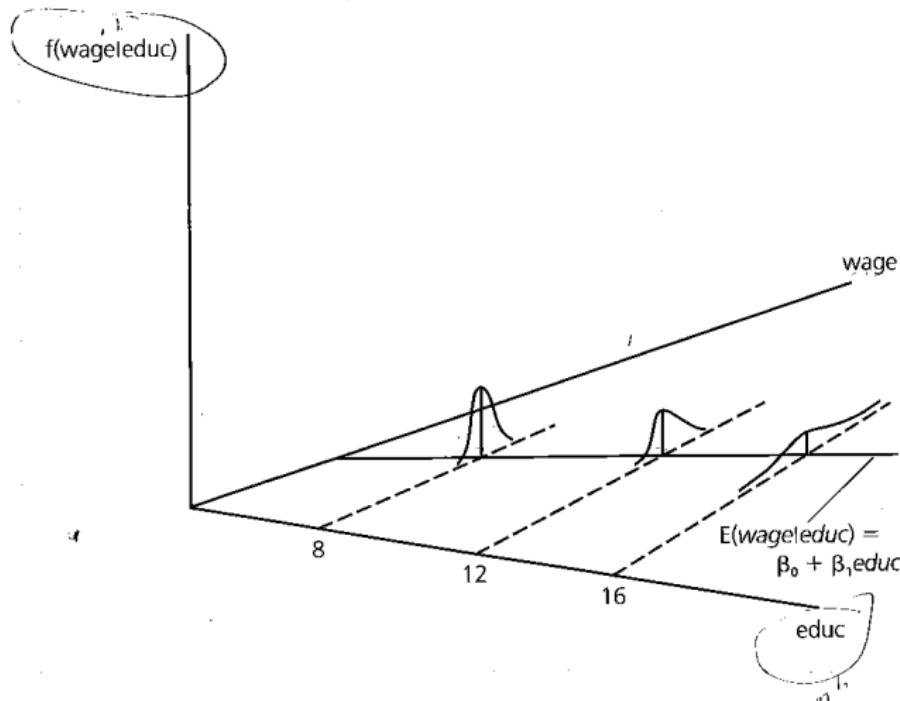
where both the numerator and denominator are computed from the data.

- For reasons we will see, it is useful to report the standard errors below the corresponding coefficient, usually in parentheses.

- OLS inference is generally faulty in the presence of heteroskedasticity

Figure 2.9

Var (wage|educ) increasing with educ.



- Fortunately, OLS is still useful
- Assume SLR.1-4 hold, but not SLR.5. Therefore

$$Var(u_i|x_i) = \sigma_i^2$$

- The variance of our estimator, $\hat{\beta}_1$ equals:

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$$

- When $\sigma_i^2 = \sigma^2$ for all i , this formula reduces to the usual form, $\frac{\sigma^2}{SST_x^2}$

- A valid estimator of $\text{Var}(\hat{\beta}_1)$ for heteroskedasticity of any form (including homoskedasticity) is

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \hat{u}_i^2}{SST_x^2}$$

which is easily computed from the data after the OLS regression

- As a rule, you should always use the `, robust` command in STATA.

Clustered data

- Observations are related to each other within certain groups
 - You want to regress kids' grades on class size to determine the effect of class size on grades
 - The **unobservables** of kids belonging to the same classroom will be correlated (e.g., teacher quality, recess routines) while will not be correlated with kids in far away classrooms
- This means we are assuming independence across “clusters” but correlation within clusters.

Cluster robust standard errors

- Using cluster-robust standard errors changes somewhat
- Instead of summing over each individual, we first sum over groups

Clustered data

- Let's stack the observations and use matrix notation. Stack observations by cluster

$$y_g = x_g \beta + u_g$$

- The OLS estimator of β is:

$$\hat{\beta} = [X'X]^{-1}X'y$$

we just stacked the data is all

- The variance is given by:

$$Var(\beta) = E[[X'X]^{-1}X'\Omega X[X'X]^{-1}]$$

Clustered data

With this in mind, we can now write the variance-covariance matrix for clustered data

$$Var(\hat{\beta}) = [X'X]^{-1} \left[\sum_{i=1}^G x_g' \hat{u}_g \hat{u}_g' x_g \right] [X'X]^{-1}$$

- In STATA: `vce(cluster clustervar)`. Where `clustervar` is a variable that identifies the groups in which unobservables are allowed to correlate

The importance of knowing your data

- In real world you should never go with the “independent and identically distributed” (i.e., homoskedasticity) case. Life is not that simple.
- You need to know your data in order to choose the correct error structure and then infer the required SE calculation
- If you have aggregate variables, like class size, clustering at that level is *required*

Outline: CEF and Linear regression

- Properties of the conditional expectation function (CEF)
- Reasons for using linear regression
- Regression anatomy theorem
- Omitted variable bias

Properties of the conditional expectation function

- Now we relax the assumption that x is fixed across samples and instead assume it is a random variable
- Assume we are interested in the returns to schooling in a wage regression. We can summarize the predictive power of schooling's effect on wages with the **conditional expectation function**

$$E(y_i|x_i) \tag{88}$$

- The CEF for a dependent variable, y_i , given covariates X_i , is the expectation, or population average, of y_i with x_i held constant.

- $E(y_i|x_i)$ gives the expected value of y for given values of x
 - It provides a reasonable representation of how y changes with x
 - If x is random, then $E(y_i|x_i)$ is a random function

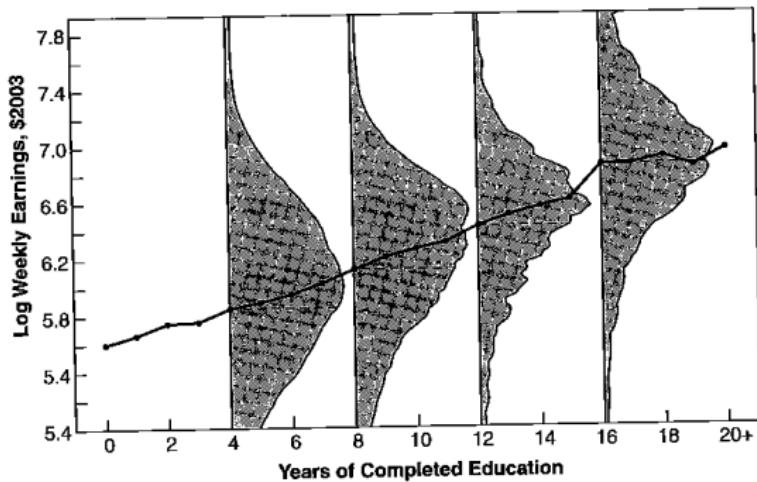


Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

- We write the CEF as $E(y_i|x_i)$ and note that it's explicitly a function of x_i , and because x_i is a random variable, so is the CEF
- We're often interested in CEFs that are functions of many variables, conveniently subsumed in the vector x_i , and for a specific value of x_i , we will write

$$E(y_i|x_i = x)$$

Helpful result: Law of Iterated Expectations

One result we will use in the following theorems is called **law of iterated expectations**. It's fairly intuitive.

Definition of Law of Iterated Expectations (LIE)

The unconditional expectation of a random variable is equal to the expectation of the conditional expectation of the random variable conditional on some other random variable

$$E(Y) = E(E[Y|X])$$

- Say that the population is divided by gender. We could take conditional expectations by gender and combine them (properly weighted) to get the unconditional expectation

$$\begin{aligned}
 E[IQ] &= E(E[IQ|Sex]) \\
 &= \sum_{Sex_i} Pr(Sex_i) \cdot E[IQ|Sex_i] \\
 &= Pr(Male) \cdot E[IQ|Male] \\
 &\quad + Pr(Female) \cdot E[IQ|Female]
 \end{aligned}$$

- In other words, the average of the conditional averages is the unconditional average.

Proof.

For the continuous case:

$$\begin{aligned} E[E(Y|X)] &= \int E(Y|X=u)g_x(u)du \\ &= \int \left[\int tf_{y|x}(t|X=u)dt \right] g_x(u)du \\ &= \int \int tf_{y|x}(t|X=u)g_x(u)dudt \\ &= \int t \left[\int f_{y|x}(t|X=u)g_x(u)du \right] dt \\ &= \int t [f_{x,y}du] dt \\ &= \int tg_y(t)dt \\ &= E(y) \end{aligned}$$



Proof.

For the discrete case,

$$\begin{aligned} E(E[Y|X]) &= \sum_x E[Y|X=x]p(x) \\ &= \sum_x \left(\sum_y y p(y|x) \right) p(x) \\ &= \sum_x \sum_y y p(x,y) \\ &= \sum_y y \sum_x p(x,y) \\ &= \sum_y y p(y) \\ &= E(Y) \end{aligned}$$



Property 1: CEF Decomposition Property

Theorem 3.1.1: The CEF Decomposition Property

$$y_i = E(y_i|x_i) + u_i$$

where

- ① u_i is mean independent of x_i ; that is

$$E(u_i|x_i) = 0$$

- ② u_i is uncorrelated with any function of x_i

In words: The theorem says that any random variable, y_i , can be decomposed into two parts: the part that can be explained by x_i and the part left over that can't be which is by definition unexplained by x_i . Proof is in Angrist and Pischke (ch. 3)

Property 2: CEF Prediction Property

Theorem 3.1.2: The CEF Prediction Property

Let $m(x_i)$ be any function of x_i . The CEF solves

$$E(y_i|x_i) = \arg \min_{m(x_i)} E[(y_i - m(x_i))^2].$$

In words: The CEF is the minimum mean squared error predictor of y_i given x_i .

Proof is in Angrist and Pischke (ch. 3)

Property 3: ANOVA Theorem

Theorem 3.1.3: The ANOVA (Analysis of variance) Theorem

The final property of the CEF is the ANOVA theorem which states

$$V(y_i) = V[E(y_i|x_i)] + E[V(y_i|x_i)]$$

where $V(\cdot)$ is the variance and $V(y_i|x_i)$ is the conditional variance of y_i given x_i .

Proof is in Angrist and Pischke (ch. 3).

3 reasons why linear regression may be of interest

Angrist and Pischke argue that the linear regression may be of interest to you – even if the underlying CEF is not linear! We review some of the linear theorems now. These are merely to justify the use of linear models to approximate the CEF.

Theorem 3.1.4 - The Linear CEF Theorem

Suppose the CEF is linear. Then the population regression is it.

Comment: Trivial theorem imho because if the population CEF is linear, then of course you should linear regression to estimate it.
Duh.

Proof in Angrist and Pischke (ch. 3). Proof uses the CEF Decomposition Property from earlier.

Theorem 3.1.5 - The Best Linear Predictor Theorem

- ① The CEF, $E(y_i|x_i)$, is the minimum mean squared error (MMSE) predictor of y_i given x_i in the class of all functions x_i by the CEF prediction property
- ② The population regression function, $E(x_i y_i) E(x_i x_i')^{-1}$, is the best we can do in the class of all linear functions

Proof is in Angrist and Pischke (ch. 3).

Theorem 3.1.6: The Regression CEF Theorem

The function $x_i\beta$ provides the minimum mean squared error (MMSE) linear approximation to $E(y_i|x_i)$, that is

$$\beta = \arg \min_b E\{(E(y_i|x_i) - x_i'b)^2\}$$

Again, proof in Angrist and Pischke (ch. 3).

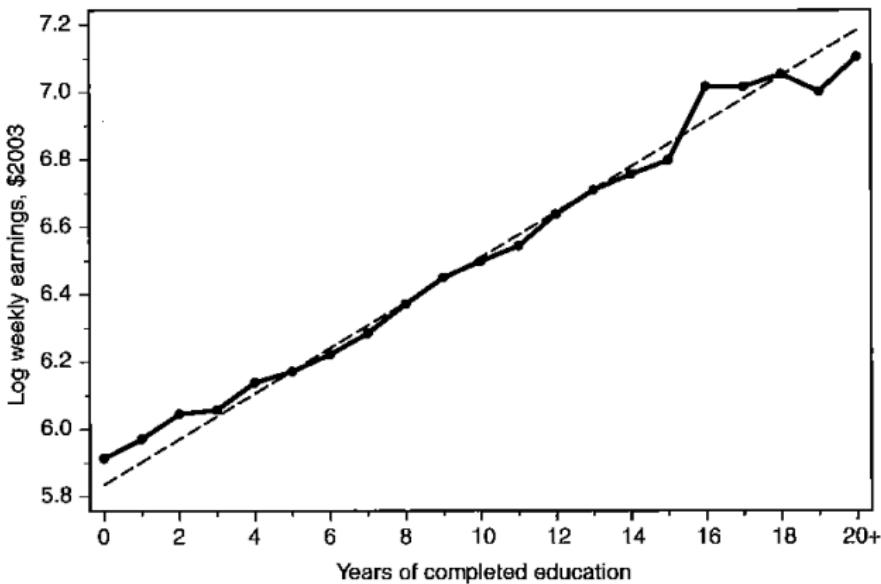


Figure 3.1.2 Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

Theorem 3.1.7: Regression anatomy theorem

- Regression anatomy theorem is maybe more intuitive with an example and some data visualization. It concerns multiple linear regression.
- Can we estimate the causal effect of family size on labor supply by regressing labor supply (`workforpay`) on family size (`numkids`)?

$$workforpay_i = \beta_0 + \beta_1 numkids_i + u_i$$

. regress workforpay numkids

where the first line is the causal / econometric model, and the second line is the regression command in STATA

- If family size is random, then number of kids is uncorrelated with the unobserved error term, which means we can interpret $\hat{\beta}_1$ as the causal effect.
 - Example: if Melissa has no children in reality (i.e., $\text{numkids} = 0$) and we wanted to know what the effect on labor supply will be if we surgically manipulated her family size (i.e., $\text{numkids} = 1$) then $\hat{\beta}_1$ would be our answer
 - Visual: Even better, we could just plot the regression coefficient in a scatter plot showing all i (workforpay , numkids) pairs and the slope coefficient would be the best fit of the data through these points, as well as tell us the average causal effect of family size on labor supply
- But how do we interpret $\hat{\beta}_1$ if numkids is non-random?

- Assume that family size is random once we condition on race, age, marital status and employment. Then the model is:

$$\text{Workforpay}_i = \beta_0 + \beta_1 \text{Numkids}_i + \gamma_1 \text{White}_i + \gamma_2 \text{Married}_i + \gamma_3 \text{Age}_i + \gamma_4 \text{Employed}_i + u_i$$

- If we want to estimate average causal effect of family size on labor supply, we will need two things:
 - 1 a data set with all 6 variables;
 - 2 numkids must be randomly assigned conditional on the other 4 variables
- Now how do we interpret $\hat{\beta}_1$? And can we visualize $\hat{\beta}_1$ when there's multiple dimensions to the data? Yes, using the regression anatomy theorem, we can.

Theorem 3.1.7: Regression Anatomy Theorem

Assume your main multiple regression model of interest:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + \epsilon_i$$

and an auxiliary regression in which the variable x_{1i} is regressed on all the remaining independent variables

$$x_{1i} = \gamma_0 + \gamma_{k-1} x_{k-1i} + \gamma_{k+1} x_{k+1i} + \cdots + \gamma_K x_{Ki} + f_i$$

and $\tilde{x}_{1i} = x_{1i} - \hat{x}_{1i}$ being the residual from the auxiliary regression.
The parameter β_1 can be rewritten as:

$$\beta_1 = \frac{\text{Cov}(y_i, \tilde{x}_{1i})}{\text{Var}(\tilde{x}_{1i})}$$

In words: The regression anatomy theorem is about interpretation.
It says that $\hat{\beta_1}$ is simply a scaled covariance with the \tilde{x}_1 residual used instead of the actual data x .

I think a more detailed proof could be helpful, so I'm leaving it in the slides for now.

Regression Anatomy Proof

To prove the theorem, note $E[\tilde{x}_{ki}] = E[x_{ki}] - E[\hat{x}_{ki}] = E[f_i]$, and plug y_i and residual \tilde{x}_{ki} from x_{ki} auxiliary regression into the covariance $\text{cov}(y_i, \tilde{x}_{ki})$

$$\begin{aligned}\beta_k &= \frac{\text{cov}(y_i, \tilde{x}_{ki})}{\text{var}(\tilde{x}_{ki})} \\ &= \frac{\text{cov}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, \tilde{x}_{ki})}{\text{var}(\tilde{x}_{ki})} \\ &= \frac{\text{cov}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i, f_i)}{\text{var}(f_i)}\end{aligned}$$

- ① Since by construction $E[f_i] = 0$, it follows that the term $\beta_0 E[f_i] = 0$.
- ② Since f_i is a linear combination of all the independent variables with the exception of x_{ki} , it must be that

$$\beta_1 E[f_i x_{1i}] = \cdots = \beta_{k-1} E[f_i x_{k-1i}] = \beta_{k+1} E[f_i x_{k+1i}] = \cdots = \beta_K E[f_i x_{Ki}] = 0$$

- ③ Consider now the term $E[e_i f_i]$. This can be written as:

$$\begin{aligned}
 E[e_i f_i] &= E[e_i f_i] \\
 &= E[e_i \tilde{x}_{ki}] \\
 &= E[e_i (x_{ki} - \hat{x}_{ki})] \\
 &= E[e_i x_{ki}] - E[e_i \tilde{x}_{ki}]
 \end{aligned}$$

Since e_i is uncorrelated with any independent variable, it is also uncorrelated with x_{ki} : accordingly, we have $E[e_i x_{ki}] = 0$. With regard to the second term of the subtraction, substituting the predicted value from the x_{ki} auxiliary regression, we get

$$E[e_i \tilde{x}_{ki}] = E[e_i (\hat{\gamma}_0 + \hat{\gamma}_1 x_{1i} + \cdots + \hat{\gamma}_{k-1} x_{(k-1)i} + \hat{\gamma}_{k+1} x_{(k+1)i} + \cdots + \hat{\gamma}_K x_{Ki})]$$

Once again, since e_i is uncorrelated with any independent variable, the expected value of the terms is equal to zero. Then, it follows $E[e_i f_i] = 0$.

Regression Anatomy Proof (cont.)

- ④ The only remaining term is $E[\beta_k x_{ki} f_i]$ which equals $E[\beta_k x_{ki} \tilde{x}_{ki}]$ since $f_i = \tilde{x}_{ki}$. The term x_{ki} can be substituted using a rewriting of the auxiliary regression model, x_{ki} , such that

$$x_{ki} = E[x_{ki}|X_{-k}] + \tilde{x}_{ki}$$

This gives

$$\begin{aligned} E[\beta_k x_{ki} \tilde{x}_{ki}] &= E[\beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})]] \\ &= \beta_k E[\tilde{x}_{ki}(E[x_{ki}|X_{-k}] + \tilde{x}_{ki})] \\ &= \beta_k \{E[\tilde{x}_{ki}^2] + E[(E[x_{ki}|X_{-k}] \tilde{x}_{ki})]\} \\ &= \beta_k \text{var}(\tilde{x}_{ki}) \end{aligned}$$

which follows directly from the orthogonality between $E[x_{ki}|X_{-k}]$ and \tilde{x}_{ki} . From previous derivations we finally get

$$\text{cov}(y_i, \tilde{x}_{ki}) = \beta_k \text{var}(\tilde{x}_{ki})$$

which completes the proof. □

STATA command: reganat (i.e., regression anatomy)

```
. ssc install reganat, replace
. sysuse auto
. regress price length weight headroom mpg
. reganat price length weight headroom mpg, dis(weight length) biline
```

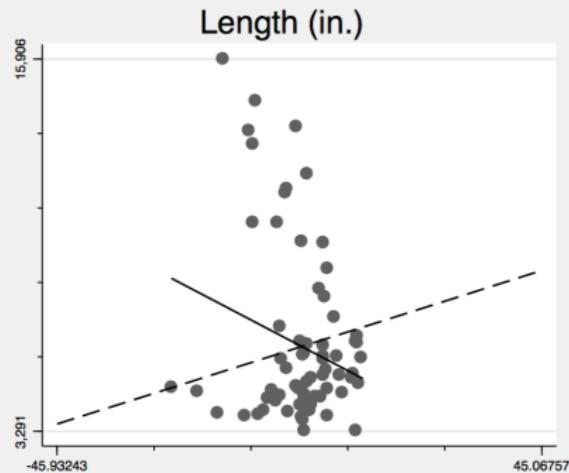
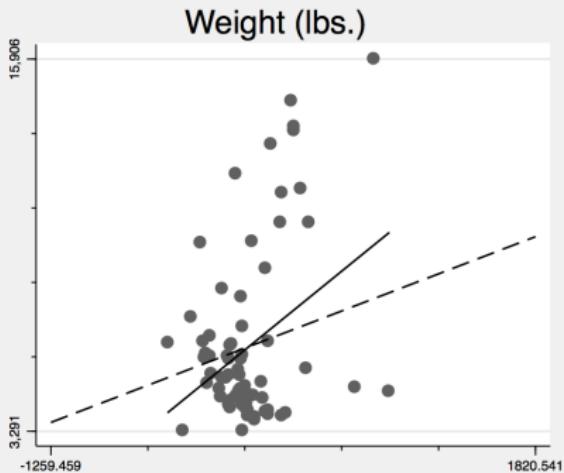
```
. regress price length weight headroom mpg
```

Source	SS	df	MS	Number of obs	=	74
Model	236190226	4	59047556.6	F(4, 69)	=	10.21
Residual	398875170	69	5780799.56	Prob > F	=	0.0000
Total	635065396	73	8699525.97	R-squared	=	0.3719

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
price						
length	-94.49651	40.39563	-2.34	0.022	-175.0836	-13.90944
weight	4.335045	1.162745	3.73	0.000	2.015432	6.654657
headroom	-490.9667	388.4892	-1.26	0.211	-1265.981	284.048
mpg	-87.95838	83.5927	-1.05	0.296	-254.7213	78.80449
_cons	14177.58	5872.766	2.41	0.018	2461.735	25893.43

Regression Anatomy

Dependent variable: Price



Covariates: Length (in.), Weight (lbs.), Headroom (in.), Mileage (mpg).

Regression lines: Solid = Multivariate, Dashed = Bivariate.

Big picture

- ① Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable
- ② If we prefer to think of approximating $E(y_i|x_i)$ as opposed to predicting y_i , the regression CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.
- ③ Regression anatomy theorem helps us interpret a single slope coefficient in a multiple regression model by the aforementioned decomposition.

Omitted Variable Bias

- A typical problem is when a key variable is omitted. Assume schooling causes earnings to rise:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 A_i + u_i$$

Y_i = log of earnings

S_i = schooling measured in years

A_i = individual ability

- Typically the econometrician cannot observe A_i ; for instance, the Current Population Survey doesn't present adult respondents' family background, intelligence, or motivation.

- What are the consequences of leaving ability out of the regression? Suppose you estimated this short regression instead:

$$Y_i = \beta_0 + \beta_1 S_i + \eta_i$$

where $\eta_i = \beta_2 A_i + u_i$; β_0 , β_1 , and β_2 are population regression coefficients; S_i is correlated with η_i through A_i only; and u_i is a regression residual uncorrelated with all regressors by definition.

Derivation of Ability Bias

- Suppressing the i subscripts, the OLS estimator for β_1 is:

$$\hat{\beta}_1 = \frac{\text{Cov}(Y, S)}{\text{Var}(S)} = \frac{E[YS] - E[Y]E[S]}{\text{Var}(S)}$$

- Plugging in the true model for Y , we get:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{Cov}[(\beta_0 + \beta_1 S + \beta_2 A + u), S]}{\text{Var}(S)} \\ &= \frac{E[(\beta_0 S + \beta_1 S^2 + \beta_2 S A + u S)] - E(S)E[\beta_0 + \beta_1 S + \beta_2 A + u]}{\text{Var}(S)} \\ &= \frac{\beta_1 E(S^2) - \beta_1 E(S)^2 + \beta_2 E(SA) - \beta_2 E(S)E(A) + E(uS) - E(S)E(u)}{\text{Var}(S)} \\ &= \beta_1 + \beta_2 \frac{\text{Cov}(A, S)}{\text{Var}(S)}\end{aligned}$$

- If $\beta_2 > 0$ and $\text{Cov}(A, S) > 0$ the coefficient on schooling in the shortened regression (without controlling for A) would be upward biased

Summary

- When $\text{Cov}(A, S) > 0$ then ability and schooling are correlated.
- When ability is unobserved, then not even multiple regression will identify the causal effect of schooling on wages.
- Here we see one of the main justifications for this class – what will we do when the treatment variable is endogenous?
Because endogeneity means the causal effect has not been identified.

Introduction to Selection and Randomization

- Potential outcomes concepts and definitions
- Random assignment and selection bias
- Krueger (1999)

Introduction to the Selection Problem

- A lot of the language we use in the class is from the medical literature (e.g., “treatment”, “control”)
- Simple example introducing *potential outcomes* notation and the selection problem: “Do hospitals make people healthier?”
- National Health Interview Survey (NHIS) 2005
 - Health status measured from 1 (poor health) to 5 (excellent health)

Group	Sample Size	Mean Health Status	Std. Error
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

Introductory Notation

- Think of hospitalization as a “treatment” and health status as an “outcome”.
- Let the treatment (hospitalization) be a binary variable:

$$D_i = \begin{cases} 1 & \text{if hospitalized} \\ 0 & \text{if not hospitalized} \end{cases}$$

where i indexes an individual observation, such as a person

- Observed outcomes (health status) is Y_i
- Causal question: $D \rightarrow Y$, “Does hospitalization cause health to improve?”
 - Note: we did not ask “is hospitalization correlated with health?” $\frac{1}{n} \frac{\text{Cov}(D, Y)}{\sqrt{\text{Var}_D} \sqrt{\text{Var}_Y}}$
 - What’s the difference between the two questions? What does cause mean you think? What does correlation mean?

Introduction to Rubin Causal Model

- Observed variables:
 - Treatment, D_i , is observed as either 0 or 1 for each i unit.
 - Actual outcomes, Y , are observed for each unit i
- Unobserved counterfactual variables:
 - Each individual i has one *counterfactual* (potential) outcome that in principle exists but *is not observed*:

$$\text{Potential outcome} = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

- Y_i^1 is the health of individual i (e.g. Jill) in the hospital
- Y_i^0 is the health of the *same individual* i (e.g. also Jill) not in the hospital

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i , equals $Y_i^1 - Y_i^0$

Definition 3: Switching equation

An individual's observed health outcomes, Y , is determined by treatment assignment, D_i , and corresponding potential outcomes:

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

$$Y_i = \begin{cases} Y_i^1 & \text{if } D_i = 1 \\ Y_i^0 & \text{if } D_i = 0 \end{cases}$$

Definition 2: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i^1 - Y_i^0] \\ &= E[Y_i^1] - E[Y_i^0] \end{aligned}$$

Definition 4: Fundamental problem of causal inference

It is impossible to observe both Y_i^1 and Y_i^0 for the same individual and so individual causal effects, δ_i , are unknowable.

Counterfactuals

- Consider the following two situations:
 - 1 Jack is in the hospital ($D_{Jack} = 1$) and his health is a 2 ($Y_{Jack} = 2$).
 - 2 Jill is not in the hospital ($D_{Jill} = 0$) and her health is a 4 ($Y_{Jill} = 4$)
- According to definition 3 (switching equation):
 - 1 Jack's observed health outcome, $Y_{Jack} = 2$ equals his potential health outcome under treatment, $Y_{Jack}^1 = 2$
 - 2 Jill's observed health outcome, $Y_{Jill} = 4$ equals her potential health outcome under control, $Y_{Jill}^0 = 4$

- According to definitions 1, 2 and 4:
 - 1 We do not know Jack's potential health outcome under control (i.e., Y_{Jack}^0) so therefore we do not know his treatment effect since $\delta_{Jack} = Y_{Jack}^1 - Y_{Jack}^0$
 - 2 We do not know Jill's potential health outcome under treatment (i.e., Y_{Jill}^1) so therefore we do not know her treatment effect since $\delta_{Jill} = Y_{Jill}^1 - Y_{Jill}^0$
 - 3 Therefore we do not know the average treatment effect (i.e., $E[\delta]$) because we are missing Y_{Jill}^1 and Y_{Jack}^0 – the *missing counterfactuals* (definition 4) – which are needed to calculate ATE
- We cannot calculate the average treatment effect because we are missing individual treatment effects, and we are missing individual treatment effects because we are missing counterfactuals for each unit i

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the treatment group is equal to the average treatment effect conditional on being a treatment group member:

$$\begin{aligned} E[\delta|D = 1] &= E[Y^1 - Y^0|D = 1] \\ &= E[Y^1|D = 1] - E[Y^0|D = 1] \end{aligned}$$

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the untreated group is equal to the average treatment effect conditional on being untreated:

$$\begin{aligned} E[\delta|D = 0] &= E[Y^1 - Y^0|D = 0] \\ &= E[Y^1|D = 0] - E[Y^0|D = 0] \end{aligned}$$

Simple difference in means

- The fundamental problem of causal inference states that since we cannot observe both Y_{Jack}^1 and Y_{Jack}^0 (or any i for that matter) at the same moment in time, then we cannot calculate the causal effect of hospitalization on health outcomes.
- But we have to make due with observable health outcomes, Y , since that's all we got
- If we subtract Jill's observed health outcome ($Y_{Jill} = 4$) from Jack's observed health outcome, ($Y_{Jack} = 2$), we get:

$$(Y_{Jack}|D_{Jack} = 1) - (Y_{Jill}|D_{Jill} = 0) = 2 - 4 = -2$$

which implies that hospitalization *caused* health to go from good to poor which is probably ridiculous.

- Do hospitalizations make people sick? Or do sick people go to hospitals? This is the “selection problem” – without the full potential outcomes, the simple difference between treatment and control group won’t tell us the causal effect of hospitalization on health outcomes.
- Same problem carries over to the simple difference in means, $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$, too
- So what are we actually measuring if we compare average health status for the hospitalized with that of the non-hospitalized?

Definition 7: Simple difference in mean outcomes (SDO)

A simple difference in mean outcomes (SDO) is the difference between the population average outcome for the treatment and control groups, and can be approximated by the sample averages:

$$E[Y^1|D = 1] - E[Y^0|D = 0] = E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$$

in large samples. It is called the *naive average treatment effect* in Morgan and Winship.

Notice that . . . :

- All individuals in the population contribute twice to ATE, whereas a sampled individual is used only once to estimate SDO by contributing to either $E_N[y_i|d_i = 1]$ or $E_N[y_i|d_i = 0]$.

- Statistical models, such as SDO, are valuable insofar as they can provide unbiased and/or consistent estimates of the parameter of interest (i.e., ATE). But notice the subtle difference between the LHS and RHS:

$$\begin{array}{ccc} SDO & \leq & ATE \\ E[y|d=1] - E[y|d=0] & \leq & E[Y^1] - E[Y^0] \end{array}$$

- The LHS term is the *estimator* of the RHS term which is a *parameter*, and estimators can be biased.

Biased simple difference in mean outcomes

Decomposition of the simple difference in mean outcomes (SDO)

The simple difference in mean outcomes can be decomposed into three parts (ignoring sample average notation):

$$\begin{aligned} E[Y^1|D=1] - E[Y^0|D=0] &= \text{ATE} \\ &\quad + E[Y^0|D=1] - E[Y^0|D=0] \\ &\quad + (1 - \pi)(\text{ATT} - \text{ATU}) \end{aligned} \tag{89}$$

- How do we interpret this?
 - First line is the parameter of interest: the average treatment effect, $E[Y^1 - Y^0]$
 - Second line is one of the most important sources of bias: the “selection bias” which is the difference in potential health outcome for the treatment and control group had neither received any hospitalization. In other words, how did Jack and Jill’s health status compare to one another had Jack *not* gone to the hospital? Was he sicker than her already?
 - Third line notes that part of the bias is due to heterogeneous treatment effects weighted by the share of the population in the control group
- Proof from Morgan and Winship, p. 26 on following slides

Decomposition of SDO

ATE is equal to sum of conditional average expectations by LIE

$$\begin{aligned}\text{ATE} &= E[Y^1] - E[Y^0] \\ &= \{\pi E[Y^1|D=1] + (1-\pi)E[Y^1|D=0]\} \\ &\quad - \{\pi E[Y^0|D=1] + (1-\pi)E[Y^0|D=0]\}\end{aligned}$$

Use simplified notations

$$\begin{aligned}E[Y^1|D=1] &= a \\ E[Y^1|D=0] &= b \\ E[Y^0|D=1] &= c \\ E[Y^0|D=0] &= d \\ \text{ATE} &= e\end{aligned}$$

Rewrite ATE

$$e = \{\pi a + (1-\pi)b\} - \{\pi c + (1-\pi)d\}$$

Move SDO terms to LHS

$$\begin{aligned}e &= \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\} \\e &= \pi a + b - \pi b - \pi c - d + \pi d \\e &= \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d}) \\&= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d} \\\mathbf{a} - \mathbf{d} &= e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} \\\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d \\\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)a - (1 - \pi)b + (1 - \pi)d - (1 - \pi)c \\\mathbf{a} - \mathbf{d} &= e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)(a - c) - (1 - \pi)(b - d)\end{aligned}$$

Substitute conditional means

$$\begin{aligned}E[Y^1|D=1] - E[Y^0|D=0] &= \text{ATE} \\&\quad + (E[Y^0|D=1] - E[Y^0|D=0]) \\&\quad + (1 - \pi)(\{E[Y^1|D=1] - E[Y^0|D=1]\} \\&\quad - (1 - \pi)\{E[Y^1|D=0] - E[Y^0|D=0]\})\end{aligned}$$

$$\begin{aligned}E[Y^1|D=1] - E[Y^0|D=0] &= \text{ATE} \\&\quad + (E[Y^0|D=1] - E[Y^0|D=0]) \\&\quad + (1 - \pi)(\text{ATT} - \text{ATU})\end{aligned}$$

Decomposition of difference in means

$$\underbrace{E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]}_{\text{SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0|D = 1] - E[Y^0|D = 0]}_{\text{Selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

where $E_N[y_i|d_i = 1] \rightarrow E[Y^1|D = 1]$,
 $E_N[y_i|d_i = 0] \rightarrow E[Y^0|D = 0]$ and $(1 - \pi)$ is the share of the population in the control group.

Independence assumption

Independence assumption

Treatment is independent of potential outcomes

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

In words: Random assignment means that the treatment has been assigned to units independent of their potential outcomes. Thus, mean potential outcomes for the treatment group and control group are the same *for a given state of the world*

$$E[Y^1|D = 1] = E[Y^1|D = 0]$$

$$E[Y^0|D = 1] = E[Y^0|D = 0]$$

Random Assignment Solves the Selection Problem

$$\underbrace{E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]}_{\text{SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{E[Y^0|D = 1] - E[Y^0|D = 0]}_{\text{Selection bias}} + \underbrace{(1 - \pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}$$

- Notice that the selection bias from the second line of the decomposition of SDO was:

$$E[Y^0|D = 1] - E[Y^0|D = 0]$$

- If treatment is independent of potential outcomes, then swap out equations and **selection bias** zeroes out:

$$\begin{aligned} E[Y^0|D = 1] - E[Y^0|D = 0] &= E[Y^0|D = 0] - E[Y^0|D = 0] \\ &= 0 \end{aligned}$$

Random Assignment Solves the Heterogenous Treatment Effects

- How does randomization affect heterogeneity treatment effects bias from the third line? Rewrite definitions for ATT and ATU:

$$\text{ATT} = E[Y^1|D = 1] - E[Y^0|D = 1]$$

$$\text{ATU} = E[Y^1|D = 0] - E[Y^0|D = 0]$$

- Rewrite the third row bias after $1 - \pi$:

$$\begin{aligned} \text{ATT} - \text{ATU} &= \mathbf{E}[\mathbf{Y}^1 \mid \mathbf{D}=\mathbf{1}] - E[Y^0|D = 1] \\ &\quad - \mathbf{E}[\mathbf{Y}^1 \mid \mathbf{D}=\mathbf{0}] + E[Y^0|D = 0] \\ &= 0 \end{aligned}$$

- If treatment is independent of potential outcomes, then:

$$\begin{aligned} E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0] &= E[Y^1] - E[Y^0] \\ SDO &= ATE \end{aligned}$$

What independence does not mean

- Notice – independent treatment assignment means that

$$E[Y^1|D = 1] = E[Y^1|D = 0]$$

and the equivalent for Y^0 .

- But that does not in any way imply that there is no causal effect. Independence does not imply, in other words, that $E[Y^1|D = 1]$ is equal to $E[Y^0|D = 0]$.
- Independence only implies that the average values for a given potential outcome (i.e., Y^1 or Y^0) are the same for the groups who did receive the treatment as those who did not

SUTVA

- The potential outcomes model presupposes a set of bundled assumptions called the “stable unit-treatment value assumption”
 - ① **S**: *stable*
 - ② **U**: across all **units**, or the population
 - ③ **TV**: **treatment-value** (“treatment effect”, “causal effect”)
 - ④ **A**: *assumption*
- SUTVA means that average treatment effects are parameters that assume (1) homogenous dosage and (2) potential outcomes are invariant to who else (and how many) is treated.

SUTVA: Homogenous dose

- SUTVA constrains what the treatment can be.
- Individuals are receiving the same treatment – i.e., the “dose” of the treatment to each member of the treatment group is the same. That’s the “stable unit” part.
- If we are estimating the effect of hospitalization on health status, we assume everyone is getting the same dose of the hospitalization treatment.
- Easy to imagine violations if hospital quality varies, though, across individuals. But, that just means we have to be careful what we are and are not defining as *the treatment*

SUTVA: No externalities

- What if hospitalizing Jack (hospitalized, $D = 1$) is actually about vaccinating Jack from small pox?
- If Jack is vaccinated for small pox, then Jill's potential health status (without vaccination) may be higher than when he isn't vaccinated.
- In other words, Y_{Jill}^0 , may vary with what Jack does *regardless of whether she herself receives treatment*.
- SUTVA means that you don't have a problem like this.
- If there are no externalities from treatment, then δ_i is stable for each i unit regardless of whether someone else receives the treatment too.

SUTVA: Partial equilibrium only

Easier to imagine this with a different example.

- Let's say we are estimating the effect of some technological innovation that lowers the cost functions to firms in competitive markets.
- A decrease in cost raises profits in the short-run, but positive profits leads to firm entry in the long-run.
- Firm entry in the long-run causes the supply curve to shift right, pushing market prices down until price equals average total cost.
- The first effect – short-run responses to decreases in cost – are the only things we can estimate with potential outcomes.

Example 1: Krueger (1999)

- Krueger (1999) econometrically re-analyzes a randomized experiment to determine the causal effect of class size on student achievement
- The project is the Tennessee Student/Teacher Achievement Ratio (STAR) experiment from the 1980s
- 11,600 students and their teachers were *randomly* assigned to one of the following three groups:
 - ① Small class of 13-17 students
 - ② Regular class of 22-25 students
 - ③ Regular class of 22-25 students with a full-time teacher's aide
- After the assignment, the design called for students to remain in the same class type for four years
- Randomization occurred within schools

Regression analysis of experiments

- With randomization one could simply calculate SDO (simple difference in mean outcomes) for the treatment and control group and know that SDO=ATE because of independence
- Nonetheless, it is often useful to analyze experimental data with regression analysis (see MW section 3.2.2; MHE ch. 2)
- Assume that treatment effects are constant – i.e., $Y_i^1 - Y_i^0 = \delta \forall i$
- Substitute into a rearranged switching equation (Definition 2):

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0) D_i$$

$$Y_i = Y_i^0 + \delta D_i$$

$$Y_i = E[Y_i^0] + \delta D_i + Y_i^0 - E[Y_i^0]$$

$$Y_i = \alpha + \delta D_i + \eta_i$$

where η_i is the random part of Y_i^0

- This is a regression equation that could be used to estimate the causal effect of D on Y

Regression analysis of experiments (cont.)

- The conditional expectation, $E[Y_i|D_i]$, with treatment status switched on and off gives:

$$E[Y_i|D_i = 1] = \alpha + \delta + E[\eta_i|D_i = 1]$$

$$E[Y_i|D_i = 0] = \alpha + E[\eta_i|D_i = 0]$$

- Subtracting the latter from the former, we get:

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{\text{SDO}} = \underbrace{\delta}_{\text{Treatment Effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{Selection bias}}$$

- We can estimate SDO using least squares but there's other options as well
- In the STAR experiment, D_i , equalled one if the student was enrolled in a small class and had been *randomly* assigned
- Recall that randomization implies that treatment is independent of potential outcomes, and therefore the selection bias vanishes

Why Include Control Variables?

- To evaluate experimental data, one may want to add additional controls in the multivariate regression model. So, instead of estimating the prior equation, we might estimate:

$$Y_i = \alpha + \delta D_i + X'_i \gamma + \eta_i$$

- There are 2 main reasons for including additional controls in the regression models:
 - ① Conditional random assignment. Sometimes randomization is done *conditional* on some observable. (here that's the school). We'll discuss "conditional independence assumption" when we cover matching.
 - ② Additional controls increase precision. Although control variables X_i are uncorrelated with D_i , they may have substantial explanatory power for Y_i . Including controls thus reduces variance in the residuals which lowers the standard errors of the regression estimates.

Regression in Krueger (1999)

Krueger estimates the following econometric model

$$Y_{ics} = \beta_0 + \beta_1 \text{SMALL}_{cs} + \beta_2 \text{REG}/A_{cs} + \alpha_s + \varepsilon_{ics}$$

- i indexes a student, c indexes a class, and s indexes a school
- Y_{ics} is the student's percentile score
- SMALL_{cs} is a dummy equalling 1 if she is assigned to a small class.
- REG/A_{cs} is a dummy equalling 1 if she was assigned to a regular class with an aide
- α is a “school fixed effect” which is a vector of school-specific dummy variables. He conditions on school fixed effects because randomized classroom assignment occurred *within* schools.

Regression results Kindergarten

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

Regression results 1st grade

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
B. First grade				
Small class	8.57 (1.97)	8.43 (1.21)	7.91 (1.17)	7.40 (1.18)
Regular/aide class	3.44 (2.05)	2.22 (1.00)	2.23 (0.98)	1.78 (0.98)
White/Asian (1 = yes)	—	—	6.97 (1.18)	6.97 (1.19)
Girl (1 = yes)	—	—	3.80 (.56)	3.85 (.56)
Free lunch (1 = yes)	—	—	-13.49 (.87)	-13.61 (.87)
White teacher	—	—	—	-4.28 (1.96)
Male teacher	—	—	—	11.82 (3.33)
Teacher experience	—	—	—	.05 (0.06)
Master's degree	—	—	—	.48 (1.07)
School fixed effects	No .02	Yes .24	Yes .30	Yes .30
<i>R</i> ²				

Problem 1: Attrition

A common problem in randomized experiments on humans is *attrition* – i.e., people leaving the experiment

- If attrition is random, then attrition affects the treatment and control groups in the same way and our estimates remain unbiased
- But in this application, attrition is probably non-random: especially good students from large classes may have enrolled in private schools creating a selection bias problem
- Krueger addresses this concern by imputing the test scores (from their earlier test scores) for all children who leave the sample and then re-estimates the model including students with imputed test scores.

Problem 1: Attrition

TABLE VI
EXPLORATION OF EFFECT OF ATTRITION DEPENDENT VARIABLE: AVERAGE
PERCENTILE SCORE ON SAT

Grade	Actual test data		Actual and imputed test data	
	Coefficient on small class dum.	Sample size	Coefficient on small class dum.	Sample size
K	5.32 (.76)	5900	5.32 (.76)	5900
1	6.95 (.74)	6632	6.30 (.68)	8328
2	5.59 (.76)	6282	5.64 (.65)	9773
3	5.58 (.79)	6339	5.49 (.63)	10919

Estimates of reduced-form models are presented. Each regression includes the following explanatory variables: a dummy variable indicating initial assignment to a small class; a dummy variable indicating initial assignment to a regular/aide class, unrestricted school effects; a dummy variable for student gender; and a dummy variable for student race. The reported coefficient on small class dummy is relative to regular classes. Standard errors are in parentheses.

- Non-random attrition hardly biases the results.

Problem 2: Switch Classrooms after Random Assignment

"It is virtually impossible to prevent some students from switching between class types over time." (Krueger 1999, p. 506)

B. First grade to second grade

First grade	Second grade			
	Small	Regular	Reg/aide	All
Small	1435	23	24	1482
Regular	152	1498	202	1852
Aide	40	115	1560	1715
All	1627	1636	1786	5049

- Interpreting Krueger's "transition matrix" (above)
 - If students remained in their same class type over time, all the off-diagonal elements would be zero
 - Interpretation: Of the 1,482 first graders assigned to small classrooms, 1,435 remained in small classes; 23 and 24 switched into regular and regular with aide classes in the second grade
- If students with stronger expected academic potential were more likely to move into the small classes, then these transitions would bias a simple comparison of outcomes across class types.

Problem 2: Switch Classrooms after Random Assignment

- Subjects moved between treatment and control groups. How to address this?
- A common solution to this problem is to use initial classroom assignment (i.e., small or regular classes) as an *instrument* for actual assignment. We will discuss instrumental variables later, so I will hold off on that now.
- Krueger reports regression results where instead of the student's actual status as the treatment variable, he regresses performance against their *randomly assigned class size*. This is called the “reduced form” model, and we learn more about this when we cover IV.
- In Kindergarten, OLS and reduced form are the same because students remained in their initial class for at least one year.
- From grade 1 onwards, OLS and reduced form results differ.

Problem 2: Switch Classrooms after Random Assignment

Explanatory variable	OLS: actual class size				Reduced form: initial class size			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
B. First grade								
Small class	8.57 (1.97)	8.43 (1.21)	7.91 (1.17)	7.40 (1.18)	7.54 (1.76)	7.17 (1.14)	6.79 (1.10)	6.37 (1.11)
Regular/aide class	3.44 (2.05)	2.22 (1.00)	2.23 (0.98)	1.78 (0.98)	1.92 (1.12)	1.69 (0.80)	1.64 (0.76)	1.48 (0.76)
White/Asian (1 = yes)	—	—	6.97 (1.18)	6.97 (1.19)	—	—	6.86 (1.18)	6.85 (1.18)
Girl (1 = yes)	—	—	3.80 (.56)	3.85 (.56)	—	—	3.76 (.56)	3.82 (.56)
Free lunch (1 = yes)	—	—	-13.49 (.87)	-13.61 (.87)	—	—	-13.65 (.88)	-13.77 (.87)
White teacher	—	—	—	-4.28 (1.96)	—	—	—	-4.40 (1.97)
Male teacher	—	—	—	11.82 (3.33)	—	—	—	13.06 (3.38)
Teacher experience	—	—	—	.05 (0.06)	—	—	—	.06 (.06)
Master's degree	—	—	—	.48 (1.07)	—	—	—	.63 (1.09)
School fixed effects	No .02	Yes .24	Yes .30	Yes .30	No .01	Yes .23	Yes .29	Yes .30
<i>R</i> ²								

Problem 3: Heterogeneous Treatment Effects

- Heterogeneous treatment effect bias occurs if treatment effects differ across the population (i.e., $ATT \neq ATU$)
- If people selecting to take part of the randomized trial have different returns compared to the population average, then the experiment will only identify a localized average treatment effect for the sub-population participants in the experiment.

Problem 4 and 5

- **Supply Side Changes**

- If programs are scaled up, the supply side implementing the treatment may be different
- In the trial phase, the supply side may be more motivated than during the large scale roll-out of a program

- **Attrition bias**

- Attrition rates (i.e., leaving the sample between the baseline and the follow-up surveys) may be different in treatment and control groups
- The estimated treatment effect may therefore be biased

Problem 6 and 7

- **Hawthorne effects**

- People behave differently if they are being observed in an experiment. Similar to “placebo effects” in that this is a false positive result.
- If they operate differently on treatment and control groups, then they may introduce biases
- If people from the control group behave differently, these effects are sometimes called “John Henry” effects

- **Substitution bias**

- Control group members may seek substitutes for treatment
- This would bias the estimated treatment effects *downward*. Can you see why?
- Can also occur if the experiment frees up resources that can now be concentrated on the control group.

History of graphical causal modeling in science

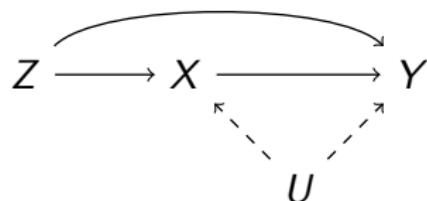
- Path diagrams were developed by Sewell Wright, early 20th century geneticist, for causal inference
- Sewell Wright's father, Phillip Wright, used them to prove the existence of the "instrumental variable" estimator (see Stock and Trebbi 2003)
- Judea Pearl and colleagues in Artificial Intelligence at UCLA developed DAG modeling to create a formalized causal inference methodology

Three biases in observational studies DAGs help illustrate

- ➊ Omitted variable bias
 - Something causes the outcome which is correlated with the treatment
- ➋ Reverse causality
 - Your logic about cause and effect is backwards
- ➌ Conditioning on collider
 - Conditioning on some variables can introduce spurious correlations

Simple DAG

A **Directed Acyclic Graph** (DAG) is a set of nodes and arrows with no directed cycles

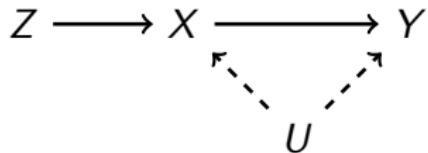


- Nodes represent variables
- Arrows represent direct causal effects (“direct” means not mediated by other variables in the graph)

Complete DAGs

- A DAG is a complete set of all causal relationships, but we emphasize the causal relationships that surround D and Y
 - All direct causal effects among the variables in the graph
 - All common causes of any pair of variables in the graph
- Do not leave out anything, even if you can't measure it or don't have data on it.
- DAGs are based on theory, a model, observation, experience, prior studies, intuition
- Try to get into the habit of accompanying every regression you will ever run with a DAG of your own creation to help guide your identification strategy

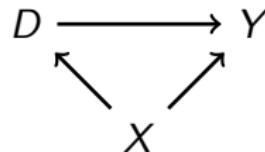
Some DAG concepts



- U is a **parent** of X and Y
- X and Y are **descendants** of Z
- There is a **directed path** from Z to Y
- There are two **paths** from Z to U (but no directed path)
- X is a **collider** of the path $Z \rightarrow X \leftarrow U$
- X is a **noncollider** of the path $Z \rightarrow X \rightarrow Y$

Confounding

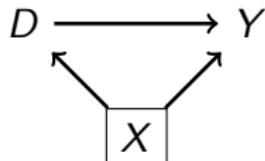
- Confounding occurs when the treatment and the outcomes have a common cause or parent which creates spurious correlation between D and Y



The *correlation* between D and Y no longer reflects the causal effect of D on Y

Backdoor Paths

- Confounding creates **backdoor paths** between treatment and outcome ($D \leftarrow X \rightarrow Y$)
- However, we can “block” the backdoor path by conditioning on the common cause X
- Once we condition on X , the correlation between D and Y estimates the causal effect of D on Y
- Conditioning means calculating $E[Y|D = 1, X] - E[Y|D = 0, X]$ for each value of X



Blocked backdoor paths

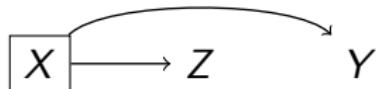
A backdoor path is blocked if and only if:

- It contains a noncollider that has been conditioned on
- Or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on

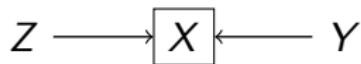
Blocked paths

Examples:

- ① Conditioning on a noncollider blocks a path:



- ② Conditioning on a collider opens a path:



- ③ *Not* conditioning on a collider blocks a path:



Backdoor criterion

Backdoor criterion

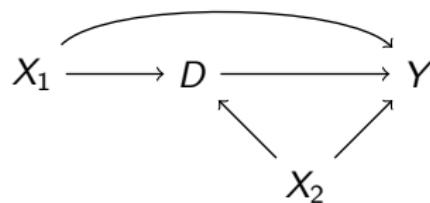
Conditioning on X satisfies the backdoor criterion with respect to (D, Y) directed path if:

- ① All backdoor paths are blocked by X
- ② No element of X is a collider

In words: If X satisfies the backdoor criterion with respect to (D, Y) , then matching on X identifies the causal effect of D on Y

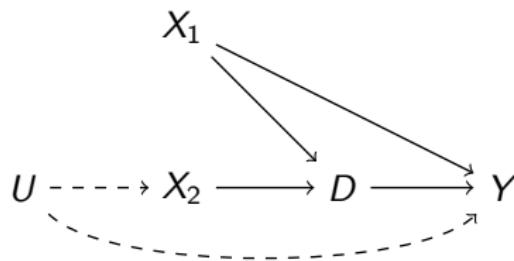
- **Matching on all common causes is sufficient**

- There are two backdoor paths from D to Y



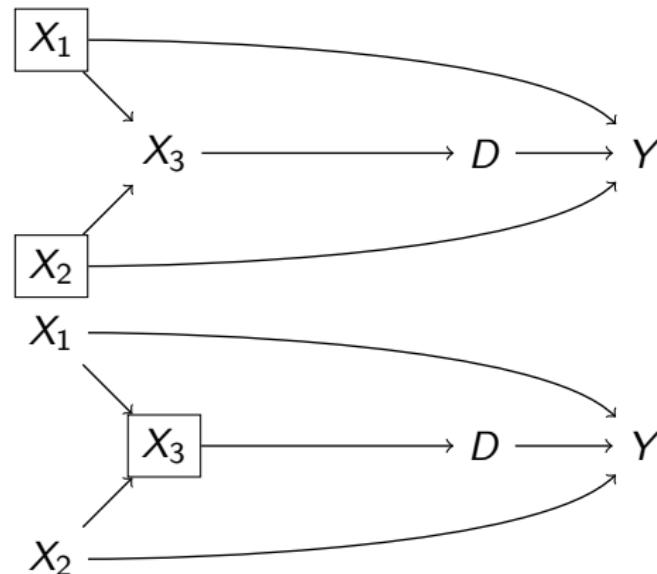
- Conditioning on X_1 and X_2 blocks both backdoor paths

- Matching may work even if not all common causes are observed
 - U and X_1 are common causes



- Conditioning on X_1 and X_2 is enough

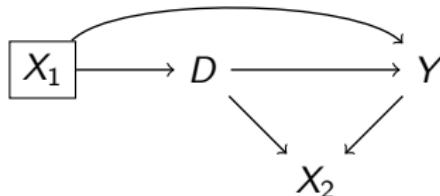
- There may be more than one set of conditioning variables that satisfy the backdoor criterion



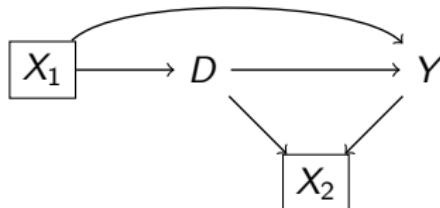
- Conditioning on the common causes, X_1 and X_2 , is sufficient
- ... but so is conditioning on X_3

Collider bias

- Matching on an outcome may create bias
 - There is only one backdoor path from D to Y

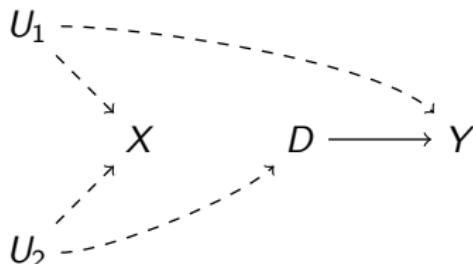


- Conditioning on X_1 blocks the backdoor path
- But what if we also condition on X_2 ?

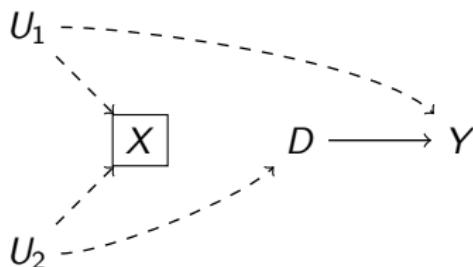


- Conditioning on X_2 opens up a new path!

- Matching on all pretreatment covariates can create bias
 - There is one backdoor path and it is closed



- No confounding – D is identified
- But what if we condition on X ? Now a backdoor path opens.



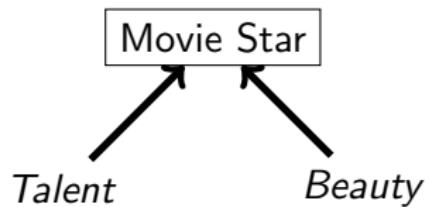
- Ultimately, we can't know if we have a collider bias problem, or whether we've satisfied the backdoor criterion, without a model
- There's no getting around it – all empirical work requires theory to guide the work. Otherwise how do you know if you've conditioning on a noncollider or a collider?
- Put differently, you cannot identify treatment effects without making assumptions about the process that generated your data

Simple example of collider bias

Important: Since unconditioned colliders block back-door paths, what exactly does conditioning on a collider do? Let's illustrate with a fun example and some made-up data

- CNN.com headline: Megan Fox voted worst – but sexiest – actress of 2009 ([link](#))
- Assume talent and beauty are independent, but each causes someone to become a movie star. What's the correlation between talent and beauty for a sample of movie stars compared to the population as a whole (stars and non-stars)?

- What if the sample consists *only* of movie stars?



STATA code

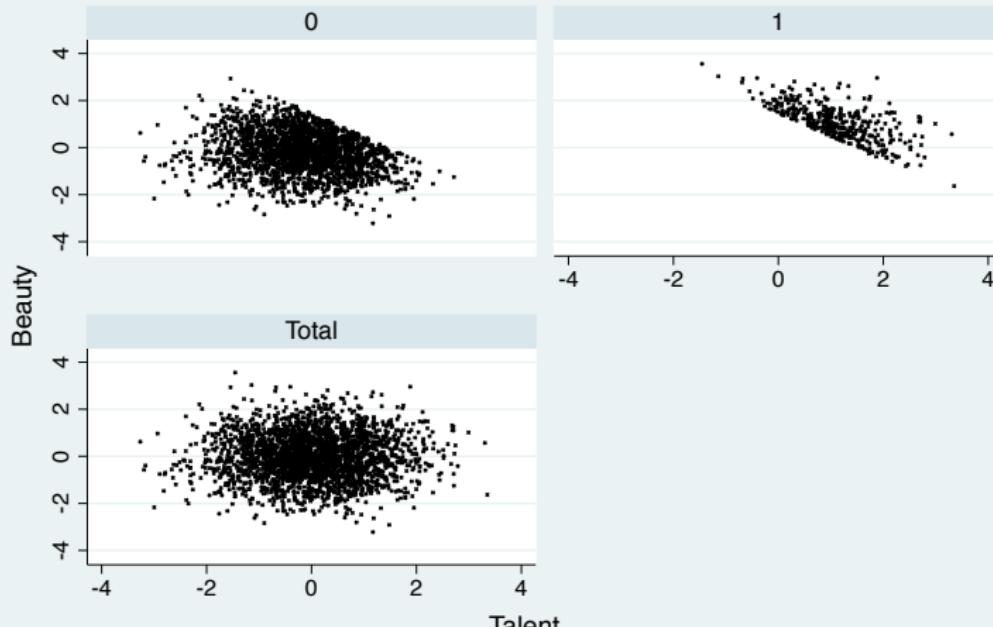
```
clear all
set seed 3444

* 2500 independent draws from standard normal distribution
set obs 2500
generate beauty=rnormal()
generate talent=rnormal()

* Creating the collider variable (star)
gen score=(beauty+talent)
egen c85=pctile(score), p(85)
gen star=(score>=c85)
label variable star "Movie star"

* Conditioning on the top 15%
twoway (scatter beauty talent, mcolor(black) msize(small) msymbol(smx)),
ytitle(Beauty) xtitle(Talent) subtitle(Aspiring actors and actresses)
by(star, total)
```

Aspiring actors and actresses



Graphs by Movie star

Figure: Top left figure: Non-star sample scatter plot of beauty (vertical axis) and talent (horizontal axis). Top right figure: Star sample scatter plot of beauty and talent. Bottom left figure: Entire (stars and non-stars combined) sample scatter plot of beauty and talent.

Final Remarks

- The backdoor criterion provides a useful graphical guide that can be used to select matching variables
 - The backdoor criterion is only a set of sufficient conditions, but covers most of the interesting cases
- Applying the backdoor criterion requires knowledge of the DAG
 - Suppose that there is a set of observed pretreatment covariates and that we know if they are causes of the treatment and/or the outcomes
 - Suppose that there exists a subset of the observed covariates such that matching on them is enough to control for confounding
 - Then, matching on the subset of the observed covariates that are either a cause of the treatment or a cause of the outcome or both is also enough to control for confounding

What if you can't conduct randomized experiment?

- Problems with the experimental design itself:
 - non-compliance by administrators
 - non-compliance by members of the treatment group
 - non-compliance by members of the control group
- Experiments may be impractical due to:
 - Too expensive
 - Unethical
 - Not feasible for some other reason

Observational alternatives to experiments

- ① Selection on observables: Treatment and control groups differ from each other on observable characteristics only
 - Stratification/subclassification
 - Matching
 - Propensity score matching
- ② Selection on unobservables: treatment and control groups differ from each other on *unobservable* characteristics
 - exogenous variable induces variation in treatment – instrumental variables
 - selection mechanism is known – regression discontinuity designs
 - treatment and controls are observed before and after treatment – panel estimators, differences-in-differences, synthetic control, even studies

Readings

- Textbook readings
 - Read MW chapter 3-4
 - Read MHE chapter 3
- Article readings
 - See website under *Matching estimation (Job Trainings Papers)*

Figure 1
Lung Cancer at Autopsy: Combined Results from 18 Studies

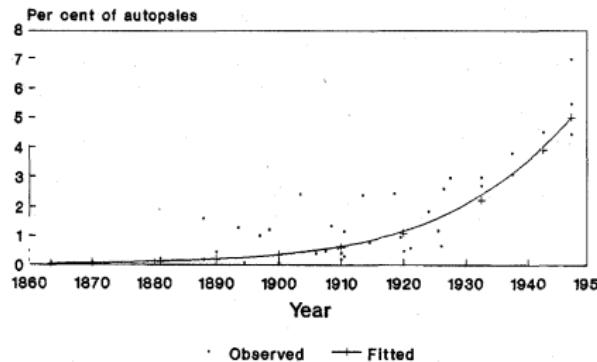


Figure 2(a)
Mortality from Cancer of the Lung in Males

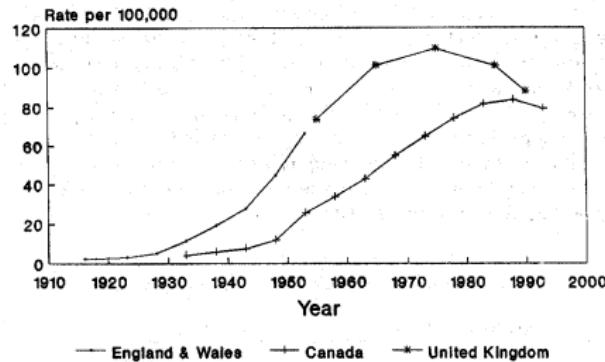
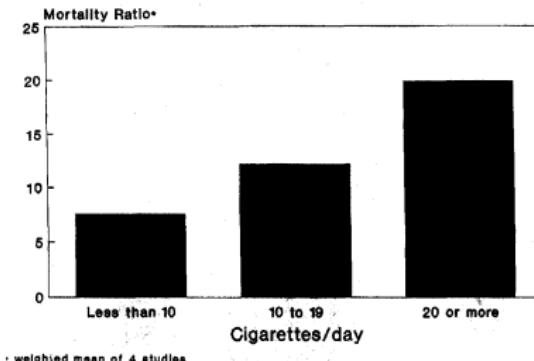


Figure 4
Smoking and Lung Cancer Case-control Studies

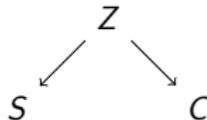


Figure 5
Smoking and Lung cancer Cohort Studies in Males



The Great Debate

- Smoking, S , causes lung cancer, C ($S \rightarrow C$) versus spurious correlation due to backdoor path:



- Criticisms from Joseph Berkson, Jerzy Neyman and Ronald Fisher: (Hill, Millar and Connelly 2003)
 - ① Correlation b/w smoking and lung cancer is spurious due to biased selection of subjects (e.g., conditioning on collider problem)
 - ② Functional form complaints about using “risk ratios” and “odds ratios”
 - ③ Confounder, Z , creates backdoor path between smoking and cancer
 - ④ Implausible magnitudes
 - ⑤ No experimental evidence to incriminate smoking as a cause of lung cancer
- Fisher's confounding theory
 - Fisher, equally famous as a geneticist, argued from logic, statistics and genetic evidence for a hypothetical confounding genome, Z , and therefore smokers and non-smokers were not exchangeable (violation of independence assumption)
 - Other studies showed that cigarette smokers and non-smokers were different on observables – more extraverted than non-smokers and pipe smokers, differed in age, differed in income, differed in education, etc.

Uh, I thought you told me Fisher was smart?

- Fisher's arguments were actually based on sound science.
- It may be too easy for us to criticize Fisher for his stance because we look back to a different time when the smoking/lung cancer link was not universally accepted, and evidence for the *causal* link was shallow:

"the [the epidemiologists] turned out to be right, but only because bad logic does not necessarily lead to wrong conclusions." Robert Hooke's (1983), How to Tell the Liars from the Statisticians.
- But FWIW, Fisher was a chain smoking pipe smoker, he died of lung cancer, and he was a paid expert witness for the tobacco industry.

Motivation: Smoking and Mortality

Table: Death rates per 1,000 person-years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

- Are cigars dangerous?

Non-smokers and smokers differ in mortality and age

Table: Mean ages, years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

- Older people die at a higher rate, and for reasons other than just smoking cigars
- Maybe cigar smokers higher observed death rates is because they're older on average

Subclassification

- One way to think about the problem is that the covariates are *not balanced* – their mean values differ for treatment and control group. So let's try to balance them.
- Subclassification (also called stratification): Compare mortality rates across the different smoking groups *within* age groups so as to neutralize covariate imbalances in the observed sample

Subclassification

- Divide the smoking group samples into age groups
- For each of the smoking group samples, calculate the mortality rates for the age group
- Construct probability weights for each age group as the proportion of the sample with a given age
- Compute the weighted averages of the age groups mortality rates for each smoking group using the probability weights

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What is the average death rate for pipe smokers?

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What is the average death rate for pipe smokers?

$$15 \cdot \left(\frac{11}{40}\right) + 35 \cdot \left(\frac{13}{40}\right) + 50 \cdot \left(\frac{16}{40}\right) = 35.5$$

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

Subclassification: example

	Death rates	Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers
Age 20-50	15	11	29
Age 50-70	35	13	9
Age +70	50	16	2
Total		40	40

Question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

$$15 \cdot \left(\frac{29}{40} \right) + 35 \cdot \left(\frac{9}{40} \right) + 50 \cdot \left(\frac{2}{40} \right) = 21.2$$

Table: Adjusted death rates using 3 age groups (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

Covariates

Definition: Predetermined Covariates

Variable X is predetermined with respect to the treatment D (also called “pretreatment”) if for each individual i , $X_i^0 = X_i^1$, i.e., the value of X_i does not depend on the value of D_i . Such characteristics are called *covariates*.

Comment I: Does not imply X and D are independent

Comment II: Predetermined variables are often time invariant (e.g., sex, race), but time invariance is not a necessary condition

Outcomes

Definition: Outcomes

Those variables, Y , that are (possibly) not predetermined are called *outcomes* (for some individual i , $Y_i^0 \neq Y_i^1$)

Comment: Recall the “collider bias”; in general, one shouldn’t condition on outcomes because it may induce bias

Adjustment for Observables

- Subclassification
- Matching
- Propensity score methods
- Regression

Identification under independence

- Recall that randomization implies

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

- and therefore:

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= \underbrace{E[Y^1|D=1] - E[Y^0|D=0]}_{\text{by the switching equation}} \\ &= \underbrace{E[Y^1] - E[Y^0]}_{\text{by independence}} \\ &= \underbrace{E[Y^1 - Y^0]}_{\text{ATE}} \end{aligned}$$

- As well as that $ATT = ATE$:

$$E[Y^1 - Y^0] = E[Y^1 - Y^0|D=1]$$

Identification under conditional independence

Identification assumptions:

- ① $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (conditional independence)
- ② $0 < Pr(D = 1|X) < 1$ with probability one (common support)

Identification result:

- Given assumption 1:

$$\begin{aligned} E[Y^1 - Y^0|X] &= E[Y^1 - Y^0|X, D = 1] \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

- Given assumption 2:

$$\begin{aligned} \delta_{ATE} &= E[Y^1 - Y^0] \\ &= \int E[Y^1 - Y^0|X, D = 1] dPr(X) \\ &= \int (E[Y|X, D = 1] - E[Y|X, D = 0]) dPr(X) \end{aligned}$$

Identification under conditional independence

Identification assumptions:

- ① $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (conditional independence)
- ② $0 < Pr(D = 1|X) < 1$ with probability one (common support)

Identification result:

- Similarly

$$\begin{aligned}\delta_{ATT} &= E[Y^1 - Y^0 | D = 1] \\ &= \int (E[Y|X, D = 1] - E[Y|X, D = 0]) dPr(X|D = 1)\end{aligned}$$

- To identify δ_{ATT} the conditional independence and common support assumptions can be relaxed to:

- ① $Y^0 \perp\!\!\!\perp D|X$
- ② $Pr(D = 1|X) < 1$ (with $Pr(D = 1) > 0$)

Subclassification estimator

- The identification result is:

$$\delta_{ATE} = \int (E[Y|X, D=1] - E[Y|X, D=0]) dPr(X)$$

$$\delta_{ATT} = \int (E[Y|X, D=1] - E[Y|X, D=0]) dPr(X|D=1)$$

- Assume X takes on K different cells $\{X^1, \dots, X^k, \dots, X^K\}$.
Then the analogy principle suggests the following estimators:

$$\hat{\delta}_{ATE} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$$

$$\hat{\delta}_{ATT} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$$

where N^k is the number of obs. and N_T^k is the number of treatment observations in cell k ; $\bar{Y}^{1,k}$ is the mean outcome for the treated in cell k ; $\bar{Y}^{0,k}$ is the mean outcome for the control in cell k

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N}\right)$?

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta}_{ATE} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N}\right)$?

$$4 \cdot \left(\frac{13}{30}\right) + 6 \cdot \left(\frac{17}{30}\right) = 5.13$$

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

Subclassification by Age ($K = 2$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

$$4 \cdot \left(\frac{3}{10} \right) + 6 \cdot \left(\frac{7}{10} \right) = 5.4$$

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Not identified!

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

Subclassification by Age and Gender ($K = 4$)

X_k	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

$$4 \cdot \left(\frac{3}{10} \right) + 5 \cdot \left(\frac{3}{10} \right) + 6 \cdot \left(\frac{4}{10} \right) = 5.1$$

Curse of Dimensionality

- Subclassification may become less feasible in finite samples as the number of covariates grows (e.g., $K = 4$ was too many for this sample)
- Assume we have k covariates and we divide each into 3 coarse categories (e.g., age: young, middle age, old; income: low, medium, high, etc.)
- The number of sub classification cells (or “strata”) is 3^k . For $k = 10$, then it's $3^{10} = 59,049$

Curse of Dimensionality

- If sparseness occurs, it means many cells may contain either only treatment units or only control units but not both. If so, we cannot use sub classification.
- Subclassification is also a problem if the cells are “too coarse”. We can always use “finer” classifications, but finer cells worsens the dimensional problem, so we don’t gain much from that. ex: using 10 variables and 5 categories for each, we get $5^{10} = 9,765,625$.

Matching

- Alternatively, we could estimate δ_{ATT} by *imputing* the missing potential outcome of each treatment unit using the observed outcome from that outcome's “closest” control unit

$$\delta_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where $Y_{j(i)}$ is the observed outcome of a control unit such that $X_{j(i)}$ is the **closest** value to X_i among all of the control observations

Matching

- We could also use the average observed outcome over M closest matches:

$$\delta_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right)$$

- Works well when we can find good matches for each treatment group unit, so M is usually defined to be small (i.e., $M = 1$ or $M = 2$)

Matching

- We can also use matching to estimate δ_{ATE} . In that case, we match in both directions:
 - ① If observation i is treated, we impute Y_i^0 using the control matches, $\{Y_{j_1(i)}, \dots, Y_{j_M(i)}\}$
 - ② If observation i is control, we impute Y_i^1 using the treatment matches, $\{Y_{j_1(i)}, \dots, Y_{j_M(i)}\}$
- The estimator is:

$$\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^N (2D_i - 1) \left[Y_i - \left(\frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} \right) \right]$$

Matching example with single covariate

unit	Potential Outcome			
	under Treatment	under Control	D_I	X_i
i	Y_i^1	Y_i^0		
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta_{ATT}} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

Matching example with single covariate

unit	Potential Outcome			
	under Treatment	under Control	D_I	X_i
i	Y_i^1	Y_i^0		
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta_{ATT}} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

Match and plug in!

Matching example with single covariate

unit	Potential Outcome		D_I	X_i
	under Treatment	under Control		
i	Y_i^1	Y_i^0		
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is $\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$?

$$\widehat{\delta}_{ATT} = \frac{1}{3} \cdot (6 - 9) + \frac{1}{3} \cdot (1 - 0) + \frac{1}{3} \cdot (0 - 9) = -3.7$$

A Training Example

Trainees			Non-Trainees		
unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900
2	34	10200	2	50	31000
3	29	14400	3	30	21000
4	25	20800	4	27	9300
5	29	6100	5	54	41100
6	23	28600	6	48	29800
7	33	21900	7	39	42000
8	27	28800	8	28	8800
9	31	20300	9	24	25500
10	26	28100	10	33	15500
11	25	9400	11	26	400
12	27	14300	12	31	26600
13	29	12500	13	26	16500
14	24	19700	14	34	24200
15	25	10100	15	25	23300
16	43	10700	16	24	9700
17	28	11500	17	29	6200
18	27	10700	18	35	30200
19	28	16300	19	32	17800
Average:		28.5	16426	20	23
			21	32	25900
			Average:		20724

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
			Average:	33	20724			

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
				21	32	25900		
				Average:		20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

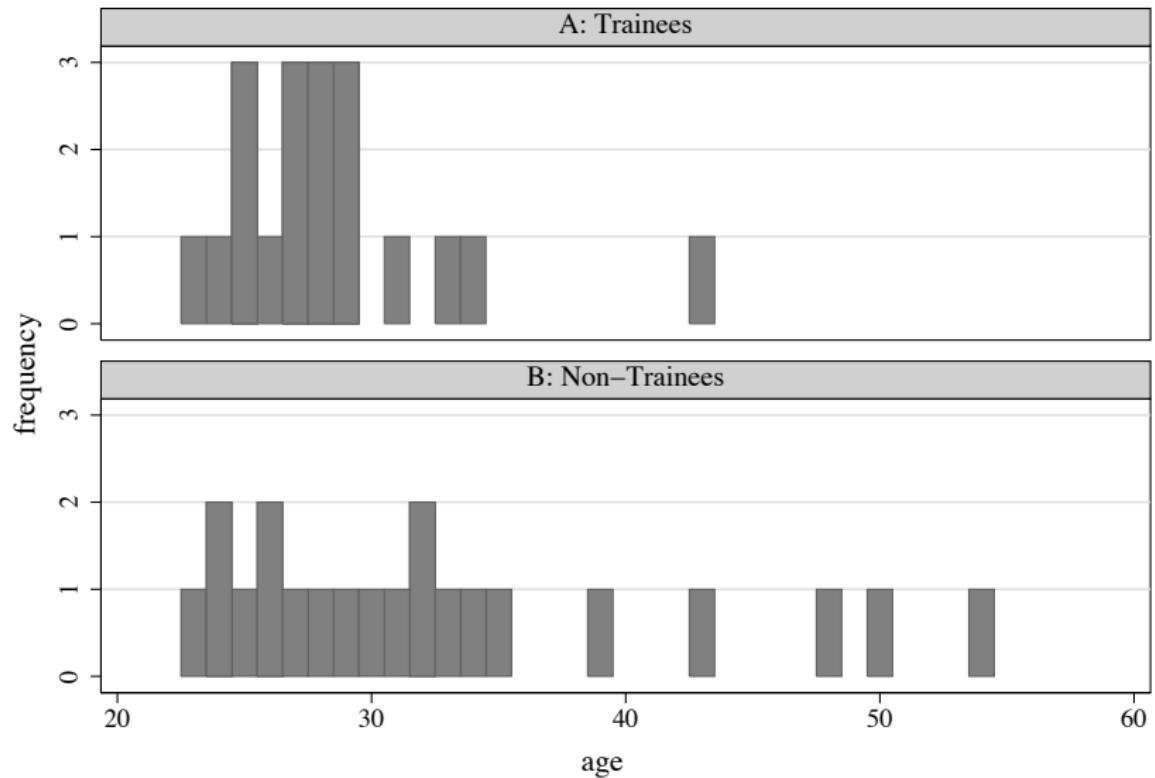
A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:		20724			

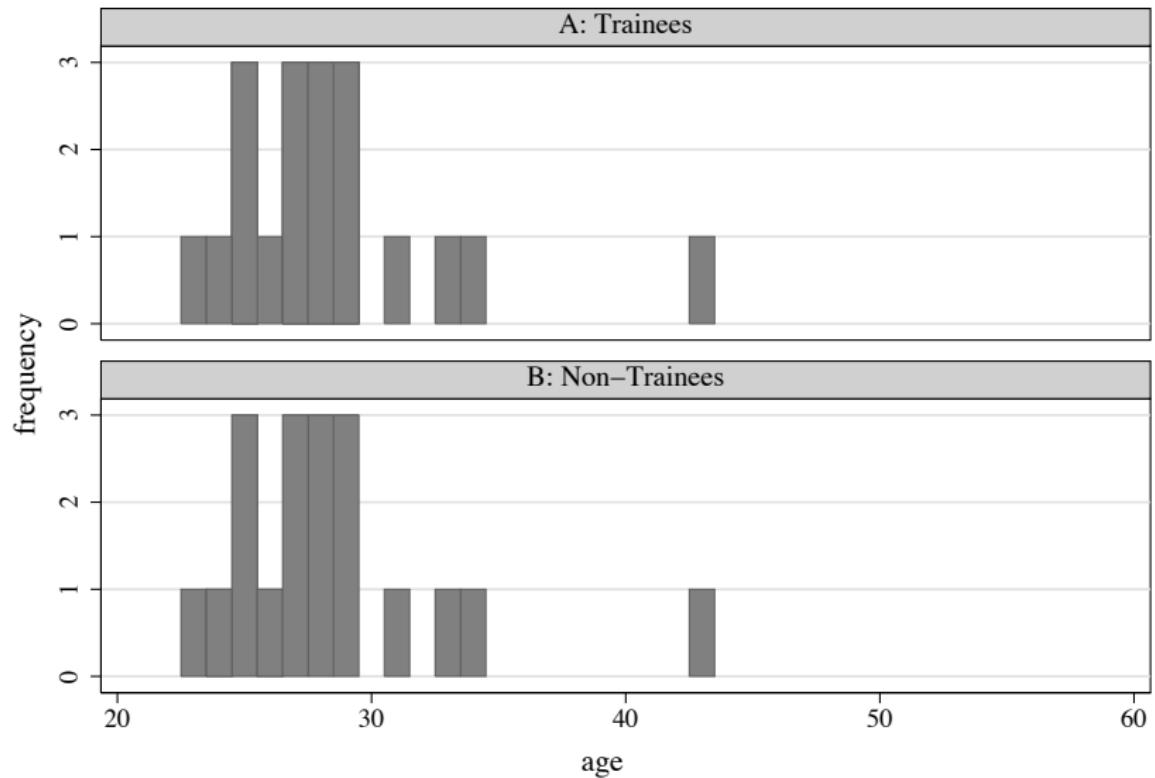
A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:	28.5	13982
			21	32	25900			
			Average:		20724			

Age Distribution: Before Matching



Age Distribution: After Matching



Training Effect Estimates

Difference in average earnings between trainees and non-trainees

- Before matching

$$16426 - 20724 = -4298$$

- After matching:

$$16426 - 13982 = 2444$$

Alternative distance metric: Euclidean distance

When the vector of matching covariates, $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$ has more

than one dimension ($k > 1$) we will need a new definition of **distance** to measure “closeness”.

Definition: Euclidean distance

$$\begin{aligned} \|X_i - X_j\| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{n=1}^k (X_{ni} - X_{nj})^2} \end{aligned}$$

Comment: The Euclidean distance is not invariant to changes in the scale of the X 's. For this reason, alternative distance metrics that are invariant to changes in scale are used.

Normalized Euclidean distance

Definition: Normalized Euclidean distance

A commonly used distance is the normalized Euclidean distance:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{V}^{-1} (X_i - X_j)}$$

where

$$\hat{V}^{-1} = \begin{pmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\sigma}_k^2 \end{pmatrix}$$

- Notice that the normalized Euclidean distance is equal to:

$$\|X_i - X_j\| = \sqrt{\sum_{n=1}^k \frac{(X_{ni} - X_{nj})^2}{\hat{\sigma}_n^2}}$$

- Thus, if there are changes in the scale of X_{ni} , these changes also affect $\hat{\sigma}_n^2$, and the normalized Euclidean distance does not change

Mahalanobis distance

Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{\Sigma}_X^{-1} (X_i - X_j)}$$

where $\hat{\Sigma}_X$ is the sample variance-covariance matrix of X .

Comments: Related, you can create arbitrary distances like:

$$\|X_i - X_j\| = \sqrt{\sum_{n=1}^k \omega_n \cdot (X_{ni} - X_{nj})^2}$$

(with all $\omega_n \geq 0$) so that we assign large ω_n 's to those covariates that we want to match particularly well.

Matching and the Curse of Dimensionality

Dimensionality creates headaches for us in matching.

- **Bad news:** Matching discrepancies $\|X_i - X_{j(i)}\|$ tend to increase with k , the dimension of X
- **Good news:** Matching discrepancies converge to zero ...
- **Bad news:** ... but they converge very slow if k is large
- **Good news?:** Mathematically, it can be shown that $\|X_i - X_{j(i)}\|$ converges to zero at the same rate as $\frac{1}{N^{\frac{1}{k}}}$
- **Bad news:** It's hard to find good matches when X has a large dimension: you need many observations if k is big.

Deriving the matching bias

Derive the matching bias by first writing out the sample ATT estimate:

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}),$$

where each i and $j(i)$ units are matched, $X_i \approx X_{j(i)}$ and $D_{j(i)} = 0$. Define potential outcomes and switching eq.

$$\begin{aligned}\mu^0(x) &= E[Y|X = x, D = 0] = E[Y^0|X = x], \\ \mu^1(x) &= E[Y|X = x, D = 1] = E[Y^1|X = x], \\ Y_i &= \mu^{D_i}(X_i) + \varepsilon_i\end{aligned}$$

Substitute and distribute terms

$$\begin{aligned}\hat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} [(\mu^1(X_i) + \varepsilon_i) - (\mu^0(X_{j(i)}) + \varepsilon_{j(i)})] \\ &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)})) + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Deriving the matching bias

Difference between sample estimate and population parameter is:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)}) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Algebraic manipulation and simplification:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_i) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\mu^0(X_i) - \mu^0(X_{j(i)})) .\end{aligned}$$

Deriving the matching bias

Apply the Central Limit Theorem and the difference,

$$\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT}),$$

converges to a Normal distribution with zero mean. But, however,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly
- The difference $\mu^0(X_i) - \mu^0(X_{j(i)})$ converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$ may not converge to zero and can be very large!
- $E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right]$ may not converge to zero because the bias of the matching discrepancy is dominating the matching estimator!

Bias is often an issue when we match in many dimensions

Solutions to matching bias problem

The bias of the matching estimator is caused by large matching discrepancies $||X_i - X_{j(i)}||$. The curse of dimensionality virtually guarantees this. However:

- ① But the matching discrepancies are observed. We can always check in the data how well we're matching the covariates.
- ② For $\widehat{\delta}_{ATT}$ we can always make the matching discrepancies small by using a large reservoir of untreated units to select the matches (that is, by making N_C large).
- ③ If the matching discrepancies are large, so we are worried about potential biases, we can apply bias correction techniques
- ④ Partial solution: propensity score methods (coming soon...)

Matching with bias correction

- Each treated observation contributes

$$\mu^0(X_i) - \mu^0(X_{j(i)})$$

to the bias.

- Bias-corrected (BC) matching:

$$\widehat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[(Y_i - Y_{j(i)}) - (\widehat{\mu^0}(X_i) - \widehat{\mu^0}(X_{j(i)})) \right]$$

where $\widehat{\mu^0}(X)$ is an estimate of $E[Y|X = x, D = 0]$. For example using OLS.

- Under some conditions, the bias correction eliminates the bias of the matching estimator without affecting the variance.

Bias adjustment in matched data

unit	Potential Outcome			
	under Treatment	under Control	D_i	X_i
i	Y_i^1	Y_i^0		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

Bias adjustment in matched data

unit	Potential Outcome		D_i	X_i
	under Treatment	under Control		
i	Y_i^1	Y_i^0		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\widehat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		D_i	X_i
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\widehat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

$$\widehat{\delta}_{ATT} = \frac{(10 - 8) - (\widehat{\mu^0}(3) - \widehat{\mu^0}(4))}{3} + \frac{(4 - 1) - (\widehat{\mu^0}(1) - \widehat{\mu^0}(0))}{3} + \frac{(10 - 9) - (\widehat{\mu^0}(10) - \widehat{\mu^0}(8))}{3} = 1.33$$

Matching bias: Implications for practice

Bias arises because of the effect of large matching discrepancies on $\mu^0(X_i) - \mu^0(X_{j(i)})$. To minimize matching discrepancies:

- ① Use a small M (e.g., $M = 1$). Larger values of M produce large matching discrepancies.
- ② Use matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller matching discrepancies than matching without replacement.
- ③ Try to match covariates with a large effect on $\mu^0(\cdot)$ particularly well.

Large sample distribution for matching estimators

- Matching estimators have a Normal distribution in large samples (provided the bias is small):

$$\sqrt{N_T}(\hat{\delta}_{ATT} - \delta_{ATT}) \xrightarrow{d} N(0, \sigma_{ATT}^2)$$

- For matching without replacement, the “usual” variance estimator:

$$\hat{\sigma}_{ATT}^2 = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \hat{\delta}_{ATT} \right)^2,$$

is valid.

Large sample distribution for matching estimators

- For matching with replacement:

$$\begin{aligned}\widehat{\sigma}_{ATT}^2 &= \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2 \\ &+ \frac{1}{N_T} \sum_{D_i=0} \left(\frac{K_i(K_i-1)}{M^2} \right) \widehat{\text{var}}(\varepsilon | X_i, D_i = 0)\end{aligned}$$

where K_i is the number of times observation i is used as a match.

- $\widehat{\text{var}}(Y_i | X_i, D_i = 0)$ can be estimated also by matching. For example, take two observations with $D_i = D_j = 0$ and $X_i \approx X_j$, then

$$\widehat{\text{var}}(Y_i | X_i, D_i = 0) = \frac{(Y_i - Y_j)^2}{2}$$

is an unbiased estimator of $\widehat{\text{var}}(\varepsilon_i | X_i, D_i = 0)$

- The bootstrap doesn't work!

Propensity score

- Overview
 - What do we use a propensity score for?
 - How do we construct the propensity score?
 - How do we implement propensity score estimation in STATA?
- Discuss several articles using propensity score matching

Joke (sort of...)

- Two heart surgeons walk into a bar.
 - Jack: "I just finished my 100th heart surgery!"
 - Jill: "I finished my 100th heart surgery last week. So I have more experience than you, and am therefore probably a better surgeon. How many of your patients died within 3 months of surgery? I've only had 10 die."
 - Jack: "Five. Which means I'm probably the better surgeon."
 - Jill: "Or maybe my patients are older and have a higher risk of mortality than your patients."
- I didn't say it was a good joke.
- They are debating about who is the better surgeon – the treatment of interest. But if the patients' characteristics differ between Jack and Jill, and those characteristics are correlated with mortality, then it confounds the inference
- To really judge who is the better surgeon, we want them to each see the same groups of patients of equivalent characteristics

Purpose of propensity scores

- When treatment is non-random, you can use propensity score matching to compare a treatment and control group who are equivalent on observable characteristics
- Propensity score matching also provides a way to summarize covariate information about treatment selection into a single scalar
- Can be used to adjust for differences via study design, or matching, or during estimation of the treatment effect (e.g., subclassification or regression)

Some caveats

- This is only relevant for selection on observables
- If you cannot write down a DAG such that conditioning on some covariate(s) meets the backdoor criterion, then propensity score matching is not the appropriate methodology
- You need to exhaustively identify the variables, X , that will block all back door paths between the treatment, D , and the outcome, Y using your knowledge of theory and the institutional details

Motivation

- If the treatment is randomly assigned, then we use the simple difference in outcomes (SDO) estimator

$$\frac{1}{N_T} \sum(Y|D=1) - \frac{1}{N_C} \sum(Y|D=0)$$

- This is called “ignobility” when we have randomized treatment assignment
- But what if ignobility is violated?
- Propensity score matching has a weaker assumption called conditional independence which we will discuss later

OLS

- In principle you could estimate the treatment effect using regression and control for X

$$Y = \delta D + \beta X + \varepsilon$$

where X is a matrix of covariates that you think affect the probability of receiving a scholarship

- OLS consistently estimates the conditional mean, but if probability of treatment is nonlinear, this conditional mean may be less informative
- Usually, we won't know how selection depended on X – only that it did (e.g., DAG and the backdoor criterion)

Propensity score

- The idea with matching on X 's was to compare units who were "close" to one another based on some distance to the nearest X neighbor. But we had some issue with matching discrepancies and the the curse of dimensionality.
- The idea with propensity score matching is to compare units who, based solely on their observables, had very similar probabilities of being placed into treatment
- If conditional on X , two units have a similar probability of treatment, then we say they have similar *propensity scores*
- If two units have equivalent propensity scores, then differences between their outcomes is attributable to the treatment
- If we compare a unit in the treatment group to a control group unit with two similar propensity scores, then conditional on the propensity score, all remaining variation between these two is randomness—if selection on observables is a correct assumption.

First: How do we practically do this?

- Estimation using propensity score matching is a two-step procedure
 - ① First step: estimate the propensity score
 - ② Second step: calculate the average causal effect of interest by averaging differences in outcomes over units with similar propensity scores
- Estimating the propensity score
 - Estimate the following equation with binary treatment D on the left-hand-side and observables X related to selection determination (e.g., backdoor criterion) on the right-hand-side using probit or logit model

$$\text{Prob}(D = 1|X) = \gamma X + \omega$$

- Second using the estimated coefficients, calculate the predicted probability of treatment

$$\hat{\rho} = \hat{\gamma} X$$

- The propensity score is just the predicted conditional probability of treatment, or the fitted value for each unit

Definition

Definition of Propensity score

Propensity score is defined as the selection probability conditional on the confounding variables: $p(X) = Pr(D = 1|X)$

Identification Assumptions:

- ① $(Y^0, Y^1) \perp\!\!\!\perp D|X$ (conditional independence)
- ② $0 < Pr(D = 1|X) < 1$ (common support)

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D | X_i$. There exists a set X of observable covariates such that after controlling for these covariates, treatment assignment is independent of potential outcomes.

- **Interpretation:** After controlling for X , the assignment of units to the treatment is 'as good as random'.
- **Why?**: Conditional independence allows us the ability to construct an unbiased counterfactual for the treatment group using the non-experimental control group units.
- Also called 'ignorable treatment assignment' or 'unconfoundedness' (statistics), backdoor criterion (Pearl), 'selection on observables', or 'exogeneity' (economics). Exogeneity in traditional econometric pedagogy

$$\begin{aligned} Y_i^0 &= \alpha + \beta' X_i + \varepsilon, \text{ and} \\ Y_i^1 &= Y_i^0 + \delta \end{aligned}$$

then we can write:

$$Y_i = \alpha + \delta D_i + \beta' X_i + \varepsilon_i$$

and conditional independence $\iff \varepsilon_i \perp\!\!\!\perp D_i | X_i$ (exogeneity)

- Conditional independence is **not testable**

Identifying assumption II: Common support

For each value of X , there is a positive probability of being both treated and untreated

$$0 < \Pr(D_i = 1 | X_i) < 1$$

- **Interpretation:** Implies that the probability of receiving treatment for every value of the vector X is strictly within the unit interval: as is the probability of not receiving treatment.
- **Why?:** Common support ensures there is sufficient overlap in the characteristics of treated and untreated units to find adequate matches
- Common support is **testable**

Identification given strong ignorability

- When both assumptions are satisfied, the treatment assignment is said to be **strongly ignorable** in the terminology of Rosenbaum and Rubin (1983).
- By definition

$$\begin{aligned}\delta_i(X_i) &= E[Y_i^1 - Y_i^0 | X_i = x] \\ &= E[Y_i^1 | X_i = x] - E[Y_i^0 | X_i = x]\end{aligned}$$

- Unconfoundedness allow us to substitute

$$E[Y_i^1 | D_i = 1, X_i = x] = E[Y_i | D_i = 1, X_i = x]$$

and similar for other term.

- Common support allows us to estimate both terms
- Then, $\delta = E[\delta(X_i)]$

Propensity score theorem (Rosenbaum and Rubin 1983)

If $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (conditional independence), then $(Y^1, Y^0) \perp\!\!\!\perp D|p(X)$ where $p(X) = \Pr(D = 1|X)$, the propensity score

Sufficient conditions: Conditioning on the propensity score is enough to have independence between the treatment indicator, D , and the potential outcomes, (Y^1, Y^0)

Dimension reduction: Extremely valuable theorem. Here's why

- Stratifying on X tends to run into sparseness-related problems (i.e., empty cells) in finite samples for even moderate number of covariates
- But the propensity score, $p(X)$, is just a scalar, and therefore stratifying across that probability is going to be a lot easier.

Proof: Straightforward application of the law of iterated expectations (LIE) with nested conditioning from MHE (pp. 80-81)

Proof

- If we can show that the probability an individual receives treatment conditional on potential outcomes and the propensity score is not a function of potential outcomes

$$\Pr(D = 1 | Y^0, Y^1, p(X)) \neq f(Y^0, Y^1)$$

then we will have proven that there is an independence between Y^0, Y^1 and D conditional on X

- Before diving into the proof, first recognize that

$$\Pr(D = 1 | Y^0, Y^1, p(X)) = E[D | Y^0, Y^1, p(X)]$$

because

$$\begin{aligned} E[D | Y^0, Y^1, p(X)] &= 1 \cdot \Pr(D = 1 | Y^0, Y^1, p(X)) \\ &\quad + 0 \cdot \Pr(D = 0 | Y^0, Y^1, p(X)) \end{aligned}$$

and the second term cancels out.

Assume that $(Y^1, Y^0) \perp\!\!\!\perp D|X$. Then:

$$\begin{aligned}
 \Pr(D = 1|Y^1, Y^0, p(X)) &= \underbrace{E[D|Y^1, Y^0, p(X)]}_{\text{See previous slide}} \\
 &= \underbrace{E[E[D|Y^1, Y^0, p(X), X]|Y^1, Y^0, p(X)]}_{\text{by LIE}} \\
 &= \underbrace{E[E[D|Y^1, Y^0, X]|Y^1, Y^0, p(X)]}_{\text{Given } X, \text{ we know } p(X)} \\
 &= \underbrace{E[E[D|X]|Y^1, Y^0, p(X)]}_{\text{by conditional independence}} \\
 &= \underbrace{E[p(X)|Y^1, Y^0, p(X)]}_{\text{propensity score definition}} \\
 &= p(X)
 \end{aligned}$$

□

- Using a similar argument, we obtain:

$$\begin{aligned}
 \Pr(D = 1|p(X)) &= \underbrace{E[D|p(X)]}_{\text{Previous slide}} = \underbrace{E[E[D|X]|p(X)]}_{\text{LIE}} \\
 &= \underbrace{E[p(X)|p(X)]}_{\text{definition}} = p(X)
 \end{aligned}$$

and $\Pr(D = 1|Y^1, Y^0, p(X)) = \Pr(D = 1|p(X))$ by conditional independence

Propensity score theorem

- Like the omitted variable bias formula for regression, the propensity score theorem says that you need only control for covariates that affect the probability of treatment
- But it also says something more
- The *only* covariate you really need to control for is the probability of treatment itself
- Direct propensity score matching works in the same way as covariate matching (e.g., nearest neighbor matching) except that we match on the *score* instead of the *covariates* directly

Corollary

If $(Y^1, Y^0) \perp\!\!\!\perp D|X$, we can estimate average treatment effects:

$$E[Y^1 - Y^0|p(X)] = E[Y|D = 1, p(X)] - E[Y|D = 0, p(X)]$$

Three-step procedure:

- ① Estimate the conditional probability of treatment, or propensity score, $p(X) = Pr(D = 1|X)$, using any standard probability model (logit or probit)
- ② Do matching, sub classification (stratification), inverse probability weighting or some other algorithmic procedures to estimate the average causal effect conditional on the estimated propensity score
- ③ Compute standard errors

Balancing property

- Because the propensity score is a function of X , we know:

$$\begin{aligned} \Pr(D = 1|X, p(X)) &= \Pr(D = 1|X) \\ &= p(X) \end{aligned}$$

\therefore conditional on $p(X)$, the probability that $D = 1$ does not depend on X .

- D and X are independent conditional on $p(X)$:

$$D \perp\!\!\!\perp X|p(X)$$

- So we obtain the **balancing property** of the propensity score:

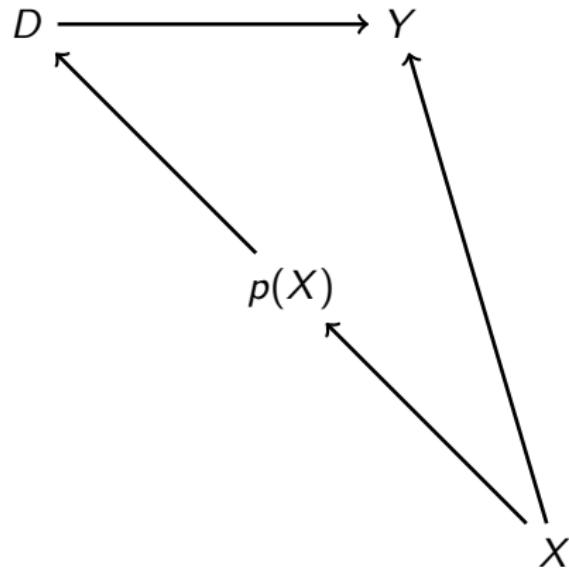
$$\Pr(X|D = 1, p(X)) = \Pr(X|D = 0, p(X))$$

conditional on the propensity score, the distribution of the covariates is the same for treatment and control group units

- We can use this to check if our estimated propensity score actually produces balance:

$$\Pr(X|D = 1, \hat{p}(X)) = \Pr(X|D = 0, \hat{p}(X))$$

Balancing Property



Inverse Probability Weighting

Proposition

If $Y^1, Y^0 \perp\!\!\!\perp D|X$, then

$$\delta_{ATE} = E[Y^1 - Y^0] = E \left[Y \cdot \frac{D - p(X)}{p(X) \cdot (1 - p(X))} \right]$$

$$\delta_{ATT} = E[Y^1 - Y^0 | D = 1] = \frac{1}{Pr(D = 1)} \cdot E \left[Y \cdot \frac{D - p(X)}{1 - p(X)} \right]$$

Proof.

$$\begin{aligned} E \left[Y \frac{D - p(X)}{p(X)(1 - p(X))} \middle| X \right] &= E \left[\frac{Y}{p(X)} \middle| X, D = 1 \right] p(X) \\ &\quad + E \left[\frac{-Y}{1 - p(X)} \middle| X, D = 0 \right] (1 - p(X)) \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

and the results follow from integrating over $P(X)$ and $P(X|D = 1)$. □

Weighting on the propensity score

ATE and ATT parameters (previous slide):

$$\delta_{ATE} = E[Y^1 - Y^0] = E \left[Y \cdot \frac{D - p(X)}{p(X) \cdot (1 - p(X))} \right]$$

$$\delta_{ATT} = E[Y^1 - Y^0 | D = 1] = \frac{1}{Pr(D = 1)} \cdot E \left[Y \cdot \frac{D - p(X)}{1 - p(X)} \right]$$

The analogy principle suggests a two-step estimator:

- 1 Estimate the propensity score: $\hat{p}(X)$
- 2 Use estimated score to produce analog estimators. Let $\hat{\delta}_{ATE}$ and $\hat{\delta}_{ATT}$ be an estimate of the ATE and ATT parameter:

$$\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{p}(X_i)}{\hat{p}(X_i) \cdot (1 - \hat{p}(X_i))}$$

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{p}(X_i)}{1 - \hat{p}(X_i)}$$

Weighting on the propensity score

Sample analogs of the population parameters (previous slide):

$$\begin{aligned}\widehat{\delta}_{ATE} &= \frac{1}{N} \sum_{i=1}^N Y_i \cdot \frac{D_i - \widehat{p}(X_i)}{\widehat{p}(X_i) \cdot (1 - \widehat{p}(X_i))} \\ \widehat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{i=1}^N Y_i \cdot \frac{D_i - \widehat{p}(X_i)}{1 - \widehat{p}(X_i)}\end{aligned}$$

Standard errors:

- We need to adjust the standard errors for first-step estimation of $p(X)$
- Parameteric first step: Newey and McFadden (1994)
- Non-parametric first step: Newey (1994)
- Or bootstrap the entire two-step procedure.

Other algorithmic methods

- ① Estimate propensity score
- ② Choose algorithm
- ③ Check overlap/common support
- ④ Matching Quality/Effect estimation
- ⑤ Sensitivity Analysis

Estimation

- Once you have determined the propensity score with the procedure above, there are several ways to use it to estimate average treatment effects
 - ① Stratification. Run the second step within each group. Calculate the weighted mean of the within-group estimates to get the average treatment effect
 - ② Matching. Match each treatment observation to its nearest neighbor(s) based on similar propensity scores.

Stratification: Achieving Balance

- Stratification requires imposing balance by grouping the data and testing for differences in covariate means
 - ① Sort the data by propensity score and divide into groups of observations with similar propensity scores (e.g., percentiles)
 - ② Within each group, test (using a t-test) whether the means of the covariates (X) are equal between treatment and control
 - ③ If so, then stop. If not, it means the covariates aren't balanced *within that group*. Divide the group in half and repeat
 - ④ If a particular covariate is unbalanced for multiple groups, modify the initial logit or probit equation by including higher order terms and/or interactions with that covariate and repeat
- This is done to satisfy the *balancing property*.
- Notice: we never even look at outcomes in this step

Matching strategy and ATT estimation

The standard matching strategy is the following:

- Pair each treatment unit i with one or more *comparable* control group unit j , where comparability is in terms of proximity to the estimated propensity score
- Associate to the treatment unit's outcome Y_i a matched outcome of $Y_{i(j)}$ given by the weighted outcomes of its *neighbors* in the control group

$$Y_{i(j)} = \sum_{j \in C(i)} w_{ij} Y_j$$

where

- $C(i)$ is the set of neighbors with $W = 0$ of the treatment unit i
- and w_{ij} is the weight of control group units j with $\sum_{j \in C(i)} w_{ij} = 1$

Estimation (cont.)

The ATT equals

$$E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1]$$

We estimate it as follows

$$\widehat{ATT} = \frac{1}{N_T} = \sum_{i: W_i=1} \left[Y_i - Y_{i(j)} \right]$$

where N_T is the number of matched treatment units in the sample.

Matching methods

- An estimate of the propensity score is not enough to estimate the ATT
- In fact, the probability of observing two units with *exactly* the same propensity score is in principle zero because $p(x)$ is continuous
- Several matching methods have been proposed in the literature, but the most widely used are:
 - Nearest-neighbor matching (with or without caliper)
 - Radius matching
 - Kernel matching
 - Stratification matching
- Typically, one treatment unit i is matched to several control units j , but sometimes one-to-one matching is used

STATA

- The STATA command `psmatch2` will perform propensity score matching
 - Rich suite of options: nearest neighbor (with or without caliper, with or without replacement), k -nearest neighbors, radius, kernel, local linear regression, and Mahalanobis matching
 - It includes routes for common support graphing (`psgraph`) and covariate imbalance testing (`pstest`)
 - But standard errors are incorrect. They use bootstrapping methods or variance approximation.
 - Has an excellent help file
 - Type in `ssc install psmatch2`

STATA

- Another useful command in STATA is `pscore`
 - Various matching methods for estimating ATT
 - Standard errors are still incorrect: bootstrapping, and variance approximation
 - Balancing tests based on stratification
 - Have to use `net search pscore` to install it.

STATA

- As of STATA 14, built-in suite of options called `teffects`
 - Fewer options. No sampling without replacement
 - Standard errors are correct
 - I'm less familiar with it

Nearest Neighbor

Pretty similar to covariate matching. Formula is

$$ATT^{NN} = \frac{1}{N_T} \sum_{i: W_i=1} \left[Y_i - \sum_{j \in C(i)_M} w_{ij} Y_j \right]$$

- N_T is the number of Treatment units i
- w_{ij} is equal to $\frac{1}{N_C}$ if j is a control unit and zero otherwise; N_C is number of control units j

NN Matching: Bias vs. Variance

- How many nearest neighbors should we use?
 - Matching just one nearest neighbor minimizes bias at the cost of larger variance
 - Matching using additional nearest neighbors increases the bias but decreases the variance
- Matching with or without replacement
 - with replacement keeps bias low at the cost of larger variance
 - without replacement keeps variance low but at the cost of potential bias

Caliper matching

- Caliper matching is a variation on NN matching that tries to build into the algorithm stops against bad matches
- It does this by imposing a tolerable maximum distance (e.g., 0.2 units in the propensity score away from a treatment unit i 's propensity score)
- If there doesn't exist any control group unit j within that "caliper", then treatment unit i is discarded
- It's difficult to know what this caliper should be *ex ante*

Radius matching

Each treatment unit i is matched **only** with the control group unit j whose propensity score falls into a predefined neighborhood of the propensity score of the treatment unit.

- **All** the control units with $\hat{\rho}_j$ falling within a radius r from $\hat{\rho}_i$ are matched to the treatment unit i
- The smaller the radius, the better the quality of the matches, but the higher the possibility some treatment units are not matched because the neighborhood does not contain control group units j

Checking the common support assumption

- We can summarize the propensity scores in the treatment and control group and count how many units are off-support
- For instance, let's say that the treatment group propensity scores range from $[0.1, 0.85]$. That means the min and max scores are in that range.
- We could check what the distribution is for the control group, and drop all the units before 0.1 and above 0.85. In other words, we're throwing away observations that don't satisfy the common support assumption.
- A histogram of propensity scores by treatment and control group also highlights the overlap problem

Examples of propensity score matching

- Example: Job Trainings Program (NSW) (LaLonde 1986, Dehejia and Wahba 1999; 2002)

Description of NSW Job Trainings Program

The National Supported Work Demonstration (NSW), operated by Manpower Demonstration Research Corp in the mid-1970s:

- was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment
- was also unique in that it **randomly assigned** qualified applicants to training positions:
 - **Treatment group:** received all the benefits of NSW program
 - **Control group:** left to fend for themselves
- admitted AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes

NSW Program

- Treatment group members were:
 - guaranteed a job for 9-18 months depending on the target group and site
 - divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance
 - paid for their work
- Other details about the NSW program:
 - Wages: NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance
 - Post-treatment: after their term expired, they were forced to find regular employment
 - Job types: varied within sites – gas station attendant, working at a printer shop – and males and females were frequently performing different kinds of work

NSW Data

- NSW data collection:
 - MDRC collected earnings and demographic information from both treatment and control at baseline and every 9 months thereafter
 - Conducted up to 4 post-baseline interviews
 - Different sample sizes from study to study can be confusing, but has simple explanations
- Estimation:
 - NSW was a randomized job trainings program; therefore estimating the average treatment effect is straightforward:

$$\frac{1}{N_t} \sum_{D_i=1} Y_i - \frac{1}{N_c} \sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$

in large samples assuming treatment selection is independent of potential outcomes (randomization) – i.e., $(Y^0, Y^1) \perp\!\!\!\perp D$.

- NSW worked: Treatment group participants' real earnings post-treatment (1978) was positive and economically meaningful – $\approx \$900$ (LaLonde 1986) to $\$1,800$ (Dehejia and Wahba 2002) depending on the sample used

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*.

- LaLonde's study: was **not** an evaluation of the NSW program, but rather an evaluation of econometric models done by:
 - replacing the experimental NSW control group with non-experimental control group drawn from two public use datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)
 - estimating the average effect using non-experimental workers as controls for the NSW trainees
 - comparing his non-experimental estimates to the experimental estimates of \$900
- LaLonde's conclusion: available econometric approaches were biased and inconsistent
 - His estimates were way off and usually the wrong sign
 - Conclusion was influential in policy circles and led to greater push for more experimental evaluations

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975-78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings Growth 1975-78 Treatments Less Comparisons		Unrestricted Difference in Differences: Quasi Difference in Earnings Growth 1975-78		Controlling for All Observed Variables and Pre-Training Earnings (10)
		Pre-Training Year, 1975		Post-Training Year, 1978		Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted ^c (9)	
		Unadjusted (2)	Adjusted ^c (3)	Unadjusted (4)	Adjusted ^c (5)					
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)
PSID-3	(\$3,322 (780)	(\$455 (539)	\$455 (704)	\$697 (760)	-\$509 (967)	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)

^aThe columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^bEstimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^cThe exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^dSee Table 3 for definitions of the comparison groups.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA^{a,b}

Name of Comparison Group ^d	Comparison Group Earnings Growth 1975-78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences:		Unrestricted Difference in Earnings Growth 1975-78 Quasi Difference in Earnings Growth 1975-78	Controlling for All Observed Variables and Pre-Training Earnings (10)		
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975-78 Treatments Less Comparisons					
		Unadjusted (2)	Adjusted ^c (3)	Unadjusted (4)	Adjusted ^c (5)	Without Age (6)	With Age (7)				
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780)	(\$455 (539)	\$455 (704)	\$697 (760)	-\$509 (967)	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

^aThe columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

^bEstimates are in 1982 dollars. The numbers in parentheses are the standard errors.

^cThe exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

^dSee Table 3 for definitions of the comparison groups.

Imbalanced covariates for experimental and non-experimental samples

covariate	All		CPS	NSW	t-stat	diff
	mean	(s.d.)	Controls	Trainees		
			$N_c = 15,992$	$N_t = 297$		
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

Dehejia and Wahba (1999)

- Dehejia and Wahba (DW) update LaLonde's original study using propensity score matching
 - ① Dehejia, Rajeev H. and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". Journal of the American Statistical Association, vol. 94(448): 1053-1062 (pdf)
- Can propensity score matching improve over the estimators that LaLonde examined?

Table 1. Sample Means of Characteristics for NSW and Comparison Samples

	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. \$)	RE75 (U.S. \$)
NSW/Lalonde:^a									
Treated	297	24.63 (.32)	10.38 (.09)	.80 (.02)	.09 (.01)	.73 (.02)	.17 (.02)	3,066 (236)	
Control	425	24.45 (.32)	10.19 (.08)	.80 (.02)	.11 (.02)	.81 (.02)	.16 (.02)	3,026 (252)	
RE74 subset:^b									
Treated	185	25.81 (.35)	10.35 (.10)	.84 (.02)	.059 (.01)	.71 (.02)	.19 (.02)	2,096 (237)	1,532 (156)
Control	260	25.05 (.34)	10.09 (.08)	.83 (.02)	.1 (.02)	.83 (.02)	.15 (.02)	2,107 (276)	1,267 (151)
Comparison groups:^c									
PSID-1	2,490	34.85 [.78]	12.11 [.23]	.25 [.03]	.032 [.01]	.31 [.04]	.87 [.03]	19,429 [991]	19,063 [1,002]
PSID-2	253	36.10 [1.00]	10.77 [.27]	.39 [.04]	.067 [.02]	.49 [.05]	.74 [.04]	11,027 [853]	7,569 [695]
PSID-3	128	38.25 [1.17]	10.30 [.29]	.45 [.05]	.18 [.03]	.51 [.05]	.70 [.05]	5,566 [686]	2,611 [499]
CPS-1	15,992	33.22 [.81]	12.02 [.21]	.07 [.02]	.07 [.02]	.29 [.03]	.71 [.03]	14,016 [705]	13,650 [682]
CPS-2	2,369	28.25 [.87]	11.24 [.19]	.11 [.02]	.08 [.02]	.45 [.04]	.46 [.04]	8,728 [667]	7,397 [600]
CPS-3	429	28.03 [.87]	10.23 [.23]	.21 [.03]	.14 [.03]	.60 [.04]	.51 [.04]	5,619 [552]	2,467 [288]

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; RE74 = earnings in calendar year 1974.

^a NSW sample as constructed by Lalonde (1986).

^b The subset of the Lalonde sample for which RE74 is available.

^c Definition of comparison groups (Lalonde 1986):

PSID-1: All male household heads under age 55 who did not classify themselves as retired in 1975.

PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.

PSID-3: Selects from PSID-2 all men who were not working in 1975.

CPS-1: All CPS males under age 55.

CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.

CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

PSID1-3 and CPS-1 are identical to those used by Lalonde. CPS2-3 are similar to those used by Lalonde, but Lalonde's original subset could not be recreated.

Table 2. Lalonde's Earnings Comparisons and Estimated Training Effects for the NSW Male Participants Using Comparison Groups From the PSID and the CPS^a

A. Lalonde's original sample				B. RE74 subsample (results do not use RE74)				C. RE74 subsample (results use RE74)					
NSW	Unrestricted differences in earnings less comparison group earnings	NSW	Unrestricted differences in earnings less comparison group earnings	NSW	Unrestricted differences in earnings less comparison group earnings	NSW	Unrestricted differences in earnings less comparison group earnings	NSW	Unrestricted differences in earnings less comparison group earnings	NSW	Unrestricted differences in earnings less comparison group earnings		
treatment	differences in earnings less comparison group earnings	treatment	differences in earnings less comparison group earnings	treatment	differences in earnings less comparison group earnings	treatment	differences in earnings less comparison group earnings	treatment	differences in earnings less comparison group earnings	treatment	differences in earnings less comparison group earnings		
1978		1975-1978		1978		1975-1978		1978		1975-1978			
Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables			
Comparison group	Unadjusted ^b	Adjusted ^c	Unadjusted ^d	Adjusted ^e	Unadjusted ^b	Adjusted ^c	Unadjusted ^d	Adjusted ^e	Unadjusted ^b	Adjusted ^c	Unadjusted ^d	Adjusted ^e	
	(1)	(2)	(3)	(4)	(5)	(6)	(3)	(4)	(5)	(6)	(3)	(4)	
NSW	886 (472)	798 (472)	879 (467)	802 (468)	820 (633)	1,794 (637)	1,672 (632)	1,750 (637)	1,631 (639)	1,612 (633)	1,794 (636)	1,688 (632)	1,750 (638)
PSID-1	-15,578 (913)	-8,057 (990)	-2,380 (680)	-2,119 (746)	-1,844 (762)	-15,205 (1155)	-7,741 (1175)	-562 (841)	-265 (881)	188 (901)	-15,205 (1155)	-879 (931)	-582 (841) (866)
PSID-2	-4,020 (781)	-3,482 (935)	-1,364 (729)	-1,694 (878)	-1,876 (885)	-3,647 (960)	-2,810 (1082)	721 (886)	298 (1004)	111 (1032)	-3,647 (960)	94 (1042)	721 (886) (1004)
PSID-3	697 (760)	-559 (967)	629 (757)	-552 (967)	-576 (968)	1,070 (900)	35 (1101)	1,370 (897)	243 (1101)	298 (1105)	1,070 (900)	821 (1100)	1,370 (897) (1101)
CPS-1	-8,870 (562)	-4,416 (577)	-1,543 (426)	-1,102 (450)	-987 (452)	-8,498 (712)	-4,417 (714)	-78 (537)	525 (557)	709 (560)	-8,498 (712)	-8 (572)	-78 (537) (547)
CPS-2	-4,195 (533)	-2,341 (620)	-1,649 (459)	-1,129 (551)	-1,149 (671)	-3,822 (746)	-2,206 (574)	-263 (662)	371 (666)	305 (671)	-3,822 (672)	615 (574)	-263 (574) (654)
CPS-3	-1,008 (539)	-1 (681)	-1,204 (532)	-263 (677)	-234 (675)	-635 (657)	375 (821)	-91 (641)	844 (808)	875 (810)	-635 (657)	1,270 (798)	-91 (641) (796)

NOTES: Panel A replicates the sample of Lalonde (1986, table 5). The estimates for columns (1)-(4) for NSW, PSID-1-3, and CPS-1 are identical to Lalonde's. CPS-2 and CPS-3 are similar but not identical, because we could not exactly recreate his subset. Column (5) differs because the observations did not contain all of the covariates used in column (10) of Lalonde's table 5.

^b Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

^c The estimates based on the NSW control groups are unbiased estimates of the treatment impacts for the original sample (8886) and for the RE74 sample (\$1,794).

^d The exogenous variables used in the regressions-adjusted equations are age, age squared, years of schooling, high school dropout status, and race (and RE74 in Panel C).

^e Regresses RE78 on a treatment indicator and RE75.

^f The same as (d), but controls for the additional variables listed under (c).

^g Controls for all pretreatment covariates.

Proposition 2

$$X \perp\!\!\!\perp D|p(X)$$

- Conditional on the propensity score, the covariates are independent of the treatment, suggesting that the distribution of covariate values should be the same for both treatment and control groups

Trimming the data

- “Trimming” throws away units from control which do not overlap with the treatment in terms of estimated propensity score
- In STATA:
 - . `su pscore if treat==1`
 - . `drop if pscore<r(mean) & treat==0`

Overlap

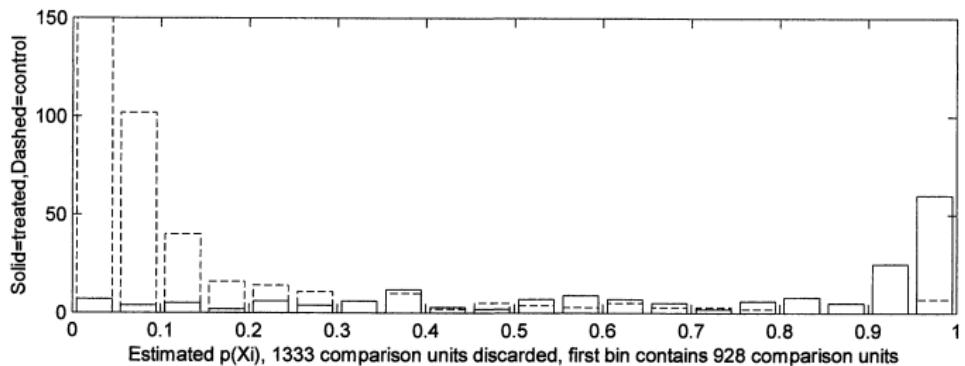


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The PSID units. There is minimal overlap between the two groups. Three bins (.8-.85, .85-.9, and .9-.95) contain no comparison treated units with an estimated propensity score greater than .8 and only 7 comparison units.

Overlap

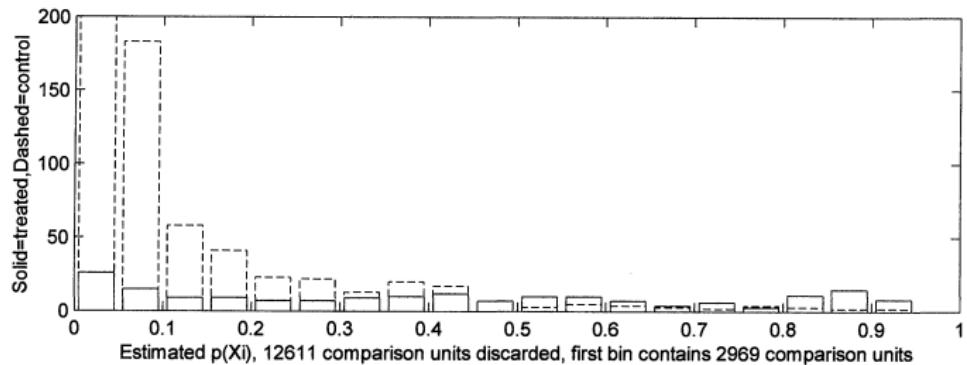


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12, estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

Results

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID a

	NSW earnings less comparison group earnings		NSW treatment earnings less comparison group earnings conditional on the estimated propensity score				Matched Unadjusted (1)	
			Quadratic in score ^b (3)	Stratifying on the score				
	(1) Unadjusted	(2) Adjusted ^a		(4) Unadjusted	(5) Adjusted	(6) Observations ^c		
NSW	1,794 (633)	1,672 (638)						
PSID-1 ^e	-15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255	1,622 (2,211)	
PSID-2 ^f	-3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,235 (1,793)	389	1,440 (2,211)	
PSID-3 ^f	1,069 (899)	825 (1,104)	647 (1,383)	2,321 (1,994)	1,870 (2,002)	247	2,111 (2,211)	
CPS-1 ^g	-8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117	1,541 (1,041)	
CPS-2 ^g	-3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493	1,711 (1,211)	
CPS-3 ^g	-635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514	5,141 (1,411)	

^a Least squares regression: RE78 on a constant, a treatment indicator, age, age², education, no degree, black, Hispanic, RE74, RE75.

^b Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note d.

^c Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

^d Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation.

Propensity scores are estimated using the logistic model, with specifications as follows:

^e PSID-1: Prob ($T_i = 1$) = $F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{married}, \text{no degree}, \text{black}, \text{Hispanic}, \text{RE74}, \text{RE75}, \text{RE74}^2, \text{RE75}^2, \text{u74*black})$.

^f PSID-2 and PSID-3: Prob ($T_i = 1$) = $F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{no degree}, \text{married}, \text{black}, \text{Hispanic}, \text{RE74}, \text{RE74}^2, \text{RE75}, \text{RE75}^2, \text{u74}, \text{u75})$.

^g CPS-1, CPS-2, and CPS-3: Prob ($T_i = 1$) = $F(\text{age}, \text{age}^2, \text{education}, \text{education}^2, \text{no degree}, \text{married}, \text{black}, \text{Hispanic}, \text{RE74}, \text{RE75}, \text{u74}, \text{u75}, \text{education}^2, \text{RE74}, \text{RE75})$.

Covariate balance

Table 4. Sample Means of Characteristics for Matched Control Samples

Matched samples	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S.)
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096
MPSID-1	56	26.39 [2.56]	10.62 [.63]	.86 [.13]	.02 [.06]	.55 [.13]	.15 [.12]	1,794 [1,406]
MPSID-2	49	25.32 [2.63]	11.10 [.83]	.89 [.14]	.02 [.08]	.57 [.16]	.19 [.16]	1,599 [1,905]
MPSID-3	30	26.86 [2.97]	10.96 [.84]	.91 [.13]	.01 [.08]	.52 [.16]	.25 [.16]	1,386 [1,680]
MCPS-1	119	26.91 [1.25]	10.52 [.32]	.86 [.06]	.04 [.04]	.64 [.07]	.19 [.06]	2,110 [841]
MCPS-2	87	26.21 [1.43]	10.21 [.37]	.85 [.08]	.04 [.05]	.68 [.09]	.20 [.08]	1,758 [896]
MCPS-3	63	25.94 [1.68]	10.69 [.48]	.87 [.09]	.06 [.06]	.53 [.10]	.13 [.09]	2,709 [1,285]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.

MPSID1-3 and MCPS1-3 are the subsamples of PSID1-3 and CPS1-3 that are matched to the treatment group.

Estimation in STATA

- We will replicate some of DW (1999) using the experimental treatment and non-experimental PSID sample
- You will need to download the following zipped dataset from Barbara Sianesi's presentation given to the 2010 German STATA Users Group in Berlin.
- Note you will need to install psmatch2

```
ssc install psmatch2, replace
```

```

. use ./nsw_psid, clear
(NSW treated and PSID non-treated)

. qui probit treated age black hispanic married educ nodegree re75

. margins, dydx(_all)

```

Average marginal effects Number of obs = 2787
 Model VCE : OIM

Expression : Pr(treated), predict()
 dy/dx w.r.t. : age black hispanic married education nodegree re75

	Delta-method					
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
age	-.0035844	.000462	-7.76	0.000	-.0044899	-.002679
black	.0766501	.0088228	8.69	0.000	.0593577	.0939426
hispanic	.0831734	.0157648	5.28	0.000	.0522751	.1140718
married	-.0850743	.0070274	-12.11	0.000	-.0988478	-.0713009
education	.0003458	.0023048	0.15	0.881	-.0041716	.0048633
nodegree	.0418875	.0108642	3.86	0.000	.0205942	.0631809
re75	-6.89e-06	5.89e-07	-11.71	0.000	-8.04e-06	-5.74e-06

```

. // compute the propensity score
. predict double score
(option pr assumed; Pr(treated))

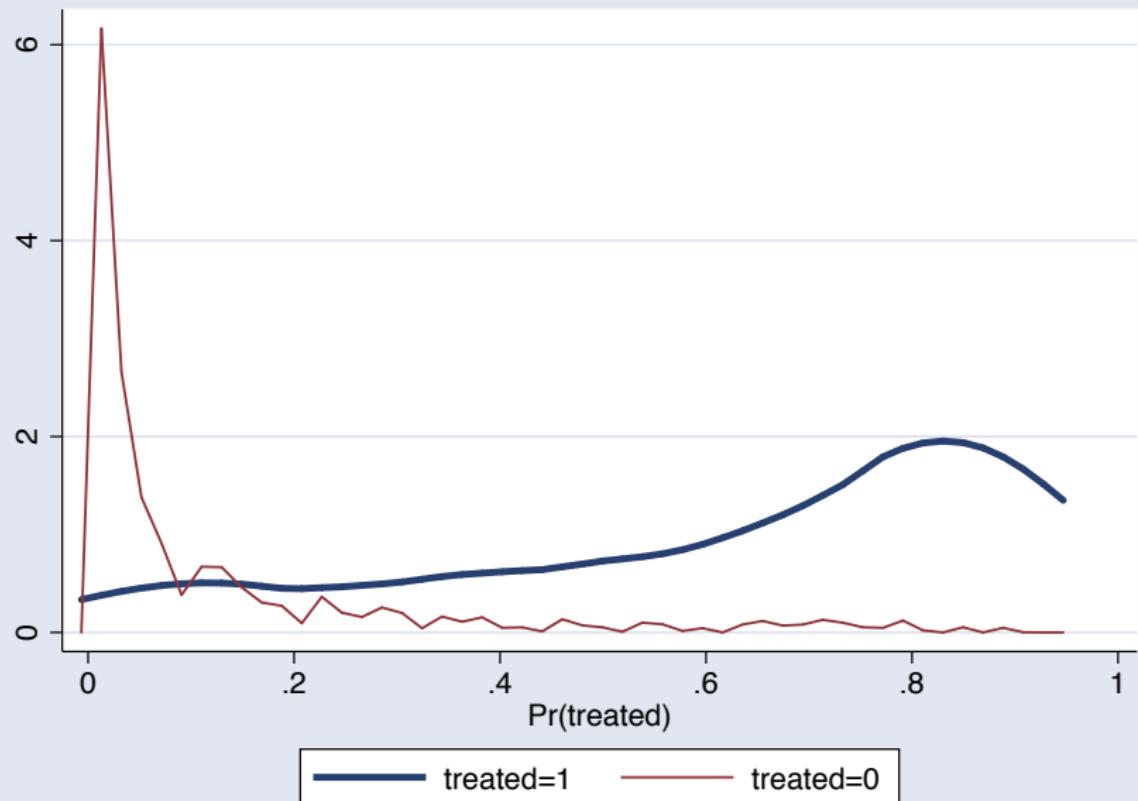
```

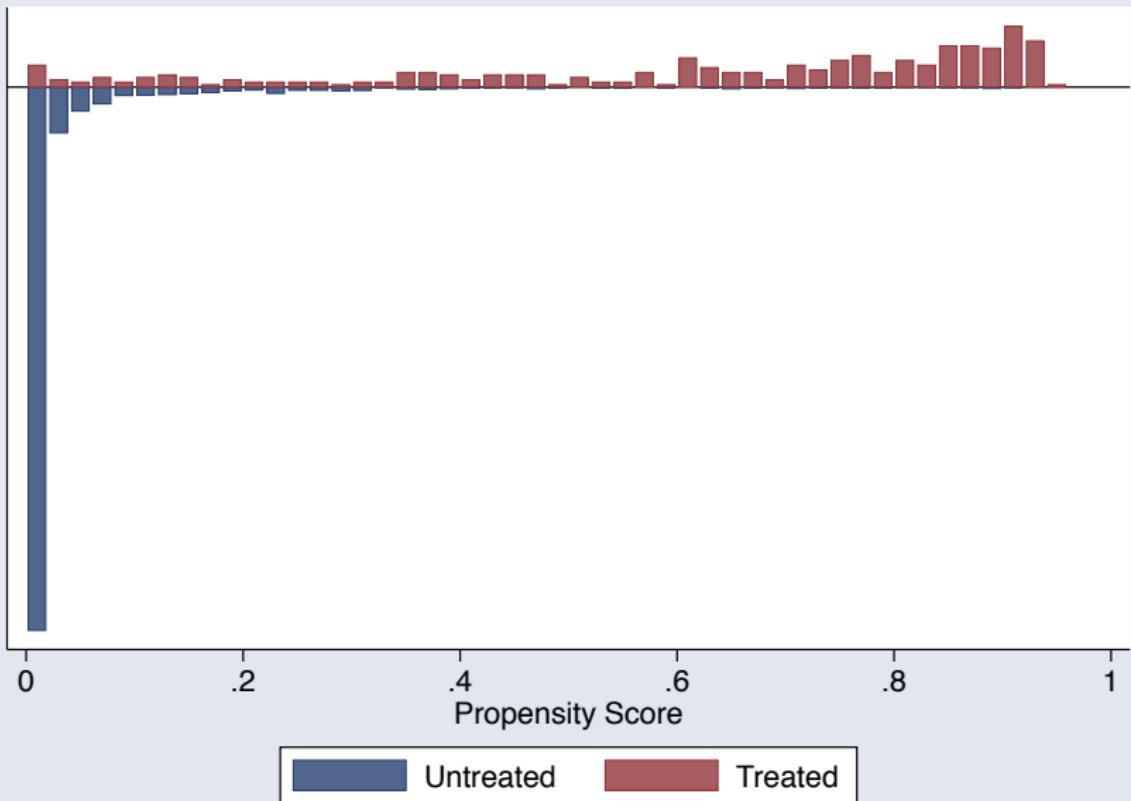
```
. // compare the densities of the estimated propensity score over groups
. density2 score, group(treated) saving("./psm2a, replace)
(file ./psm2a saved)

. graph export psm2a.pdf, replace
(file psm2a.pdf written in PDF format)

. psgraph, treated(treated) pscore(score) bin(50) saving(psm2b, replace)
(file psm2b.gph saved)

. graph export psm2b.pdf, replace
(file psm2b.pdf written in PDF format)
```





```
. // compute nearest neighbor matching with caliper and replacement
. psmatch2 treated, pscore(score) outcome(re78) caliper(0.01)
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
```

Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
<hr/>						
re78	Unmatched	5976.35202	21553.9209	-15577.5689	913.328457	-17.06
	ATT	6067.8117	5758.47686	309.334834	1080.935	0.29

Note: S.E. does not take into account that the propensity score is estimated.

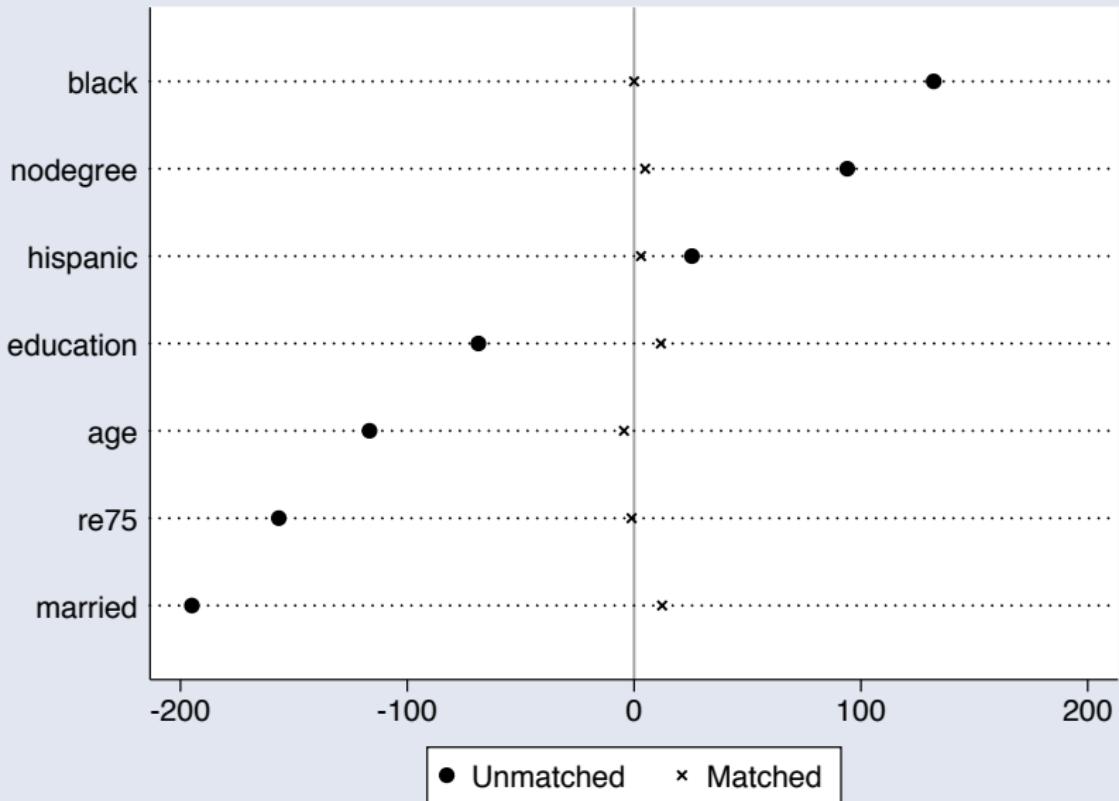
psmatch2: psmatch2: Common			
Treatment	support		
assignment	Off suppo	On suppor	Total
Untreated	0	2,490	2,490
Treated	26	271	297
Total	26	2,761	2,787

```
. //evaluate common support
. summarize _support if treated
```

Variable	Obs	Mean	Std. Dev.	Min	Max
<hr/>					
_support	297	.9124579	.2831048	0	1

```
. //evaluate quality of matching
. ptest2 age black hispanic married educ nodegree re75, sum graph
```

Variable	Sample	Mean		%reduct		t-test	
		Treated	Control	%bias	bias	t	p> t
age	Unmatched	24.626	34.851	-116.6		-16.48	0.000
	Matched	25.052	25.443	-4.5	96.2	-0.61	0.545
black	Unmatched	.80135	.2506	132.1		20.86	0.000
	Matched	.78967	.78967	0.0	100.0	-0.00	1.000
hispanic	Unmatched	.09428	.03253	25.5		5.21	0.000
	Matched	.09594	.08856	3.0	88.0	0.30	0.767
married	Unmatched	.16835	.86627	-194.9		-33.02	0.000
	Matched	.1845	.14022	12.4	93.7	1.40	0.163
education	Unmatched	10.38	12.117	-68.6		-9.51	0.000
	Matched	10.465	10.166	11.8	82.8	1.54	0.125
nodegree	Unmatched	.73064	.30522	94.0		15.10	0.000
	Matched	.71587	.69373	4.9	94.8	0.56	0.573
re75	Unmatched	3066.1	19063	-156.6		-20.12	0.000
	Matched	3197.4	3307.8	-1.1	99.3	-0.28	0.778



Kernel matching

- Alternatively we can perform propensity score matching with a kernel-based method.
- Notice on the next slide that the estimate of the ATT switches sign relative to that produced by the NN matching algorithm

```
. // compute kernel-based matching with normal kernel
. psmatch2 treated, pscore(score) outcome(re78) kernel k(normal) bw(0.01)
```

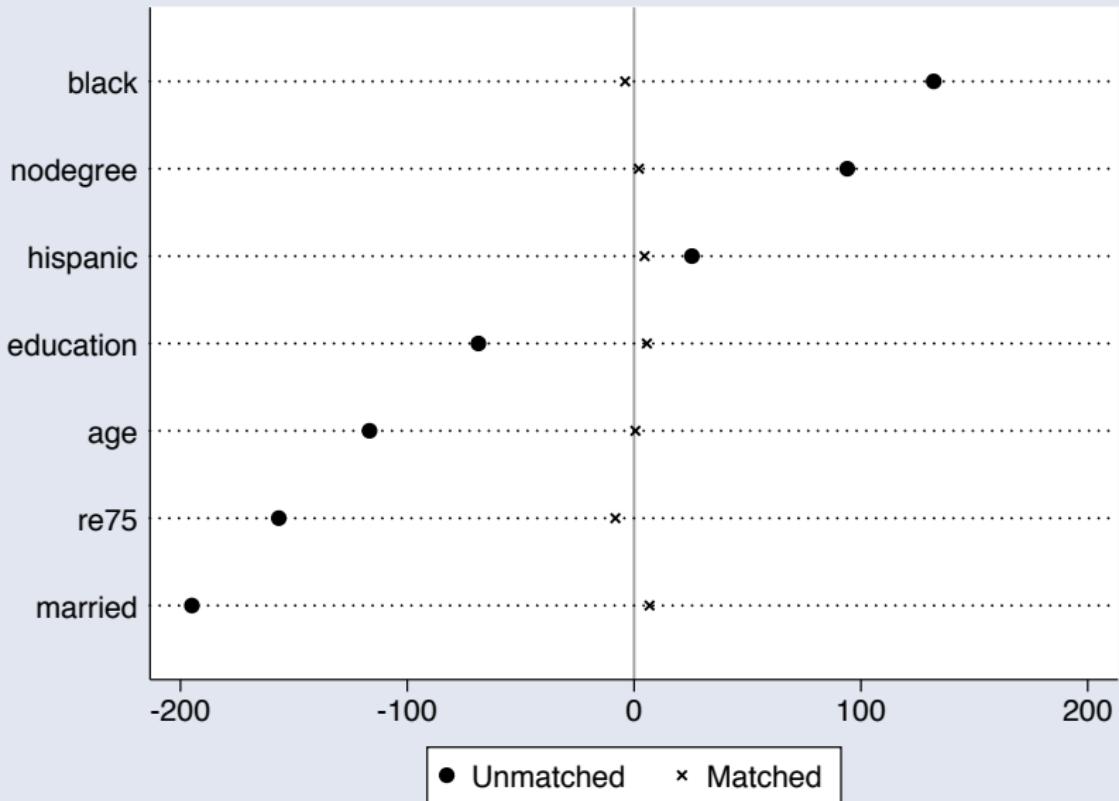
Variable	Sample	Treated	Controls	Difference	S.E.	T-stat
<hr/>						
re78	Unmatched	5976.35202	21553.9209	-15577.5689	913.328457	-17.06
	ATT	5976.35202	6882.18396	-905.831935	2151.26377	-0.42

Note: S.E. does not take into account that the propensity score is estimated.

	psmatch2:
psmatch2:	Common
Treatment	support
assignment	On support
<hr/>	
Untreated	2,490
Treated	297
<hr/>	
Total	2,787

```
. //evaluate quality of matching
. ptest2 age black hispanic married educ nodegree re75, sum graph
```

Variable	Sample	Mean		%reduct		t-test	
		Treated	Control	%bias	bias	t	p> t
age	Unmatched	24.626	34.851	-116.6		-16.48	0.000
	Matched	24.626	24.572	0.6	99.5	0.09	0.926
black	Unmatched	.80135	.2506	132.1		20.86	0.000
	Matched	.80135	.81763	-3.9	97.0	-0.50	0.614
hispanic	Unmatched	.09428	.03253	25.5		5.21	0.000
	Matched	.09428	.08306	4.6	81.8	0.48	0.631
married	Unmatched	.16835	.86627	-194.9		-33.02	0.000
	Matched	.16835	.1439	6.8	96.5	0.82	0.413
education	Unmatched	10.38	12.117	-68.6		-9.51	0.000
	Matched	10.38	10.238	5.6	91.8	0.81	0.415
nodegree	Unmatched	.73064	.30522	94.0		15.10	0.000
	Matched	.73064	.72101	2.1	97.7	0.26	0.793
re75	Unmatched	3066.1	19063	-156.6		-20.12	0.000
	Matched	3066.1	3905.8	-8.2	94.8	-1.99	0.047



Matchings vs. Propensity score

Table 2. Experimental and nonexperimental estimates for the NSW data

	$M = 1$		$M = 4$		$M = 16$		$M = 64$		$M = 2490$	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Panel A:										
Experimental estimates										
Covariate matching	1.22	(0.84)	1.99	(0.74)	1.75	(0.74)	2.20	(0.70)	1.79	(0.67)
Bias-adjusted cov matching	1.16	(0.84)	1.84	(0.74)	1.54	(0.75)	1.74	(0.71)	1.72	(0.68)
Pscore matching	1.43	(0.81)	1.95	(0.69)	1.85	(0.69)	1.85	(0.68)	1.79	(0.67)
Bias-adjusted pscore matching	1.22	(0.81)	1.89	(0.71)	1.78	(0.70)	1.67	(0.69)	1.72	(0.68)
Regression estimates										
Mean difference	1.79	(0.67)								
Linear	1.72	(0.68)								
Quadratic	2.27	(0.80)								
Weighting on pscore	1.79	(0.67)								
Weighting and linear regression	1.69	(0.66)								
Panel B:										
Nonexperimental estimates										
Simple matching	2.07	(1.13)	1.62	(0.91)	0.47	(0.85)	-0.11	(0.75)	-15.20	(0.61)
Bias-adjusted matching	2.42	(1.13)	2.51	(0.90)	2.48	(0.83)	2.26	(0.71)	0.84	(0.63)
Pscore matching	2.32	(1.21)	2.06	(1.01)	0.79	(1.25)	-0.18	(0.92)	-1.55	(0.80)
Bias-adjusted pscore matching	3.10	(1.21)	2.61	(1.03)	2.37	(1.28)	2.32	(0.94)	2.00	(0.84)
Regression estimates										
Mean difference	-15.20	(0.66)								
Linear	0.84	(0.88)								
Quadratic	3.26	(1.04)								
Weighting on pscore	1.77	(0.67)								
Weighting and linear regression	1.65	(0.66)								

NOTE: The outcome is earnings in 1978 in thousands of dollars.

Conclusions

- Propensity scores are awesome to check the balance and overlap of covariates – an under appreciated diagnostic, and one that you won't probably do if you only run regressions
- There are extensions for more than two treatments (Cattaneo 2010)
- The propensity score can make groups comparable **but** only on the variables used to estimate the propensity score in the first place. There is **NO** guarantee you are balancing on unobserved covariates.
- If you know that there are important unobservable variables, you may need another tool.
- Randomization ensure that both observable and unobservable variables are balanced

Consider this

One thing to ruminate over is that in invoking CIA, the researcher is basically saying that there exists some unknown randomness determining treatment once she conditions on X . But if there is some randomization,

- ① how does she know that exactly, and
- ② why not use that randomization itself as the treatment, such as one often does in a reduced form instrumental variables approach (more to come on that)? This is basically what Krueger (1999) does when he regresses test scores onto *original room assignment* (rather than actual room assignment).

HT to Rodney Andrews for making this point to me in a conversation

Conclusions

- Think hard about this question: *Do you really need a propensity score analysis? Why do you believe that CIA holds?*
- Why not use the conditional randomization process as the explanatory variable? If you can't do that, how confident are you in CIA?

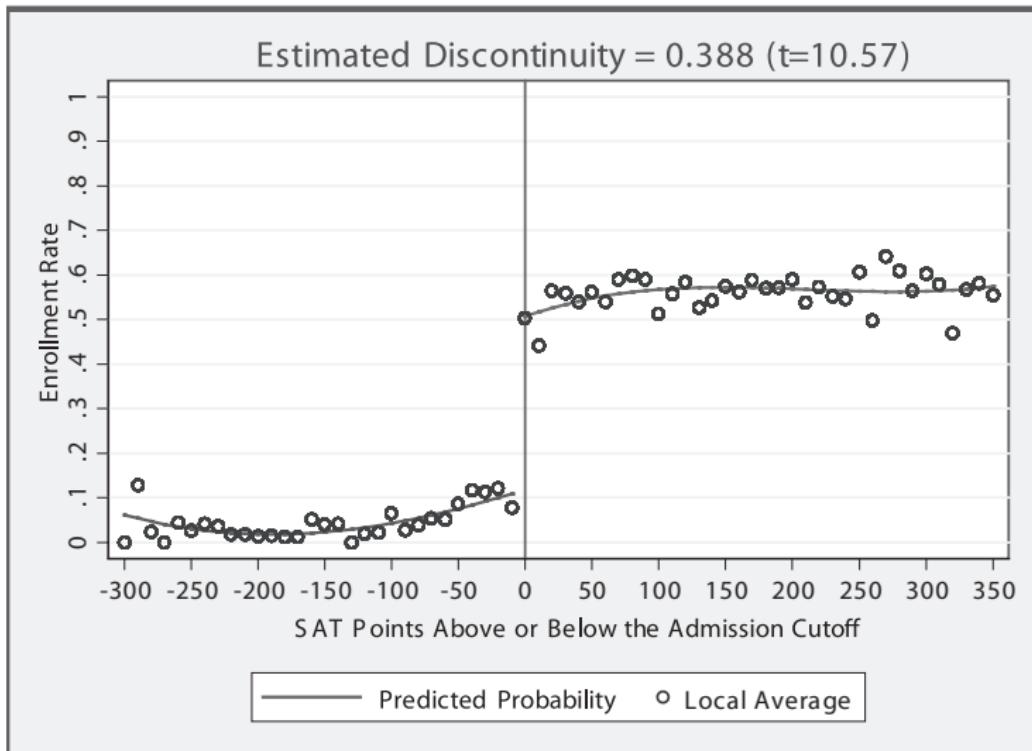
What is regression discontinuity design?

Recently there has been an increase in a particular type of research design known as *regression discontinuity design* (RDD). Cook (2008) has a fascinating history of thought on how and why.

- Donald Campbell is the originator of regression discontinuity design. First study is Thistlethwaite and Campbell (1960). Merit awards were given to students whose test scores were over some cutoff point. They compared award winners to non-winners just around the threshold to identify the causal effect of merit awards on future academic outcomes.
- Pictures are helpful for understanding the RDD research design. Tell me what you think these are saying.

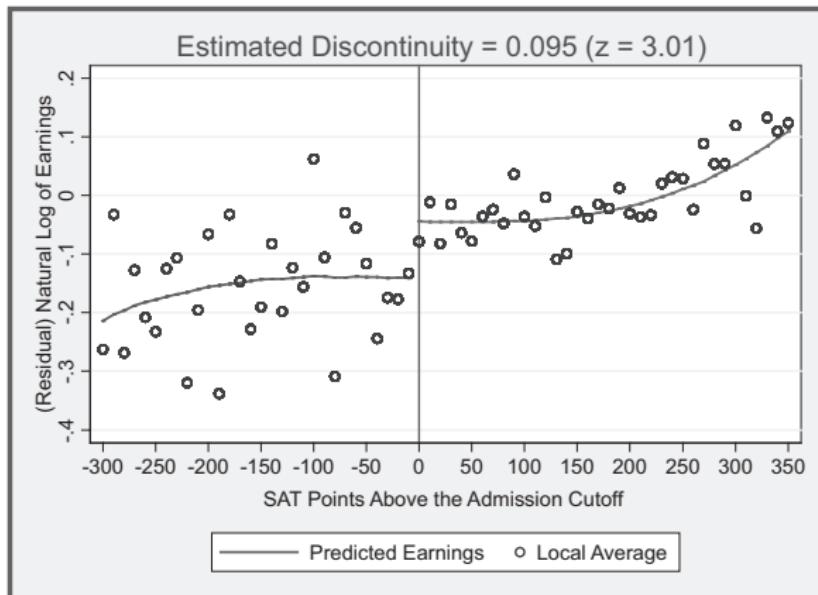
RDD Visual Example

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



RDD Visual Example

FIGURE 2.—NATURAL LOG OF ANNUAL EARNINGS FOR WHITE MEN TEN TO FIFTEEN YEARS AFTER HIGH SCHOOL GRADUATION (FIT WITH A CUBIC POLYNOMIAL OF ADJUSTED SAT SCORE)



What is a regression discontinuity design?

- RDD is based on a simple idea: the assignment of some treatment to observational units is completely known.
 - We know that the probability of treatment “jumps” when an applicant’s test scores meet university’s admission rule
 - If the students accepted and the students rejected are similar near the cutoff point, then data can be analyzed as though it were a *conditionally randomized experiment* .
- RDD is appropriate in situations where treatment assignment is sharply discontinuous in the values of a variable
- It’s best understood with an example. We’ll look at Angrist and Lavy (1999) which was one of a few papers around the same time to bring this methodology into economics

Angrist, Joshua D. and Victor Lavy (1999). "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement". *Quarterly Journal of Economics*.

Summary

- Angrist and Lavy (1999) are interested in estimating the causal effect of *class size* on *educational outcomes*
- So like Krueger (1999), Angrist and Lavy (1999) are interested in the effect of class size on pupil achievement.

Maimonides Rule and class size

- One of the earliest references to class size is the Babylonian Talmud, completed around 6th century, which discusses rules for the determination of class size and pupil-teacher ratios in bible study.
- Maimonides was a 12th century Rabbinic scholar who interpreted the Talmud's discussion of class size as follows
 - “Twenty-five children may be put in charge of one teacher. If the number in the class exceeds twenty-five but is not more than forty, two teachers must be appointed.”

Maimonedes Rule and Class size

- “The importance of Maimonedes’ rule for our purposes is that since 1969, it has been used to determine the division of enrollment cohorts into classes in Israeli public schools.”
(Angrist and Lavy 1999)

Identification

- Class size in most places and most times is strongly associated with **unobservable** differences in variables that also predict educational performance
 - Poverty, affluence, enthusiasm/skepticism about the value of education, special needs of students for remedial or advanced instruction, obscure and barely intelligible obsessions of bureaucracies
 - Each of these determines (i) class size and (ii) clouds the effect of class size on academic performance because (iii) each is also correlated independently with academic performance
- However—if adherence to Maimonides' rule is perfectly rigid, then what would separate a school with a single class of size 40 from the same school with two classes whose average size is 20.5 ($\frac{41}{2} = 20.5$) is **the enrollment of a single student**

Exogenous variation in class size

- Maimonides' rule has the largest impact on a school with about 40 students in a grade cohort
- With cohorts of size 40, 80 and 120 students, the steps down in average class size required by Maimonides' rule when an additional student enrolls are from 40 to 20.5 ($\frac{41}{2}$), 40 to 27 ($\frac{81}{3}$), and 40 to 30.25 ($\frac{121}{4}$)
- School also use the percent disadvantaged in a school "to allocate supplementary hours of instruction and other school resources" which is why Angrist and Lavy control for it

Data

- Test data came from short-lived national testing program in Israeli elementary schools
 - June 1991 and 1992 near end of school year all 4-5th graders were given achievement tests to measure math and (Hebrew) reading skills
- Average math and reading test scores (scaled from 1-100) linked with data on school characteristics and class size from other sources
 - Beginning of year (BOY) enrollment data was collected and linked
 - Class size was collected around the time of the testing and obtained from administrative source

Data

- The unit of observation in the linked data sets is the class (remember the example comparing micro and aggregate data regressions)
 - linked class-level data includes average test scores in each class, spring class size, BOY enrollment for each school and grade, a town identifier, school-level index of student socioeconomic status called “percent disadvantaged” (PD) and variables identifying the ethnic and religious composition of the school
- Study was limited to Jewish public schools, both secular and religious schools, which account for the vast majority of school children in Israel
 - Arab schools weren’t given the tests; ultra-orthodox Jewish schools were omitted because they use a curriculum that differs considerably from public schools

TABLE I
UNWEIGHTED DESCRIPTIVE STATISTICS

Variable	Mean	S.D.	Quantiles						
			0.10	0.25	0.50	0.75	0.90		
A. Full sample									
5th grade (2019 classes, 1002 schools, tested in 1991)									
Class size	29.9	6.5	21	26	31	35	38		
Enrollment	77.7	38.8	31	50	72	100	128		
Percent disadvantaged	14.1	13.5	2	4	10	20	35		
Reading size	27.3	6.6	19	23	28	32	36		
Math size	27.7	6.6	19	23	28	33	36		
Average verbal	74.4	7.7	64.2	69.9	75.4	79.8	83.3		
Average math	67.3	9.6	54.8	61.1	67.8	74.1	79.4		

Discontinuity sample

- Maimonides' rule can be used to identify the effects of class size because the rule induces a discontinuity in the relationship between enrollment and class size at enrollment multiples of 40
- Angrist and Lavy (1999) present descriptive statistics for one such “discontinuity sample” defined as only schools with enrollments \pm 5 students: 36,45; 76,85; 116,125
- Slightly fewer than 25% of classes come from schools with enrollments of this range
- Average class size is a bit larger in this \pm discontinuity sample than in the overall sample
- But otherwise remarkably similar to the full sample

B. $+\/-5$ Discontinuity sample (enrollment 36–45, 76–85, 116–124)

	5th grade		4th grade		3rd grade	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
	(471 classes, 224 schools)		(415 classes, 195 schools)		(441 classes, 206 schools)	
Class size	30.8	7.4	31.1	7.2	30.6	7.4
Enrollment	76.4	29.5	78.5	30.0	75.7	28.2
Percent disadvantaged	13.6	13.2	12.9	12.3	14.5	14.6
Reading size	28.1	7.3	28.3	7.7	24.6	6.2
Math size	28.5	7.4	28.7	7.7	24.8	6.3
Average verbal	74.5	8.2	72.5	7.8	86.2	6.3
Average math	67.0	10.2	68.7	9.1	84.2	7.0

Class size function and enrollment size

$$f_{sc} = \frac{e_s}{\text{int} \frac{e_s - 1}{40} + 1} \quad (90)$$

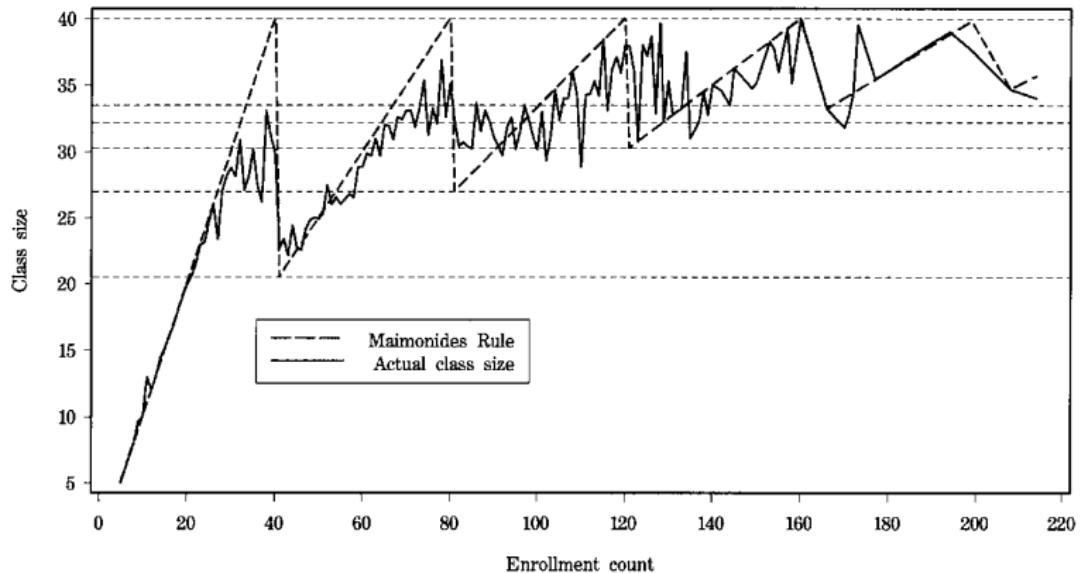
where e_s is the beginning-of-year enrollment in school s in a given grade (e.g., 5th grade); f_{sc} is class size assigned to class c in school s for that grade; $\text{int}(n)$ is the largest integer less than or equal to n

- This equation captures the fact that Maimonides' rule allows enrollment cohorts of 1-40 to be grouped in a single class, but enrollment cohorts of 41-80 are split into two classes of average size 20.5-40, enrollment cohorts of 81-120 are split into three classes of average size 27-40, and so on.

Class size function and enrollment size

- Although f_{sc} is fixed within schools, in practice enrollment cohorts are not necessarily divided into classes of equal size
- But, even though the actual relationship between class size and enrollment size involves many factors, in Israel it clearly has a lot to do with f_{sc} as we'll see

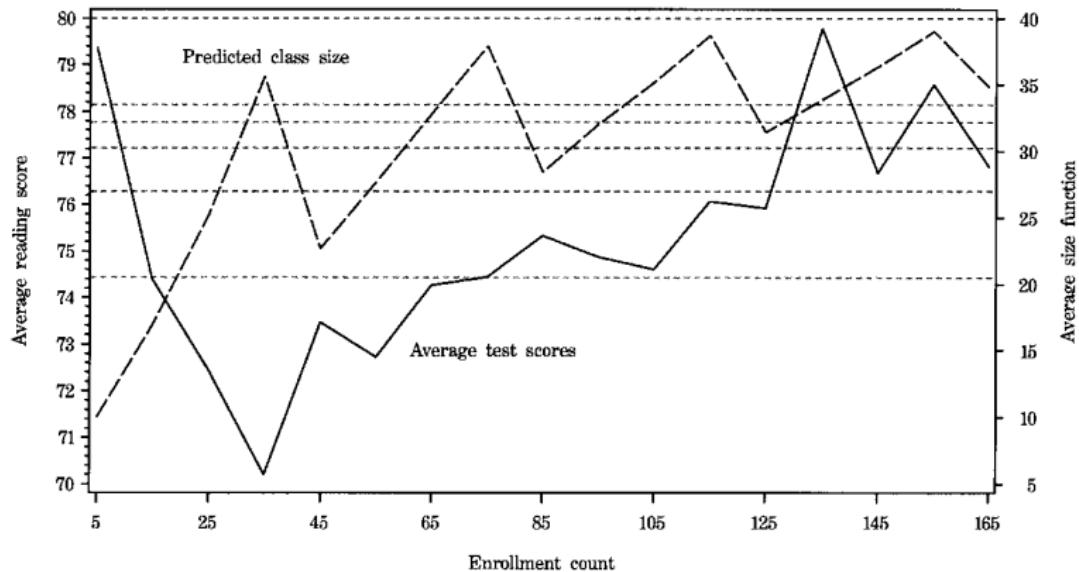
a. Fifth Grade



Class size function and test scores

- The class size function is correlated with average class size as well as average test scores of fourth and fifth graders
- The following picture plots average reading test scores and average values of f_{sc} by enrollment size in enrollment intervals of ten for fifth graders
- The figure shows that test scores are generally higher in schools with larger enrollments and larger predicted class sizes
- But, it also shows an up-and-down pattern in which average scores by enrollment size mirror the class-size function

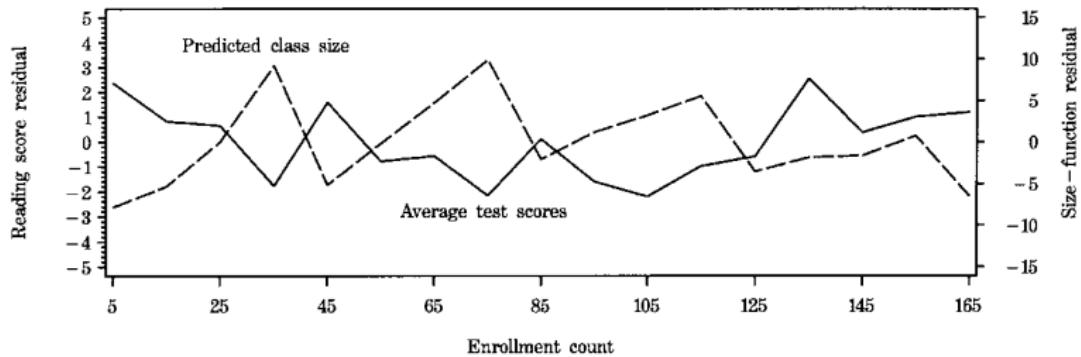
a. Fifth Grade



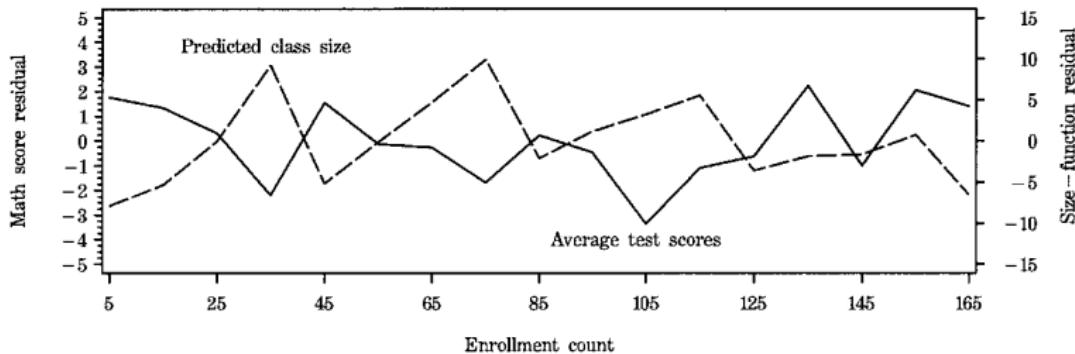
Class-size function and test scores

- The overall positive correlation between test scores and enrollment is partly attributed to larger schools in Israel being geographically concentrated in larger, more affluent cities (smaller schools in poorer “developmental towns” outside the major urban centers)
- They note that the enrollment size and the PD index measuring proportion of students from disadvantaged background is negatively correlated
- They control for the “trend” association between test scores and enrollment size and plot the residuals from regressions of average scores and the average of f_{sc} on average enrollment and PD index for each interval
- The estimates for fifth graders imply a reduction in *predicted* class size of ten students is associated with a 2.2 point increase in average reading scores – a little more than one-quarter of a standard deviation in the distribution of class averages

a. Fifth Grade (Reading)



c. Fifth Grade (Math)



OLS model

$$y_{isc} = X_s \beta + n_{sc} \delta + \mu_c \eta_s + \varepsilon_{sc} \quad (91)$$

where y_{isc} is pupil i 's score, X_s is a vector of school characteristics, sometimes including functions of enrollment and n_{sc} is the size of class c in school s .

- The term μ_c is an i.i.d. random class component, and the term η_s is an i.i.d. random school component
- The class-size coefficient, δ is the primary parameter of interest.

OLS model

- This equation describes the average potential outcomes of students under alternative assignments of n_{sc} controlling for any effects of X_s
- If n_{sc} was randomly assigned conditional on X_s , then δ would be the weighted average response to random variation in class size along the length of the individual causal response functions connecting class size and pupil scores
- Since n_{sc} is not randomly assigned, in practice it is likely that it is correlated with potential outcomes – in this case, the error components in the equation
- But while OLS may not have a causal interpretation, using regression discontinuity design might

Grouped regression

$$\bar{y}_{sc} = X_s \beta + n_{sc} \delta + \eta_s + [\mu_c + \bar{\varepsilon}_{sc}] \quad (92)$$

where overbars denote averages. The bracketed term is the class-level error term.

Step 1: “first stage”

$$n_{sc} = X_s \pi_0 + f_{sc} \pi_1 + \psi_{sc} \quad (93)$$

where π_j are parameter and the error term is defined as the residual from the population regression of η_s onto X_s and f_{sc} and captures other things associated with enrollment

TABLE II
OLS ESTIMATES FOR 1991

	5th Grade					
	Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Mean score</i>	74.3			67.3		
(<i>s.d.</i>)	(8.1)			(9.9)		
<i>Regressors</i>						
Class size	.221 (.031)	-.031 (.026)	-.025 (.031)	.322 (.039)	.076 (.036)	.019 (.044)
Percent disadvantaged		-.350 (.012)	-.351 (.013)		-.340 (.018)	-.332 (.018)
Enrollment			-.002 (.006)			.017 (.009)
Root MSE	7.54	6.10	6.10	9.36	8.32	8.30
<i>R</i> ²	.036	.369	.369	.048	.249	.252
N	2,019			2,018		

5th Graders						
	Class size		Reading comprehension		Math	
	(1)	(2)	(3)	(4)	(5)	(6)
A. Full sample						
Means (s.d.)		29.9 (6.5)		74.4 (7.7)		67.3 (9.6)
<i>Regressors</i>						
f_{sc}	.704 (.022)	.542 (.027)	-.111 (.028)	-.149 (.035)	-.009 (.039)	-.124 (.049)
Percent disadvantaged	-.076 (.010)	-.053 (.009)	-.360 (.012)	-.355 (.013)	-.354 (.017)	-.338 (.018)
Enrollment		.043 (.005)		.010 (.006)		.031 (.009)
Root MSE	4.56	4.38	6.07	6.07	8.33	8.28
R^2	.516	.553	.375	.377	.247	.255
N		2,019		2,019		2,018
B. Discontinuity sample						
Means (s.d.)		30.8 (7.4)		74.5 (8.2)		67.0 (10.2)
<i>Regressors</i>						
f_{sc}	.481 (.053)	.346 (.052)	-.197 (.050)	-.202 (.054)	-.089 (.071)	-.154 (.077)
Percent disadvantaged	-.130 (.029)	-.067 (.028)	-.424 (.027)	-.422 (.029)	-.435 (.039)	-.405 (.042)
Enrollment		.086 (.015)		.003 (.015)		.041 (.022)
Root MSE	5.95	5.58	6.24	6.24	8.58	8.53
R^2	.360	.437	.421	.421	.296	.305
N		471		471		471

Second step: use fitted values in grouped regression

$$\bar{y}_{sc} = X_s \beta + \widehat{n_{sc}} \delta + \eta_s + [\mu_c + \bar{\varepsilon}_{sc}] \quad (94)$$

where $\widehat{n_{sc}}$ is the predicted class size from the previous regression

	Reading comprehension						Math			
	Full sample				+/- 5 Discontinuity sample		Full sample			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Mean score	74.4						74.5		67.3	
(s.d.)	(7.7)						(8.2)		(9.6)	
Regressors										
Class size	-.158 (.040)	-.275 (.066)	-.260 (.081)	-.186 (.104)	-.410 (.113)	-.582 (.181)	-.013 (.056)	-.230 (.092)	-.261 (.113)	-.202 (.131)
Percent disadvantaged	-.372 (.014)	-.369 (.014)	-.369 (.013)		-.477 (.037)	-.461 (.037)	-.355 (.019)	-.350 (.019)	-.350 (.019)	
Enrollment		.022 (.009)	.012 (.026)			.053 (.028)		.041 (.012)	.062 (.037)	
Enrollment squared/100			.005 (.011)						-.010 (.016)	
Piecewise linear trend				.136 (.032)						.193 (.040)
Root MSE	6.15	6.23	6.22	7.71 1961	6.79	7.15 471	8.34	8.40 2018	8.42	9.49 1960
N										

How large is their effects?

- The Tennessee STAR experiment yielded effect sizes of about 0.13-0.27 standard deviation among pupils and about 0.32 - 0.66 standard deviation in the distribution of class means
- Angrist and Lavy (1999) compare their results by calculating their effect size associated with reducing class size by eight pupils (same as STAR)
- They then multiply this times their second step estimate for reading scores for fifth graders (-0.275) gives an effect size of around 2.2 points or 0.29 standard deviation
- Their estimates of effect size for fifth graders are at the low end of the range of those found in the Tennessee experiment
- Observational studies are often confounded by a failure to isolate a credible source of exogenous variation in school inputs, leading one researcher (Hanushek 1997) to conclude that school inputs don't matter in academic performance
- RDD overcomes problems of confounding by exploiting exogenous variation that originates in *administrative rules*
- As with STAR experiment, their study shows that smaller classes appear beneficial to student academic achievement

RDD Data Requirements

- Question: So, where can I find these “jumps”? Answer: Humans are embedding “jumps” in their rules all the time. Dumb rules – while usually bad policy – are *great* for research.
- Validity doesn’t require the assignment rule be arbitrary, only that it is known, precise and free of manipulation. The most effective RDD studies involve programs where X has a “hair trigger” that is not tightly related to the outcome being studied. Examples:
 - Probability of being tried as an adult (higher penalties for a given crime) “jumps” at age 18
 - Probability of being arrested for DWI “jumps” at blood alcohol content >0.08
 - Probability of receiving universal healthcare insurance “jumps” at age 65
 - Probability of receiving medical attention “jumps” when birthweight falls below 1,500 grams
 - Probability of having to attend summer school “jumps” when grade falls below 60
- Data requirements can be substantial. Large sample sizes are characteristic features of the RDD
 - If there are strong trends, one typically needs a lot of data
 - Researchers are typically using administrative data or settings such as birth records where there are **many** observations

More recent examples

- Almond et al. (2010): Infants with low birthweight have both higher mortality and receive more medical treatment. Medical treatment becomes more likely when infant weight falls below 1,500 grams (or 3 pounds). What's the effect of medical treatment on infant mortality?
- Anderson and Magruder (2012): At Yelp.com, individuals reviewed restaurants on a scale of 1 to 5. The aggregate score was a weighted average of all reviews. Yelp assigned an integer stars based on which quintile a restaurant was in the distribution of reviews. What's the effect of a star on reservations and revenue?

Sharp vs. Fuzzy RDD

RDD estimates the causal effect by distinguishing the discontinuous function, $1(X \geq X_0)$, from the smooth selection function, $f(X)$ (though Van der Klaauw uses S instead of X notation)

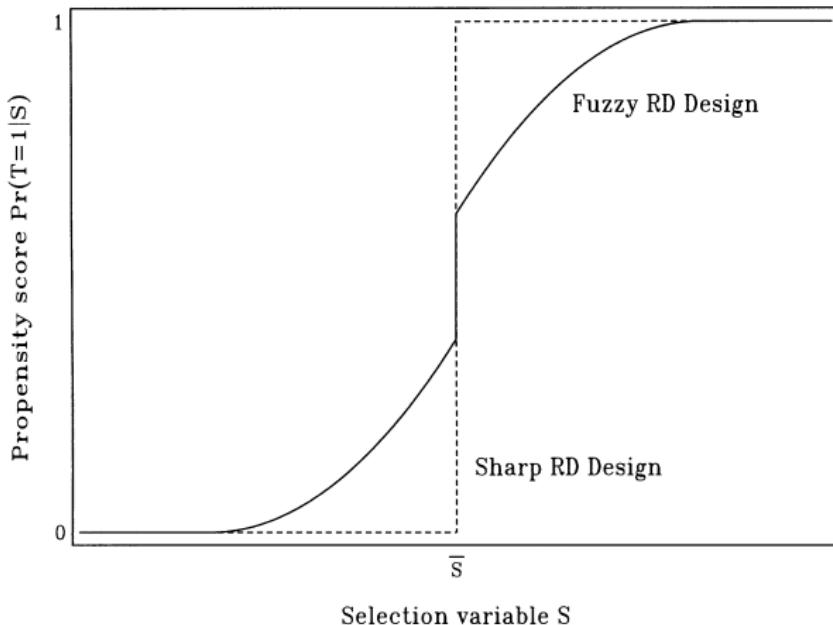


FIGURE 2

ASSIGNMENT IN THE SHARP (DASHED) AND FUZZY (SOLID) RD DESIGN

Sharp vs. Fuzzy RDD

- ① Sharp RDD: Treatment is a deterministic function of running variable, X . Example: Medicare benefits.
- ② Fuzzy RDD: Discontinuous “jump” in the *probability* of treatment when $X > X_0$. Cutoff is used as an instrumental variable for treatment. Example: Maimonides Rule (Angrist and Lavy 1999)

Treatment assignment in the sharp RDD

Deterministic treatment assignment (“sharp RDD”)

In Sharp RDD, treatment status is a deterministic and discontinuous function of a covariate, X_i :

$$D_i = \begin{cases} 1 & \text{if } X_i \geq X_0 \\ 0 & \text{if } X_i < X_0 \end{cases}$$

where X_0 is a known threshold or cutoff. In other words, if you know the value of X_i for a unit i , you know treatment assignment for unit i with certainty.

Example: Let X be age. Americans aged 64 are not eligible for Medicare, but Americans aged 65 ($X \geq X_{65}$) are eligible for Medicare (ignoring disability exemptions)

Treatment effect: definition and estimation

Definition of treatment effect

Assume constant treatment effects potential outcomes model linear in X :

$$\begin{aligned}Y_i^0 &= \alpha + \beta X_i \\Y_i^1 &= Y_i^0 + \delta\end{aligned}$$

Use the switching equation and write in terms of Y_i :

$$\begin{aligned}Y_i &= Y_i^0 + (Y_i^1 - Y_i^0)D_i \\Y_i &= \alpha + \beta X_i + \delta D_i + \varepsilon_i\end{aligned}$$

The treatment effect parameter, δ , is the discontinuity in the conditional expectation function:

$$\begin{aligned}\delta &= \lim_{X_i \rightarrow X_0} E[Y_i^1 | X_i = X_0] - \lim_{X_0 \leftarrow X_i} E[Y_i^0 | X_i = X_0] \\&= \lim_{X_i \rightarrow X_0} E[Y_i | X_i = X_0] - \lim_{X_0 \leftarrow X_i} E[Y_i | X_i = X_0]\end{aligned}$$

The sharp RDD estimation is interpreted as an average causal effect of the treatment at the discontinuity

$$\delta_{SRD} = E[Y_i^1 - Y_i^0 | X_i = X_0]$$

Notice the role of *extrapolation* in estimating treatment effects when sharp RDD

- Left of cutoff, only non-treated observations, $D_i = 0$ for $X < X_0$
- Right of cutoff, only treated observations, $D_i = 1$ for $X \geq X_0$

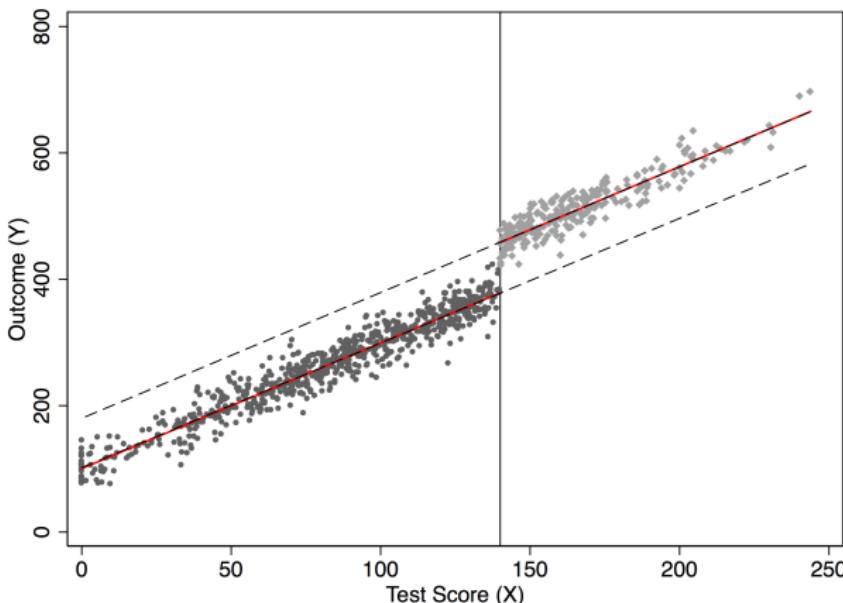


Figure: Dashed lines are extrapolations

Key identifying assumption

Continuity of conditional regression functions (Hahn, Todd and Van der Klaauw 2001; Lee 2008)

$E[Y_i^0|X = X_0]$ and $E[Y_i^1|X = X_0]$ are continuous (smooth) in X at X_0 .

Remark Alfred Marshall quotes Darwin in Principles of Economics (1890), “*Natura non facit saltum*”, which means “nature does not make jumps”. Jumps are unnatural, so when they occur, they require an explanation

Meaning If population average *potential outcomes*, Y^1 and Y^0 , are continuous functions of X at the cutoff, X_0 , then potential average outcomes *do not* jump at X_0 . All other unobserved determinants of Y are continuously related to the running variable, X

Implication This assumption allows us to use average outcome of units right below the cutoff as a valid counterfactual for units right above the cutoff. The causal effect of the treatment will be based on extrapolation from the trend, $E[Y_i^0|X < X_0]$, to those values of $X > X_0$ for the $E[Y_i^0|X > X_0]$.

```
/// --- Examples using simulated data
set obs 1000
set seed 1234567

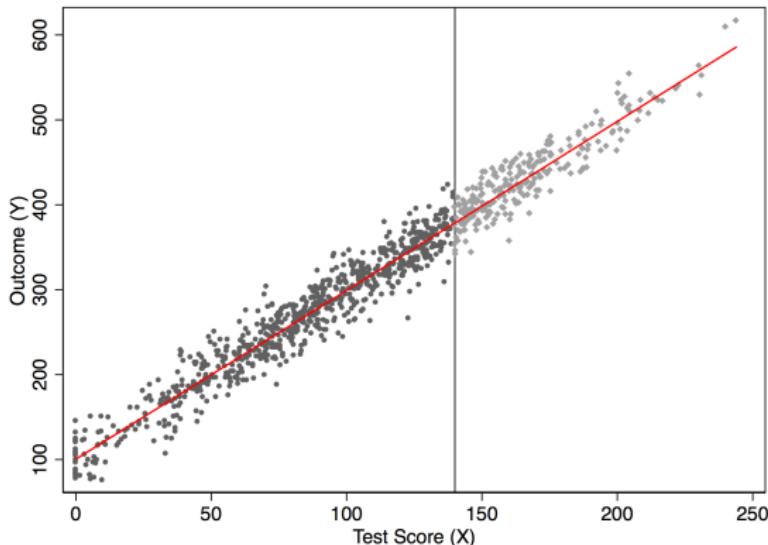
* Generate running variable
gen x = rnormal(100, 50)
replace x=0 if x < 0
drop if x > 280
sum x, det

* Set the cutoff at X=140. Treated if X > 140
gen D = 0
replace D = 1 if x > 140
```

Graphical example of continuous assumption

```
gen y1 = 100 + 0*D + 2*x + rnormal(0, 20)
```

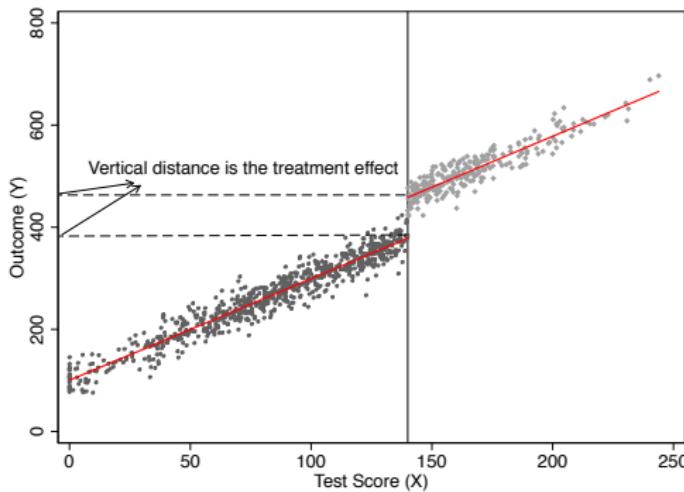
```
scatter y1 x if D==0, msize(vsmall) || scatter y1 x if D==1, msize(vsmall)
legend(off) xline(140, lstyle(foreground)) || lfit y1 x if D ==0, color(red)
|| lfit y1 x if D ==1, color(red) ytitle("Outcome (Y)") xtitle("Test
Score (X)")
```



Graphical example with treatment effect, cutoff=140

```
gen y = 100 + 80*D + 2*x + rnormal(0, 20)

scatter y x if D==0, msize(vsmall) || scatter y x if D==1, ///
msize(vsmall) legend(off) xline(140, lstyle(foreground)) || ///
lfit y x if D ==0, color(red) || lfit y x if D ==1, color(red) /// ytitle("Out(Y)") xtitle("Test Score (X)")
```



Tangent: centering at the cutoff point

- It is common for authors to transform X by “centering” at X_0 :

$$Y_i = \alpha + \beta(X_i - X_0) + \delta D_i + \varepsilon_i$$

This doesn't change the interpretation of the treatment effect – only the interpretation of the intercept.

- Example: Medicare and age 65. Center the running variable (age) by subtracting 65:

$$\begin{aligned} Y &= \beta_0 + \beta_1(Age - 65) + \beta_2 Edu \\ &= \beta_0 + \beta_1 Age - \beta_1 65 + \beta_2 Edu \\ &= \alpha + \beta_1 Age + \beta_2 Edu \end{aligned}$$

where $\alpha = \beta_0 - \beta_1 65$. All other coefficients, notice, have the same interpretation, except for the intercept.

```
. gen x_c=x-140
```

```
. reg y D x
```

Source	SS	df	MS	Number of obs	=	999
Model	15842893.9	2	7921446.97	F(2, 996)	=	19988.47
Residual	394715.557	996	396.30076	Prob > F	=	0.0000
Total	16237609.5	998	16270.1498	R-squared	=	0.9757

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	80.01418	2.144779	37.31	0.000	75.80537 84.22298
x	1.986975	.0186779	106.38	0.000	1.950322 2.023627
_cons	100.3885	1.70944	58.73	0.000	97.03397 103.743

```
. reg y D x_c
```

Source	SS	df	MS	Number of obs	=	999
Model	15842893.9	2	7921446.97	F(2, 996)	=	19988.47
Residual	394715.554	996	396.300757	Prob > F	=	0.0000
Total	16237609.5	998	16270.1498	R-squared	=	0.9756

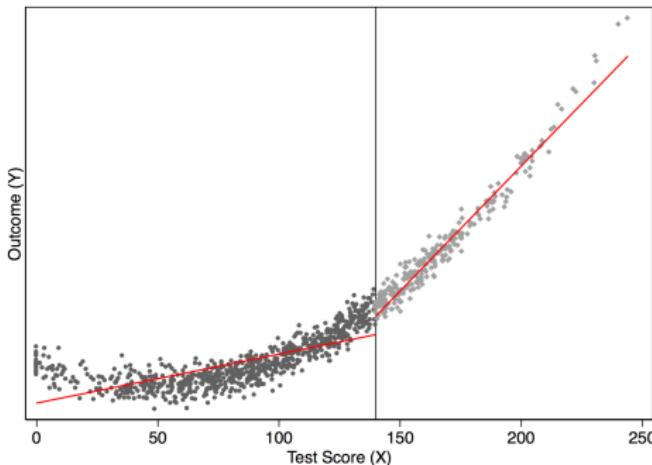
y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	80.01418	2.144779	37.31	0.000	75.80537 84.22298
x_c	1.986975	.0186779	106.38	0.000	1.950322 2.023627
_cons	378.565	1.290755	293.29	0.000	376.032 381.0979

Nonlinearity bias

Smoothness in $E[Y_i^0|X_i]$ and linearity are different things. What if the trend relation $E[Y_i^0|X_i]$ does not jump at X_0 but rather is simply nonlinear?

```
gen x2 = x**  
gen x3 = x***  
gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)
```

```
scatter y x if D==0, msize(vsmall) || scatter y x if D==1, msize(vsmall) legend(off) xline(140,  
lstyle(foreground)) ylabel(None) || lfit y x if D ==0, color(red) || lfit y x if D ==1, color(red)  
xtitle("Test Score (X)") ytitle("Outcome (Y)")
```



Sharp RDD: Nonlinear Case

- Suppose the nonlinear relationship is $E[Y_i^0|X_i] = f(X_i)$ for some reasonably smooth function $f(X_i)$. In that case we'd fit the regression model:

$$Y_i = f(X_i) + \delta D_i + \eta_i$$

- Since $f(X_i)$ is counterfactual for values of $X_i > X_0$, how will we model the nonlinearity? There are 2 ways of approximating $f(X_i)$:

- Let $f(X_i)$ equal a p^{th} order polynomial:

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \delta D_i + \eta_i$$

- Use a nonparametric kernel method (later)

Different polynomials on the 2 sides of the discontinuity

- We can generalize the function, $f(x_i)$, by allowing the x_i terms to differ on both sides of the cutoff by including them both individually and interacting them with D_i . In that case we have:

$$\begin{aligned}E[Y_i^0|X_i] &= \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \cdots + \beta_{0p}\tilde{X}_i^p \\E[Y_i^1|X_i] &= \alpha + \delta + \beta_{11}\tilde{X}_i + \beta_{12}\tilde{X}_i^2 + \cdots + \beta_{1p}\tilde{X}_i^p\end{aligned}$$

where \tilde{X}_i is the centered running variable (i.e., $X_i - X_0$). Centering at X_0 ensures that the treatment effect at $X_i = X_0$ is the coefficient on D_i in a regression model with interaction terms

- As Lee and Lemieux (2010) note, allowing different functions on both sides of the discontinuity should be the main results in an RDD paper as otherwise we use values from both sides of the cutoff to estimation the function on each side

Different polynomials on the 2 sides of the discontinuity

- To derive a regression model, first note that the observed values must be used in place of the potential outcomes:

$$E[Y|X] = E[Y^0|X] + (E[Y^1|X] - E[Y^0|X]) D$$

- Regression model you estimate is:

$$\begin{aligned} Y_i = & \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p \\ & + \delta D_i + \beta_1^* D_i \tilde{x}_i + \beta_2^* D_i \tilde{x}_i^2 + \cdots + \beta_p^* D_i \tilde{x}_i^p + \varepsilon_i \end{aligned}$$

where $\beta_1^* = \beta_{11} - \beta_{01}$, $\beta_2^* = \beta_{21} - \beta_{21}$ and $\beta_p^* = \beta_{1p} - \beta_{0p}$

- The equation we looked at earlier a few slides back was just a special case of the above equation with $\beta_1^* = \beta_2^* = \beta_p^* = 0$
- The treatment effect at x_0 is δ
- The treatment effect at $X_i - X_0 = c > 0$ is: $\delta + \beta_1^* c + \beta_2^* c^2 + \cdots + \beta_p^* c^p$

Polynomial simulation example

```
capture drop y x2 x3
gen x2 = x*x
gen x3 = x*x*x
gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)

reg y D x x2 x3

predict yhat

scatter y x if D==0, msize(vsmall) || scatter y x if D==1, msize(vsmall)
legend(off) xline(140, lstyle(foreground)) ylabel(None) || line yhat
x if D ==0, color(red) sort || line yhat x if D ==1, sort color(red)
xtitle("Test Score (X)") ytitle("Outcome (Y)")
```

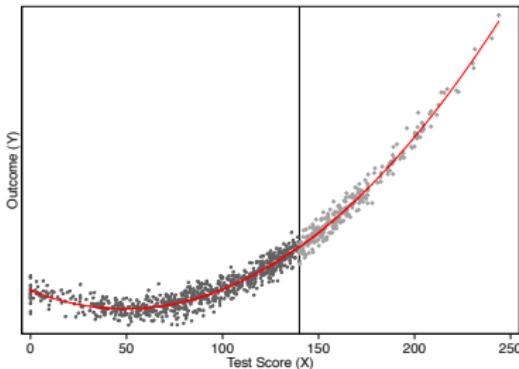


Figure: Third degree polynomial. Actual model second degree polynomial.

```

. gen x_c = x - 140
. gen x2_c = x2-140
. gen x3_c = x3-140

```

```
. reg y D x2
```

Source	SS	df	MS	Number of obs	=	999
Model	3.7863e+10	3	1.2621e+10	F(3, 995)	=	13115.22
Residual	957507024	995	962318.617	Prob > F	=	0.0000
Total	3.8821e+10	998	38898361.8	R-squared	=	0.9753
				Adj R-squared	=	0.9753
				Root MSE	=	980.98

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	-115.5381	127.4967	-0.91	0.365	-365.7314 134.6552
x	-98.57582	2.285769	-43.13	0.000	-103.0613 -94.09034
x2	1.000001	.0122767	81.45	0.000	.9759098 1.024092
_cons	9864.218	111.1206	88.77	0.000	9646.16 10082.28

```
. reg y D x_c x2_c
```

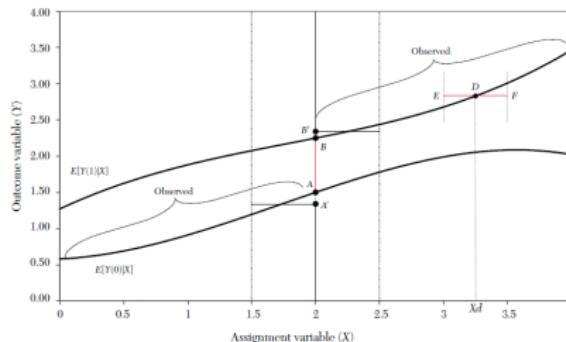
Source	SS	df	MS	Number of obs	=	999
Model	3.7863e+10	3	1.2621e+10	F(3, 995)	=	13115.22
Residual	957507020	995	962318.613	Prob > F	=	0.0000
Total	3.8821e+10	998	38898361.8	R-squared	=	0.9753
				Adj R-squared	=	0.9753
				Root MSE	=	980.98

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
D	-115.5381	127.4967	-0.91	0.365	-365.7315 134.6552
x_c	-98.57582	2.285769	-43.13	0.000	-103.0613 -94.09034
x2_c	1.000001	.0122767	81.45	0.000	.9759098 1.024092
_cons	-3796.397	227.7894	-16.67	0.000	-4243.4 -3349.394

Figure: P-order polynomial regressions with and without centering. Only intercept changes.

Kernel regression

- In addition to using p^{th} order polynomials to model the nonlinearities, we can use kernel regression. Neither is right or wrong – they have advantages and disadvantages.
- The nonparametric kernel method has problems because you are trying to estimate regressions at the cutoff point which results in a “boundary problem” (Hahn, Todd and Van der Klaauw 2001)



- While the “true” effect is AB , with a certain bandwidth a rectangular kernel would estimate the effect as $A'B'$
- There is therefore systematic bias with the kernel method if the $f(X)$ is upwards or downwards sloping

Kernel Method - Local linear regression

- The standard solution to this problem is to run local linear nonparametric regression (Hahn, Todd and Van der Klaauw 2001). In the case described by the previous slide, this would substantially reduce the bias
- Think of it as a weighted regression restricted to a window (hence “local”). The kernel provides the weights to that regression. STATA’s poly command estimates kernel-weighted local polynomial regression.
- A rectangular kernel would give the same result as taking $E[Y]$ at a given bin on X . The triangular kernel gives more importance to the observations closer to the center.
- While estimating this in a given window of width h around the cutoff is straightforward, it’s more difficult to choose this bandwidth (or window), and the method is sensitive to the choice of bandwidth. There is essentially a tradeoff between bias and efficiency. See Lee and Lemieux (2010) for two methods to choose on the bandwidth. This is an active area of research in econometrics.

Example of sharp RDD: David Card, Carlos Dobkin and Nicole Maestas (2008), “The Impact of Nearly Universal Insurance Coverage on Health Care Utilization: Evidence from Medicare”, *American Economic Review*.

Non-elderly health insurance information

- 20% of non elderly adults in US lacked health insurance coverage in 2005
- Most were from lower-income families, nearly one-half were African American or Hispanic
- Many analysts have argued that unequal insurance coverage contributes to disparities in health care utilization and health outcomes across SES
- Even among the insured, there are differences: copayments, deductibles, other features that affect use
- Evidence that better insurance causes better health outcomes is limited
- Both supply and demand for insurance depend on health status, confounding observational comparisons between people with different insurance characteristics

Elderly health insurance information

- < 1% of the elderly population are uninsured
- Most have fee-for-service Medicare coverage
- The transition occurs abruptly at age 65, which is the threshold for Medicare eligibility

Estimation

- Reduced form measure of causal effect of health insurance status on health care use, y :

$$y_{ija} = X_{ija}\alpha + f_j(\alpha; \beta) + \sum_k C_{ija}^k \delta^k + u_{ija}$$

where i indexes individuals, j indexes a socioeconomic group, a indexes an age, u_{ija} the unobserved error component, X_{ija} a set of covariates (e.g., gender and region), $f_j(\alpha; \beta)$ a smooth function representing the age profile of outcome y for group j , and $C_{ija}^k (k = 1, 2, \dots, K)$ are characteristics of the insurance coverage held by the individual such as copayment rates

- Problem: Insurance coverage is endogenous
 $\rightarrow \text{cov}(u_{ija}, C_{ija}) \neq 0$.
- Identification: age threshold for Medicare eligibility at 65 used as credibly exogenous variation in insurance status.

Insurance and age

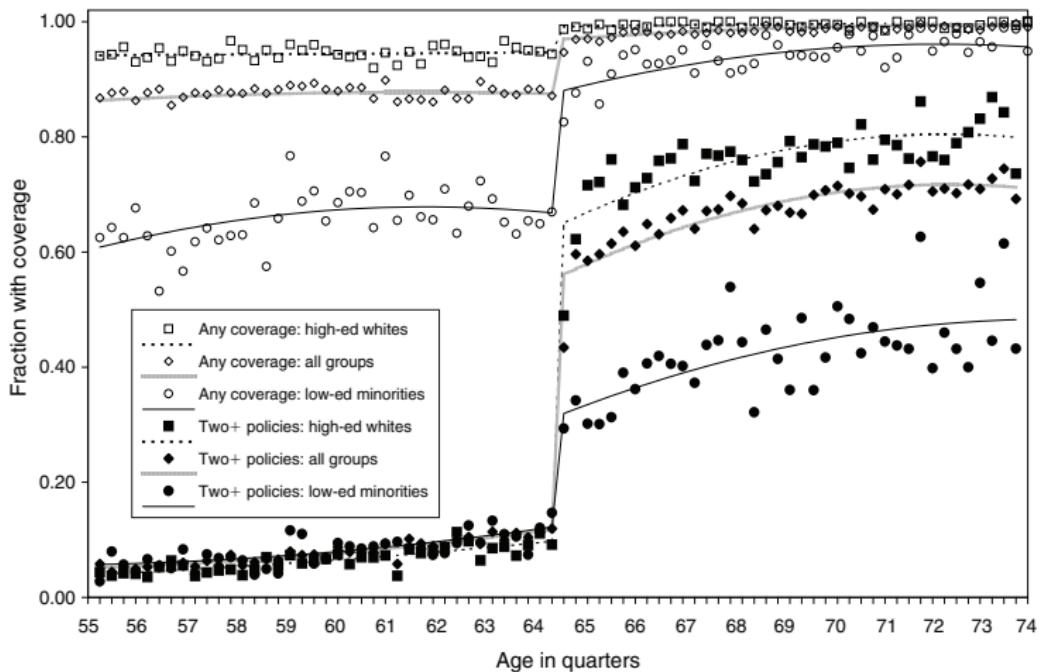


FIGURE 1. COVERAGE BY ANY INSURANCE AND BY TWO OR MORE POLICIES, BY AGE AND DEMOGRAPHIC GROUP

Estimation

- Suppose health insurance coverage can be summarized by two dummy variables
 - C_{ija}^1 (any coverage) and C_{ija}^2 (generous insurance). Linear probability model:

$$\begin{aligned}C_{ija}^1 &= X_{ija}\beta_j^1 + g_j^1(a) + D_a\pi_j^1 + v_{ija}^1 \\C_{ija}^2 &= X_{ija}\beta_j^2 + g_j^2(a) + D_a\pi_j^2 + v_{ija}^2\end{aligned}$$

where β_j^1 and β_j^2 are group-specific coefficients, $g_j^1(a)$ and $g_j^2(a)$ are smooth age profiles for group j , and D_a is a dummy for being age ≥ 65 .

- Reduced form from couple slides back

$$y_{ija} = X_{ija}\alpha + f_j(\alpha; \beta) + \sum_k C_{ija}^k \delta^k + u_{ija}$$

- Combining C_{ija} equations, and rewrite reduced for model

$$y_{ija} = X_{ija}(\alpha_j + \beta_j^1 \delta_j^1 + \beta_j^2 \delta_j^2) + h_j(a) + D_a\pi_j^y + v_{ija}^y$$

where $h(a) = f_j(a) + \delta^1 g_j^1(a) + \delta^2 g_j^2(a)$ is reduced form age profile for group j , $\pi_j^y = \pi_j^1 \delta^1 + \pi_j^2 \delta^2$ and $v_{ija}^y = u_{ija} + v_{ija}^1 \delta^1 + v_{ija}^2 \delta^2$ is the error term.

- Assuming that the profiles $f_j(a)$, $g_j^1(a)$ and $g_j^2(a)$ are continuous at age 65 (i.e., continuous assumption), then any discontinuity in y is due to insurance and magnitudes depend on the size of the insurance changes at 65 (π_j^1 and π_j^2) and on the associated causal effects (δ^1 and δ^2)

Estimation

- For some basic health care services (e.g., routine doctor visits), it may be the only thing that matters is insurance.
- But, in those situations, the implied discontinuity in y at 65 for group j will be proportional to the change in insurance status experienced by that group
- For more expensive (or elective) services, the generosity of the coverage may matter (for instance, if patients are unwilling to cover the required copay or if the managed care program won't cover the service).
- This creates a potential identification problem in interpreting the discontinuity in y for any one group
- Since π_j^y is a linear combination of the discontinuities in coverage and generosity, δ^1 and δ^2 can be estimated by a regression across groups:

$$\pi_j^y = \delta^0 + \delta^1 \pi_j^1 + \delta^2 \pi_j^2 + e_j$$

where e_j is an error term reflecting a combination of the sampling errors in π_j^y , π_j^1 and π_j^2

- RDD models like the ones on the previous slide are fit with OLS by demographic subgroup to individual data
- Estimates are combined across groups to estimate the above model

Data

- 1992-2003 National Health Interview Survey (NHIS)
 - NHIS reports respondents' birth year and birth month, and the calendar quarter of the interview. Authors construct an estimate of age in quarters at date of interview.
 - A person reaching 65 in the interview quarter is age 65 and 0 quarters
 - Assuming a uniform distribution of interview dates, one-half of these people will be 0-6 weeks younger than 65 and one-half will be 0-6 weeks older
 - Analysis is limited to people >55 and <75 . Final sample size is $n = 160,821$ (although some outcomes are not available for earlier years)
- 1992-2003 hospital discharge records for California, Florida and New York
 - Discharge files represent a complete census of discharges from all hospitals in the three states (except federally regulated institutions).
 - Data files include information on age in months at the time of admission
 - Sample selection criteria: drop records for people admitted as transfers from other institutions, and limit people between 60 and 70 years of age at admission
 - Sample sizes are 4,017,325 (California), 2,793,547 (Florida), 3,121,721 (New York)

Medicare description

- Description of Medicare eligibility
 - Medicare is available to people who are at least 65 and have worked 40 quarters or more in covered employment or have a spouse who did.
 - Coverage is available to younger people with severe kidney disease and recipients of Social Security Disability Insurance
 - Eligible individuals can obtain Medicare hospital insurance (Part A) free of charge, and medical insurance (Part B) for a modest monthly premium
 - Individuals receive notice of their impending eligibility for Medicare shortly before 65th birthday and are informed they have to enroll in it and choose whether to accept Part B coverage.
 - Coverage begins on the first day of the month in which they turn 65.

Medicare description

- Five insurance-related variables: probability of Medicare coverage, any health insurance coverage, private coverage, two or more forms of coverage, and individual's primary health insurance is managed care program.
 - Data are drawn from 1999-2003 NHIS and for each characteristic, authors show the incidence rate at ages 63-64 and the change at age 65 based on a version of the C_k equations that include a quadratic in age, fully interacted with a post-65 dummy as well as controls for gender, education, race/ethnicity, region and sample year
 - Alternative specifications were also used – a parametric model fit to a narrower age window (ages 63-67) and a local linear regression specification using a chosen bandwidth. Both show similar estimates of the change at age 65.

Changes in insurance coverage at age 65

TABLE 1—INSURANCE CHARACTERISTICS JUST BEFORE AGE 65 AND ESTIMATED DISCONTINUITIES AT AGE 65

	On Medicare		Any insurance		Private coverage		2+ Forms coverage		Managed care	
	Age 63–4 (1)	RD at 65 (2)	Age 63–4 (3)	RD at 65 (4)	Age 63–4 (5)	RD at 65 (6)	Age 63–4 (7)	RD at 65 (8)	Age 63–4 (9)	RD at 65 (10)
Overall sample	12.3	59.7 (4.1)	87.9	9.5 (0.6)	71.8	−2.9 (1.1)	10.8	44.1 (2.8)	59.4	−28.4 (2.1)
<i>Classified by ethnicity and education:</i>										
White non-Hispanic:										
High school dropout	21.1	58.5 (4.6)	84.1	13.0 (2.7)	63.5	−6.2 (3.3)	15.0	44.5 (4.0)	48.1	−25.0 (4.5)
High school graduate	11.4	64.7 (5.0)	92.0	7.6 (0.7)	80.5	−1.9 (1.6)	10.1	51.8 (3.8)	58.9	−30.3 (2.6)
At least some college	6.1	68.4 (4.7)	94.6	4.4 (0.5)	85.6	−2.3 (1.8)	8.8	55.1 (4.0)	69.1	−40.1 (2.6)
Minority:										
High school dropout	19.5	44.5 (3.1)	66.8	21.5 (2.1)	33.2	−1.2 (2.5)	11.4	19.4 (1.9)	39.1	−8.3 (3.1)
High school graduate	16.7	44.6 (4.7)	85.2	8.9 (2.8)	60.9	−5.8 (5.1)	13.6	23.4 (4.8)	54.2	−15.4 (3.5)
At least some college	10.3	52.1 (4.9)	89.1	5.8 (2.0)	73.3	−5.4 (4.3)	11.1	38.4 (3.8)	66.2	−22.3 (7.2)
<i>Classified by ethnicity only:</i>										
White non-Hispanic	10.8	65.2 (4.6)	91.8	7.3 (0.5)	79.7	−2.8 (1.4)	10.4	51.9 (3.5)	61.9	−33.6 (2.3)
(all)										
Black non-Hispanic	17.9	48.5 (3.6)	84.6	11.9 (2.0)	57.1	−4.2 (2.8)	13.4	27.8 (3.7)	48.2	−13.5 (3.7)
(all)										
Hispanic (all)	16.0	44.4 (3.7)	70.0	17.3 (3.0)	42.5	−2.0 (1.7)	10.8	21.7 (2.1)	52.9	−12.1 (3.7)

Note: Entries in odd-numbered columns are percentages of age 63–64-year-olds in group with insurance characteristic shown in column heading. Entries in even-numbered columns are estimated regression discontinuities at age 65, from models that include quadratic control for age, fully interacted with dummy for age 65 or older. Other controls include indicators for gender, race/ethnicity, education, region, and sample year. Estimates are based on linear probability models fit to pooled samples of 1999–2003 NHIS.

Other changes at age 65

- Formal identification of an RD model relating an outcome y (insurance coverage) to a treatment (Medicare age-eligibility) that depends on age (a) relies on assumption that the conditional expectation functions for both potential outcomes is continuous at $a = 65$. That is, $E[Y^0|a]$ and $E[Y^1|a]$ are continuous at $a = 65$. If so, then the average treatment effect at age 65 is identified as

$$\lim_{65 \leftarrow a} E[y^1|a] - \lim_{a \rightarrow 65} E[y^0|a]$$

- Continuity requires that all other factors that might affect the outcome of interest (insurance coverage) also trend smoothly at the cutoff $a = 65$.
- What else changes, though, at 65?
 - Employment. 65 is the traditional age of retirements. Any abrupt change in employment could lead to differences in health care utilization if non workers have more time to visit doctors.
- Test for potential discontinuities at age 65 for confounding variables using the 1996-2004 March CPS with age measured in years. No evidence for large discontinuities in employment (next slide).

Other changes at age 65

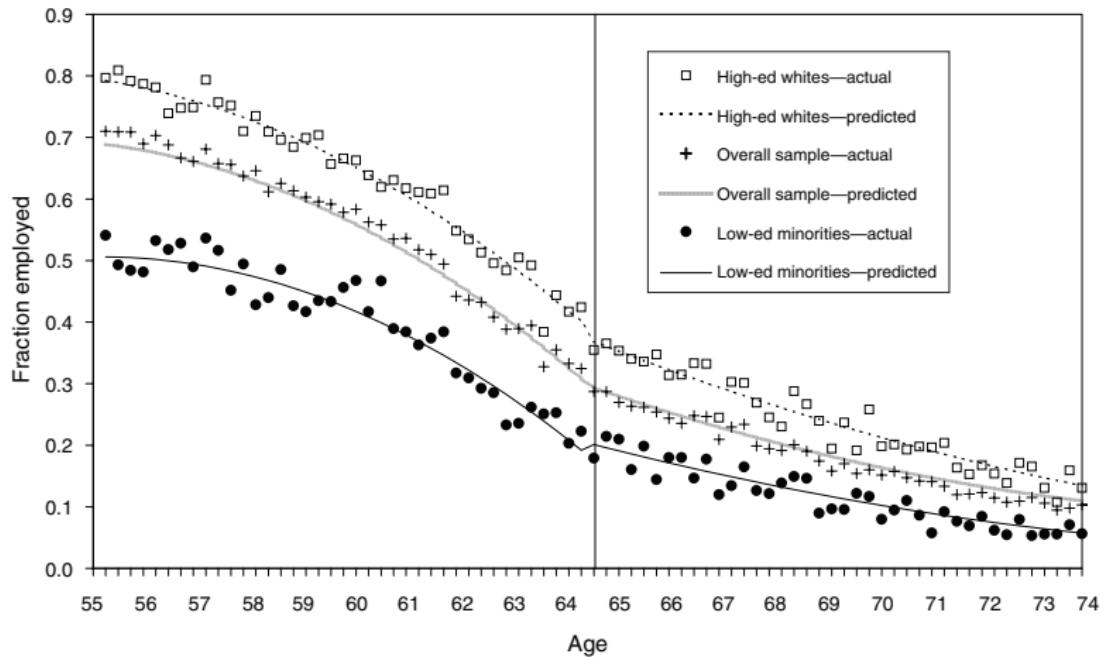


FIGURE 2. EMPLOYMENT RATES BY AGE AND DEMOGRAPHIC GROUP (1992–2003 NHIS)

Effect of cutoff on access to care and utilization

Since 1997, NHIS has asked two questions: (1) "During the past 12 months has medical care been delayed for this person because of worry about the cost?" and (2) "During the past 12 months was there any time when this person needed medical care but did not get it because (this person) could not afford it?"

TABLE 3—MEASURES OF ACCESS TO CARE JUST BEFORE 65 AND ESTIMATED DISCONTINUITIES AT 65

	1997–2003 NHIS				1992–2003 NHIS			
	Delayed care last year		Did not get care last year		Saw doctor last year		Hospital stay last year	
	Age 63–64	RD at 65	Age 63–64	RD at 65	Age 63–64	RD at 65	Age 63–64	RD at 65
Overall sample	7.2	−1.8 (0.4)	4.9	−1.3 (0.3)	84.8	1.3 (0.7)	11.8	1.2 (0.4)
<i>Classified by ethnicity and education:</i>								
White non-Hispanic:								
High school dropout	11.6	−1.5 (1.1)	7.9	−0.2 (1.0)	81.7	3.1 (1.3)	14.4	1.6 (1.3)
High school graduate	7.1	0.3 (2.8)	5.5	−1.3 (2.8)	85.1	−0.4 (1.5)	12.0	0.3 (0.7)
At least some college	6.0	−1.5 (0.4)	3.7	−1.4 (0.3)	87.6	0.0 (1.3)	9.8	2.1 (0.7)
Minority:								
High school dropout	13.6	−5.3 (1.0)	11.7	−4.2 (0.9)	80.2	5.0 (2.2)	14.5	0.0 (1.4)
High school graduate	4.3	−3.8 (3.2)	1.2	1.5 (3.7)	84.8	1.9 (2.7)	11.4	1.8 (1.4)
At least some college	5.4	−0.6 (1.1)	4.8	−0.2 (0.8)	85.0	3.7 (3.9)	9.5	0.7 (2.0)
<i>Classified by ethnicity only:</i>								
White non-Hispanic	6.9	−1.6 (0.4)	4.4	−1.2 (0.3)	85.3	0.6 (0.8)	11.6	1.3 (0.5)
Black non-Hispanic (all)	7.3	−1.9 (1.1)	6.4	−0.3 (1.1)	84.2	3.6 (1.9)	14.4	0.5 (1.1)
Hispanic (all)	11.1	−4.9 (0.8)	9.3	−3.8 (0.7)	79.4	8.2 (0.8)	11.8	1.0 (1.6)

Note: Entries in odd numbered columns are mean of variable in column heading among people ages 63–64. Entries in even numbered columns are estimated regression discontinuities at age 65, from models that include linear control for age interacted with dummy for age 65 or older (columns 2 and 4) or quadratic control for age, interacted with dummy for age 65 and older (columns 6 and 8). Other controls in models include indicators for gender, race/ethnicity, education, region, and sample year. Sample in columns 1–4 is pooled 1997–2003 NHIS. Sample in columns 5–8 is pooled 1992–2003 NHIS. Samples for regression models include people ages 55–75 only. Standard errors (in parentheses) are clustered by quarter of age.

Effect of cutoff on access to care and utilization

The right-hand columns of Table 3 present results for two key measures of healthcare utilization: (1) "Did the individual have at least one doctor visit in the past year?" and (2) "Did the individual have one or more overnight hospital stays in the past year?"

TABLE 3—MEASURES OF ACCESS TO CARE JUST BEFORE 65 AND ESTIMATED DISCONTINUITIES AT 65

	1997–2003 NHIS				1992–2003 NHIS			
	Delayed care last year		Did not get care last year		Saw doctor last year		Hospital stay last year	
	Age 63–64	RD at 65	Age 63–64	RD at 65	Age 63–64	RD at 65	Age 63–64	RD at 65
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
Overall sample	7.2	−1.8 (0.4)	4.9	−1.3 (0.3)	84.8	1.3 (0.7)	11.8	1.2 (0.4)
<i>Classified by ethnicity and education:</i>								
White non-Hispanic:								
High school dropout	11.6	−1.5 (1.1)	7.9	−0.2 (1.0)	81.7	3.1 (1.3)	14.4	1.6 (1.3)
High school graduate	7.1	0.3 (2.8)	5.5	−1.3 (2.8)	85.1	−0.4 (1.5)	12.0	0.3 (0.7)
At least some college	6.0	−1.5 (0.4)	3.7	−1.4 (0.3)	87.6	0.0 (1.3)	9.8	2.1 (0.7)
Minority:								
High school dropout	13.6	−5.3 (1.0)	11.7	−4.2 (0.9)	80.2	5.0 (2.2)	14.5	0.0 (1.4)
High school graduate	4.3	−3.8 (3.2)	1.2	1.5 (3.7)	84.8	1.9 (2.7)	11.4	1.8 (1.4)
At least some college	5.4	−0.6 (1.1)	4.8	−0.2 (0.8)	85.0	3.7 (3.9)	9.5	0.7 (2.0)
<i>Classified by ethnicity only:</i>								
White non-Hispanic	6.9	−1.6 (0.4)	4.4	−1.2 (0.3)	85.3	0.6 (0.8)	11.6	1.3 (0.5)
Black non-Hispanic (all)	7.3	−1.9 (1.1)	6.4	−0.3 (1.1)	84.2	3.6 (1.9)	14.4	0.5 (1.1)
Hispanic (all)	11.1	−4.9 (0.8)	9.3	−3.8 (0.7)	79.4	8.2 (0.8)	11.8	1.0 (1.6)

Note: Entries in odd numbered columns are mean of variable in column heading among people ages 63–64. Entries in even numbered columns are estimated regression discontinuities at age 65, from models that include linear control for age interacted with dummy for age 65 or older (columns 2 and 4) or quadratic control for age, interacted with dummy for age 65 and older (columns 6 and 8). Other controls in models include indicators for gender, race/ethnicity, education, region, and sample year. Sample in columns 1–4 is pooled 1997–2003 NHIS. Sample in columns 5–8 is pooled 1992–2003 NHIS. Samples for regression models include people ages 55–75 only. Standard errors (in parentheses) are clustered by quarter of age.

Changes in hospitalizations

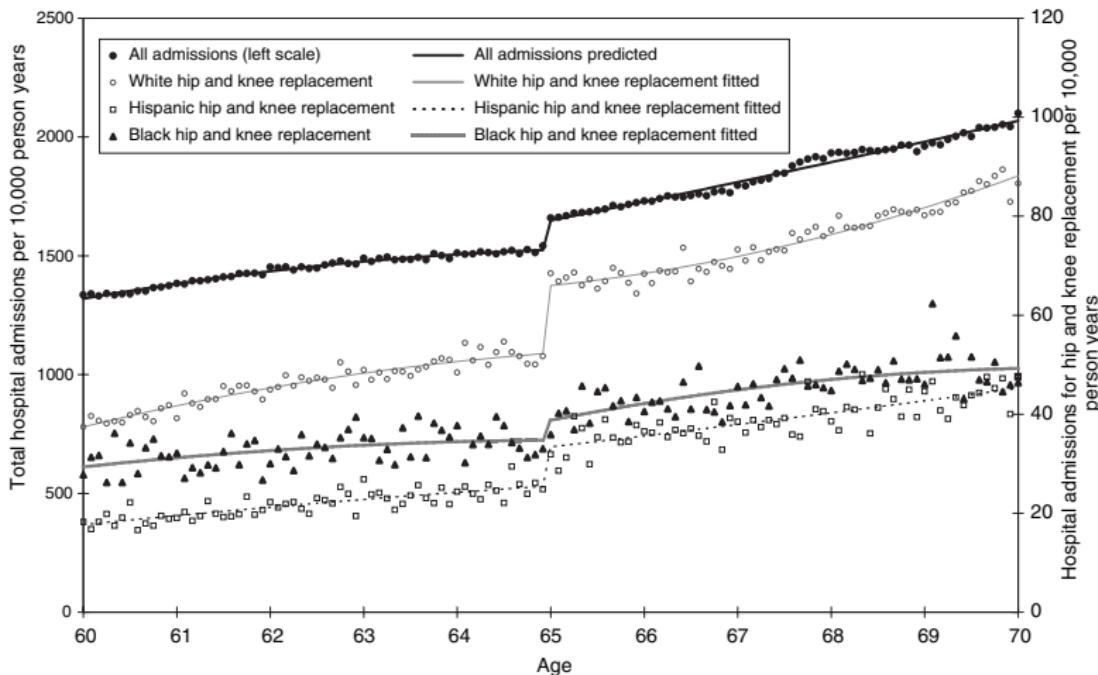


FIGURE 3. HOSPITAL ADMISSION RATES BY RACE/ETHNICITY

Effect of cutoff on access to care and utilization

TABLE 4—HOSPITAL ADMISSIONS AND INSURANCE COVERAGE AT AGE 65: CALIFORNIA, FLORIDA, AND NEW YORK

	All		Whites		Hispanics		Blacks	
	Rate age 60–64 (1)	RD at 65 (2)	Rate age 60–64 (3)	RD at 65 (4)	Rate age 60–64 (5)	RD at 65 (6)	Rate age 60–64 (7)	RD at 65 (8)
<i>Hospital admissions</i>								
All admissions	1,443	7.57 (0.29)	1,407	7.74 (0.33)	1,262	9.47 (0.55)	2,008	4.39 (0.71)
By route into hospital								
ER admission	761	3.30 (0.39)	688	3.70 (0.40)	774	2.63 (0.92)	1,313	1.93 (0.95)
Non-ER admission	682	12.16 (0.46)	718	11.51 (0.49)	488	19.89 (1.05)	695	8.92 (1.04)
By admission diagnosis								
Chronic ischemic heart disease	83	11.58 (0.96)	90	11.05 (1.16)	59	18.45 (2.45)	66	8.29 (2.78)
AMI	48	4.41 (1.43)	50	5.31 (1.65)	38	3.90 (3.33)	45	–3.43 (4.78)
Heart failure	56	0.44 (1.11)	45	2.33 (1.24)	62	–4.85 (2.63)	130	–1.47 (2.43)
Chronic bronchitis	34	7.50 (1.51)	36	6.50 (1.52)	19	9.76 (5.58)	38	13.05 (4.43)
Osteoarthritis	34	26.97 (1.39)	38	27.16 (1.64)	18	29.27 (5.05)	27	22.08 (4.01)
Pneumonia	34	2.44 (1.42)	32	2.05 (1.74)	30	3.39 (4.34)	51	3.81 (3.21)
By primary procedure								
None	419	5.70 (0.33)	400	5.73 (0.40)	388	7.23 (1.23)	614	3.86 (1.25)
Diagnostic procedures on heart	51	9.18 (1.03)	53	8.17 (1.20)	40	16.78 (3.21)	58	8.76 (3.32)
Removal of coronary artery obstruction	38	10.67 (1.46)	43	10.49 (1.60)	23	18.77 (3.94)	22	0.49 (5.33)
Bypass anastomosis of heart	26	15.91 (1.39)	29	16.17 (1.44)	17	18.97 (5.62)	13	5.15 (6.09)
Joint replacement lower extremity	41	22.69 (1.47)	46	23.16 (1.60)	22	26.40 (4.69)	33	12.14 (4.20)
Diagnostic procedure on small intestine	35	7.35 (1.27)	31	6.60 (1.47)	37	13.07 (3.27)	58	4.09 (3.13)
Cholecystectomy (gall bladder removal)	26	17.93 (2.10)	26	16.00 (1.84)	29	29.25 (5.11)	18	12.27 (7.50)
<i>Insurance coverage</i>								
Probability of coverage (March CPS data)	82.7	15.0 (0.8)	86.7	12.7 (0.8)	69.1	20.3 (2.7)	79.0	17.6 (2.7)

Notes: Insurance estimates are based on pooled March CPS 1996–2004 data for California, Florida, and New York. Entries in top row columns 1, 3, 5, 7 are fractions of 60- and 64-year-olds with insurance coverage. Entries in top row columns 2, 4, 6, 8 represent regression discontinuity estimates ($\times 100$) of the increase in coverage at age 65 from a model with a quadratic in age, fully interacted with a post-65 dummy. Entries in lower rows columns 1, 3, 5, 7 are hospital admission rates (per 10,000 person years for 60- to 64-year-olds) for California, Florida, and New York 1992–2002. Entries in lower rows columns 2, 4, 6, 8 are regression discontinuity estimates ($\times 100$) of the increase in the log of the number of admissions at age 65, from models with a quadratic in age, fully interacted with a post-65 dummy. Standard errors are in parentheses.

Changes in ownership

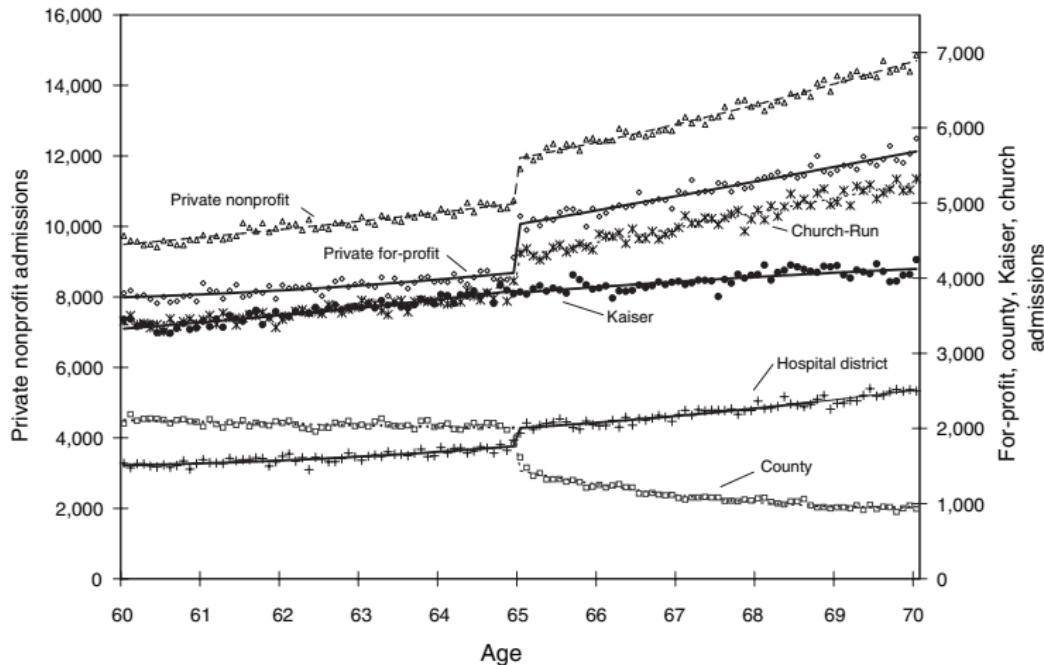


FIGURE 4. HOSPITAL ADMISSION IN CALIFORNIA BY OWNERSHIP TYPE (1992–2002)

Socioeconomic disparities

TABLE 5—SUMMARY OF EFFECTS OF INSURANCE COVERAGE ON SOCIOECONOMIC DISPARITIES

Outcome	Coefficient on coverage RD (1)	R ² (2)	Disparities at ages 63–64			Percent change in disparity due to change in coverage at 65		
			Low-ed minority-Hi-ed Whites (3)	Black-White (4)	Hispanic-White (5)	Low-ed minority-Hi-ed Whites (6)	Black-White (7)	Hispanic-White (8)
<i>Panel A: Based on change in insurance coverage at 65</i>								
Delay in care last year	-0.19 (0.06)	0.72	7.6	—	—	-42.8	—	—
No care last year	-0.12 (0.06)	0.39	8.0	—	—	-25.6	—	—
Regular doctor visit last year	0.32 (0.09)	0.77	-7.4	—	—	-73.9	—	—
Hospital stay last year	-0.09 (0.08)	0.26	4.7	—	—	-32.7	—	—
Total hospital admissions	0.06 (0.18)	0.74	—	724	-193	—	1.8	-0.5
Diagnostic procedures of the heart	0.59 (0.31)	0.62	—	9	-25	—	35.6	4.5
Bypass anastomosis of heart	-0.54 (0.93)	0.54	—	-18	-14	—	-4.9	-4.8
Joint replacement of lower extremity	-0.19 (0.64)	0.89	—	-7	-28	—	2.7	-2.3
<i>Panel B: Based on change in incidence of multiple coverage at 65</i>								
Total hospital admissions	0.03 (0.08)	0.74	—	724	-193	—	-0.1	5.8
Diagnostic procedures of the heart	-0.21 (0.14)	0.55	—	9	-25	—	22.3	-18.7
Bypass anastomosis of heart	0.46 (0.34)	0.64	—	-18	-14	—	29.4	37.1
Joint replacement lower of extremity	0.42 (0.21)	0.94	—	-7	-28	—	59.4	25.7

Notes: Each entry in panel A, column 1, is estimated coefficient from regression of RDs in listed health outcome on RDs in insurance coverage over six ethnicity/education groups (rows 1–4) or nine state-ethnicity groups (rows 5–8). All regressions weighted by the inverse sampling variance of the estimated discontinuity in each outcome, and regressions in rows 5–8 include state dummies. Entries in column 2 are corresponding R-squared coefficients from each regression. Entries in columns 3, 4, and 5 are the observed disparities in each health outcome at ages 63–64, and entries in columns 6, 7, and 8 are the percent change in the disparity attributable to the change in insurance coverage based on the coefficient in column 1. Health disparities measured in the NHIS are characterized in terms of low-ed minorities versus hi-ed whites, whereas health disparities measured in the hospital discharge data are characterized in terms of black-white or hispanic-white differences. Panel B is similar to panel A except that the RDs in each health outcome are regressed on the RDs in the incidence of multiple coverage at 65. Panel B regressions are based on data for New York and Florida only (i.e., six state-ethnicity groups).

Summary of Card, et al. (2009)

- David Card, Carlos Dobkin and Nicole Maestas (2009), “Does Medicare Save Lives?”, *American Economic Review*

Impact of Medicare on admissions

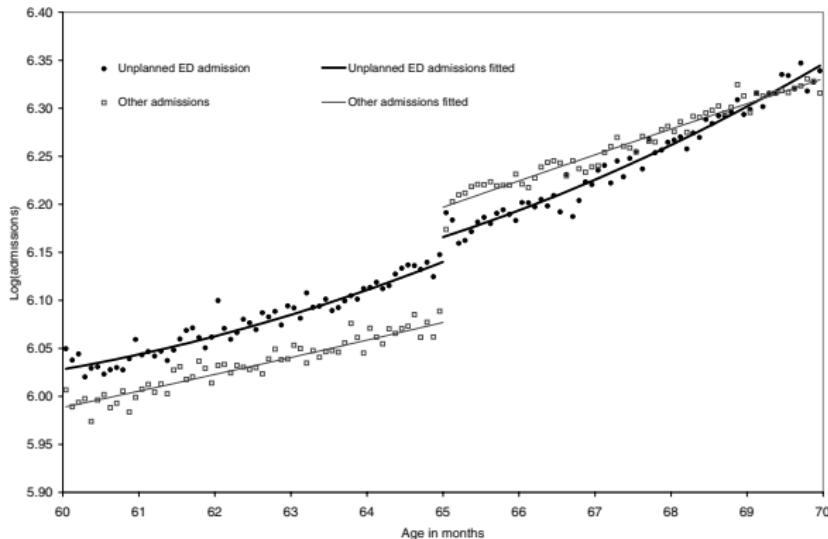


FIGURE II

Number of Admissions by Route into Hospital, California, 1992–2002

The lines are fitted values from regressions that include a second-order polynomial in age fully interacted with a dummy for age ≥ 65 and a dummy variable for the month before people turn 65. The dependent variable is the log of the number of admissions by patient's age (in days) at admission, for patients between 60 and 70 years of age. The count of admissions is based on hospital discharge records for California and includes admissions from January 1, 1992, to November 30, 2002. The points represent means of the dependent variable for 30-day cells. The age profile for unplanned ED admissions includes admissions that occurred through the emergency department and were unplanned. The category "Other Admissions" includes all other admissions.

Proportion with coverage

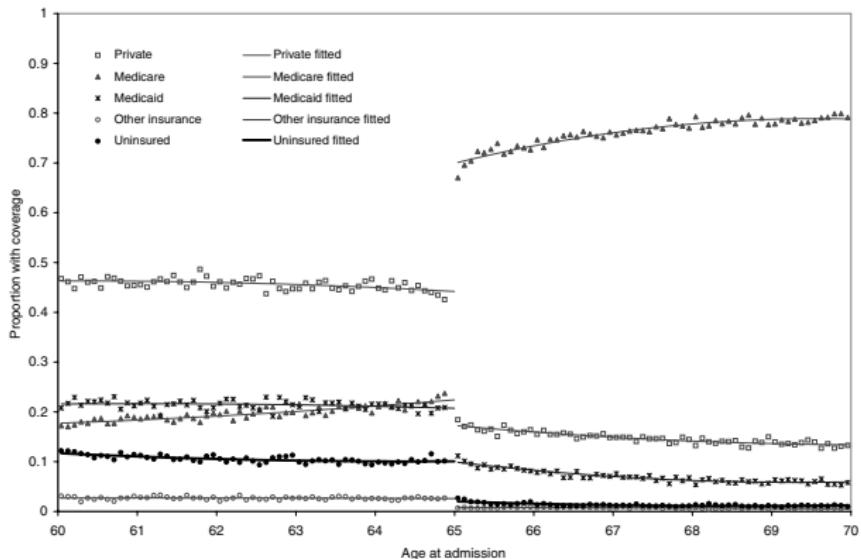


FIGURE IV
Primary Insurance Coverage of Admitted Patients

See notes for Figure II. In this figure the y-axis represents the fraction of patients with different classes of primary insurance coverage. Sample includes 425,315 patients with nondeferrable primary diagnoses, defined as unplanned admissions through the emergency department for diagnoses with a t -statistic for the test of equal weekday and weekend admission rates of 0.965 or less. Medicare eligibility status of patients within one month of their 65th birthdays is uncertain and we have excluded these observations.

Impact of Medicare on type of coverage

TABLE III
REGRESSION DISCONTINUITY MODELS FOR PROBABILITY OF DIFFERENT FORMS OF PRIMARY INSURANCE COVERAGE

	Medicare		Private		Medicaid		Uninsured	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age over 65 ($\times 100$)	43.9 (0.4)	47.5 (0.4)	-24.8 (0.4)	-26.8 (0.4)	-10.1 (0.3)	-10.8 (0.3)	-7.4 (0.2)	-8.0 (0.2)
Additional controls	No	Yes	No	Yes	No	Yes	No	Yes
Mean of dependent variable for patients aged 64–65 ($\times 100$)	24.0		43.3		43.3		9.7	

Notes. Standard errors in parentheses. Dependent variable is indicator for type of insurance listed as "primary insurer" on discharge record. Sample includes 425,315 observations on patients between the ages of 60 and 70 admitted to California hospitals between January 1, 1992, and November 30, 2002 for an unplanned admission through the emergency department, with a diagnosis (ICD-9) for which the *t*-test for equality of weekend and weekday admission rates is less than 0.96 in absolute value. All models include second-order polynomial in age (in days) fully interacted with dummy for age over 65 and are fit by OLS. Models in even-numbered columns include the following additional controls: a dummy for people who are within one month of their 65th birthday; dummies for month, year, sex, race/ethnicity, and admission on Saturday or Sunday; and a complete set of unrestricted fixed effects for each ICD-9 admission diagnosis. In columns (1)–(8) the coefficient on "age over 65" and its standard error have been multiplied by 100.

Treatment intensity

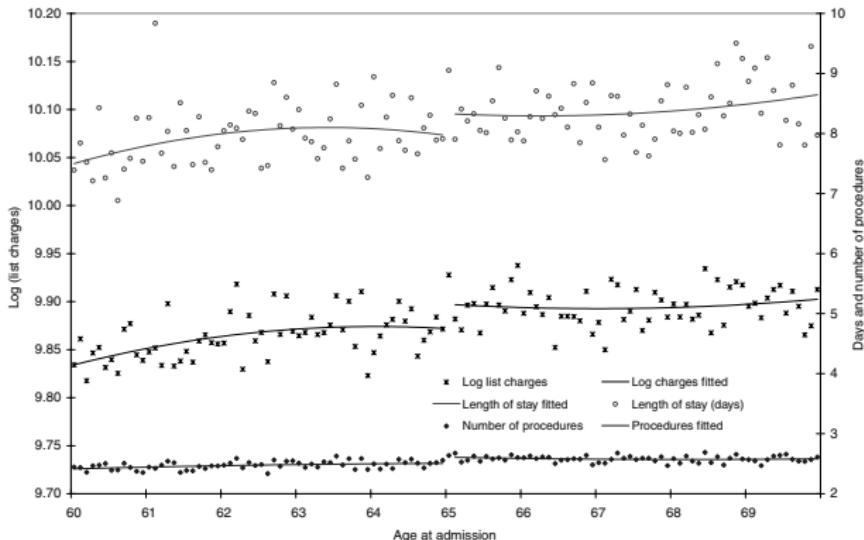


FIGURE V
Three Measures of Inpatient Treatment Intensity

See notes to Figure IV. Sample includes unplanned admissions through the emergency department for diagnoses with a *t*-statistic for the test of equal weekday and weekend admission rates of 0.965 or less. In this figure the sample is further restricted to patients with valid SSNs (407,386 observations). Sample for log list charges excludes patients admitted to Kaiser hospitals. Length of stay, number of procedures, and list charges are cumulated over all consecutive hospitalizations. List charges are measured in 2002 dollars.

Treatment intensity

TABLE IV
REGRESSION DISCONTINUITY MODELS FOR CHANGES IN TREATMENT INTENSITY

	Length of stay (days)		Number of procedures		Log list charges ($\times 100$)	
	(1)	(2)	(3)	(4)	(5)	(6)
Age over 65	0.37 (0.24)	0.35 (0.26)	0.09 (0.03)	0.11 (0.03)	2.5 (1.1)	2.6 (1.0)
Additional controls	No	Yes	No	Yes	No	Yes
Mean of dependent variable for patients aged 63 or 64	8.12		2.50		9.87	

Notes. Standard errors in parentheses. Dependent variable is length of stay in days (columns (1) and (2)), number of procedures performed (columns (3) and (4)), and log of total list charges (columns (5) and (6)). Sample includes 407,386 (352,652 in columns (5) and (6)) observations on patients with valid SSNs between the ages of 60 and 70 admitted to California hospitals between January 1, 1992, and November 30, 2002 for an unplanned admission through the ED. Data on list charges are missing for Kaiser hospitals. See note to Table III for additional details on sample, and list of additional covariates included in even-numbered columns. In columns (5) and (6) the coefficient on "age over 65" and its standard error have been multiplied by 100.

Mortality

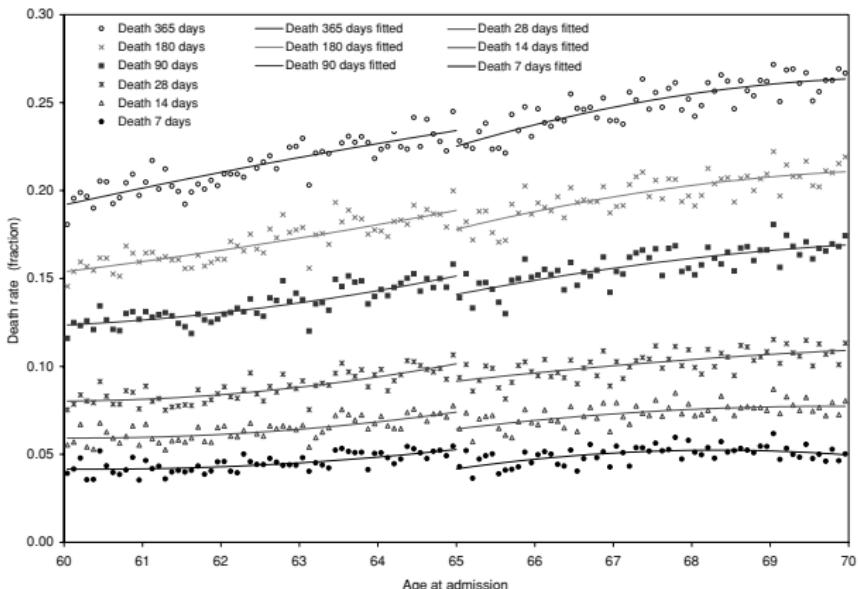


FIGURE VI
Patient Mortality Rates over Different Follow-Up Intervals

See notes to Figure IV. Sample includes unplanned admissions through the emergency department for diagnoses with a t -statistic for the test of equal weekday and weekend admission rates of 0.965 or less. In this figure the sample is further restricted to patients with valid SSNs (407,386 observations). Deaths include in-hospital and out-of-hospital deaths.

Mortality

TABLE V
REGRESSION DISCONTINUITY ESTIMATES OF CHANGES IN MORTALITY RATES

	Death rate in					
	7 days	14 days	28 days	90 days	180 days	365 days
<i>Estimated discontinuity at age 65 ($\times 100$)</i>						
Fully interacted quadratic with no additional controls	-1.1 (0.2)	-1.0 (0.2)	-1.1 (0.3)	-1.1 (0.3)	-1.2 (0.4)	-1.0 (0.4)
Fully interacted quadratic plus additional controls	-1.0 (0.2)	-0.8 (0.2)	-0.9 (0.3)	-0.9 (0.3)	-0.8 (0.3)	-0.7 (0.4)
Fully interacted cubic plus additional controls	-0.7 (0.3)	-0.7 (0.2)	-0.6 (0.4)	-0.9 (0.4)	-0.9 (0.5)	-0.4 (0.5)
Local linear regression procedure fit separately to left and right with rule-of-thumb bandwidths	-0.8 (0.2)	-0.8 (0.2)	-0.8 (0.2)	-0.9 (0.2)	-1.1 (0.3)	-0.8 (0.3)
Mean of dependent variable (%)	5.1	7.1	9.8	14.7	18.4	23.0

Notes. Standard errors in parentheses. Dependent variable is indicator for death within interval indicated by column heading. Entries in rows (1)–(3) are estimated coefficients of dummy for age over 65 from models that include a quadratic polynomial in age (rows (1) and (2)) or a cubic polynomial in age (row (3)) fully interacted with a dummy for age over 65. Models in rows (2) and (3) include the following additional controls: a dummy for people who are within 1 month of their 65 birthdays, dummies for year, month, sex, race/ethnicity, and Saturday or Sunday admissions, and unrestricted fixed effects for each ICD-9 admission diagnosis. Entries in row (4) are estimated discontinuities from a local linear regression procedure, fit separately to the left and right, with independently selected bandwidths from a rule-of-thumb procedure suggested by Fan and Gijbels (1996). Sample includes 407,386 observations on patients between the ages of 60 and 70 admitted to California hospitals between January 1, 1992, and November 30, 2002, for unplanned admission through the ED who have nonmissing Social Security numbers. All coefficients and their SEs have been multiplied by 100.

Probabilistic treatment assignment (i.e. "fuzzy RDD")

The probability of receiving treatment changes discontinuously at the cutoff, X_0 , but need not go from 0 to 1

$$\lim_{X_i \rightarrow X_0} \Pr(D_i = 1 | X_i = X_0) \neq \lim_{X_0 \leftarrow X_i} \Pr(D_i = 1 | X_i = X_0)$$

Examples: Incentives to participate in some program may change discontinuously at the cutoff but are not powerful enough to move everyone from non participation to participation.

- In the sharp RDD, D_i was *determined* by $X_i \geq X_0$; in the fuzzy RDD, the conditional probability of treatment *jumps* at X_0 .
- The relationship between the probability of treatment and X_i can be written as:

$$P[D_i = 1 | X_i] = g_0(X_i) + [g_1(X_i) - g_0(X_i)] T_i$$

where $T_i = 1$ if $(X_i \geq X_0)$ and 0 otherwise.

Visualization of Fuzzy RDD identification strategy

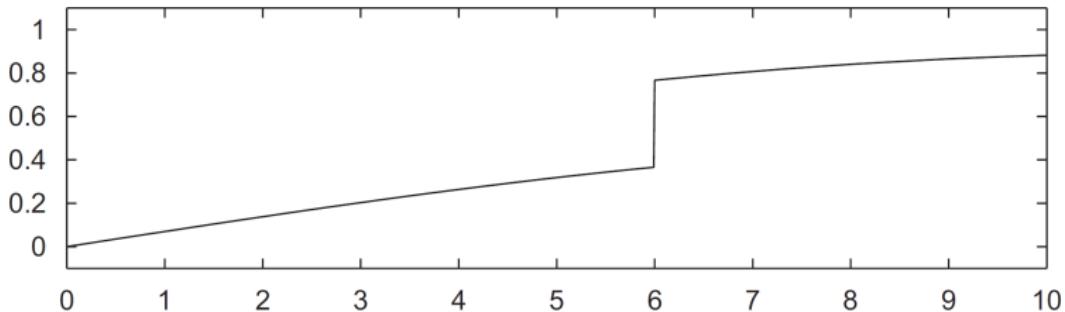


Fig. 3. Assignment probabilities (FRD).

Figure: Imbens and Lemieux (2007), figure 3. Horizontal axis is the running variable. Vertical axis is the conditional probability of treatment at each value of the running variable.

Visualization of identification strategy (i.e. smoothness)

- Conditional expectation of the 2 potential outcomes given X for $D = 0, 1$
 - Partly dashed, partly solid \rightarrow continuous functions of the covariate
- Conditional expectation of the observed outcome, Y
 - Solid \rightarrow jumps at $X = X_0$

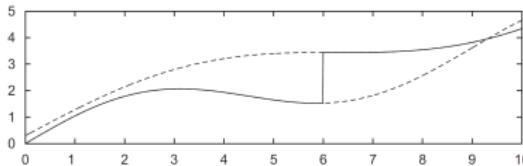


Figure: Potential and observed outcome regressions (Imbens and Lemieux 2007)

Use the discontinuity as IV

Wald estimator of treatment effect under Fuzzy RDD

Average causal effect of the treatment is the Wald IV parameter

$$\delta_{\text{Fuzzy RDD}} = \frac{\lim_{X \rightarrow X_0} E[Y|X = X_0] - \lim_{X_0 \leftarrow X} E[Y|X = X_0]}{\lim_{X \rightarrow X_0} E[D|X = X_0] - \lim_{X_0 \leftarrow X} E[D|X = X_0]}$$

- Fuzzy RDD is numerically equivalent and conceptually similar to instrumental variables
 - Numerator: “jump” in the regression of the outcome on the running variable, X . “Reduced form”.
 - Denominator: “jump” in the regression of the treatment indicator on the running variable X . “First stage”.
- Use software package to estimate (e.g., `ivregress 2sls` in STATA).
- Same IV assumptions, caveats about compliers vs. defiers, and statistical tests that we discussed with instrumental variables apply here – e.g., check for weak instruments using F test on instrument in first stage, etc.

First stage relationship between X and D

- One can use both T_i as well as the interaction terms as instruments for D_i . If one uses only T_i as IV, then it is a “just identified” model which usually has good finite sample properties.
- In the just identified case, the first stage would be:

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \cdots + \gamma_p X_i^p + \pi T_i + \zeta_{1i}$$

where π is the causal effect of T on the conditional probability of treatment.

- The fuzzy RD reduced form is:

$$Y_i = \mu + \kappa_1 X_i + \kappa_2 X_i^2 + \cdots + \kappa_p X_i^p + \rho \pi T_i + \zeta_{2i}$$

Fuzzy RDD with varying Treatment Effects - Second Stage

- As in the sharp RDD case one can allow the smooth function to be different on both sides of the discontinuity.
- The second stage model with interaction terms would be the same as before:

$$Y_i = \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p + \rho D_i + \beta_1^* D_i \tilde{x}_i + \beta_2^* D_i \tilde{x}_i^2 + \cdots + \beta_p^* D_i \tilde{x}_i^p + \eta_i$$

- Where \tilde{x} are now not only normalized with respect to x_0 but are also fitted values obtained from the first stage regression.

Fuzzy RDD with Varying Treatment Effects - First Stages

- Again one can use both T_i as well as the interaction terms as instruments for D_i ;
- Only using T the estimated first stages would be:

$$D_i = \gamma_{00} + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p + \pi T_i + \gamma_1^* \tilde{X}_i T_i + \gamma_2^* \tilde{X}_i^2 T_i + \cdots + \gamma_p^* T_i + \zeta_{1i}$$

- We would also construct analogous first stages for $\tilde{X}_i D_i$, $\tilde{X}_i^2 D_i$, \dots , $\tilde{X}_i^p D_i$.

What does Fuzzy RDD Estimate?

- As Hahn, Todd and van der Klaauw (2001) point out, one needs the same assumptions as in the standard IV framework
- As with other binary IVs, the fuzzy RDD is estimating LATE: the average treatment effect for the compliers
- In RDD, the compliers are those whose treatment status changed as we moved the value of x_i from just to the left of x_0 to just to the right of x_0

Challenges to RDD

- Treatment is not as good as randomly assigned around the cutoff, X_0 , when agents are able to manipulate their running variable scores. This happens when:
 - ① the assignment rule is known in advance
 - ② agents are interested in adjusting
 - ③ agents have time to adjust
 - ④ Examples: re-take an exam, self-reported income, etc.
- Some other unobservable characteristic changes at the threshold, and this has a direct effect on the outcome.
 - In other words, the cutoff is endogenous
 - Example: Age thresholds used for policy (i.e., person turns 18, and faces more severe penalties for crime) is correlated with other variables that affect the outcome (i.e., graduation, voting rights, etc.)

Econometricians and applied social scientists have developed several formalized tests to evaluate the severity of these problems which we now discuss.

Test 1: Manipulation of the running variable

Sorting on the running variable (i.e., Manipulation)

Assume a desirable treatment, D , and an assignment rule $X \geq X_0$. If individuals sort into D by choosing X such that $X \geq X_0$, then we say individuals are sorting on the running variable.

- Motivating example: Suppose a doctor plans to randomly assign heart patients to a statin and a placebo to study the effect of the statin on heart attacks within 10 years. The doctor randomly assigns patients to two different waiting rooms, A and B , and plans to give those in A the statin and those in B the placebo. If some of the patients learn of the planned treatment assignment mechanism, what would we expect to happen? And how would you check for it?

McCrary Density Test

We would expect waiting room A to become *crowded*. In the RDD context, sorting on the running variable implies heaping on the “good side” of X_0

- McCrary (2008) suggests a formal test. Under the null the density should be continuous at the cutoff point. Under the alternative hypothesis, the density should increase at the kink (where D is viewed as good)
 - ① Partition the assignment variable into bins and calculate frequencies (i.e., number of observations) in each bin
 - ② Treat those frequency counts as dependent variable in a local linear regression
- The McCrary Density Test has become **mandatory** for every analysis using RDD.
 - If you can estimate the conditional expectations, you evidently have data on the running variable. So in principle you can always do a density test
 - You can download the (no longer supported) STATA ado package, DCdensity, to implement McCrary’s density test
(<http://eml.berkeley.edu/~jmccrary/DCdensity/>)
 - You can install rdd for R too
(<http://cran.r-project.org/web/packages/rdd/rdd.pdf>)

Caveats about McCrary Density Test

- For RDD to be useful, you already need to know something about the mechanism generating the assignment variable and how susceptible it could be to manipulation. Note the rationality of economic actors that this test is built on.
- A discontinuity in the density is “suspicious” – it *suggests* manipulation of X around the cutoff is probably going on. In principle one doesn’t need continuity.
- This is a high-powered test. You need a lot of observations at X_0 to distinguish a discontinuity in the density from noise.

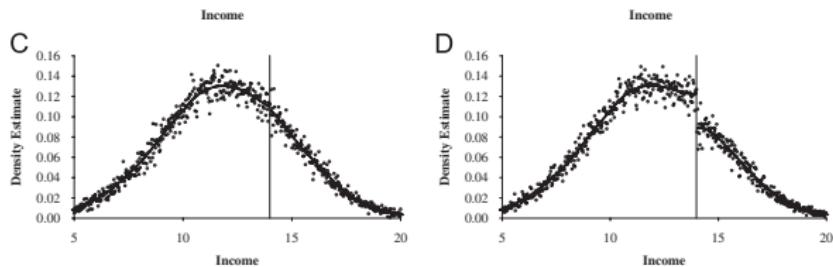
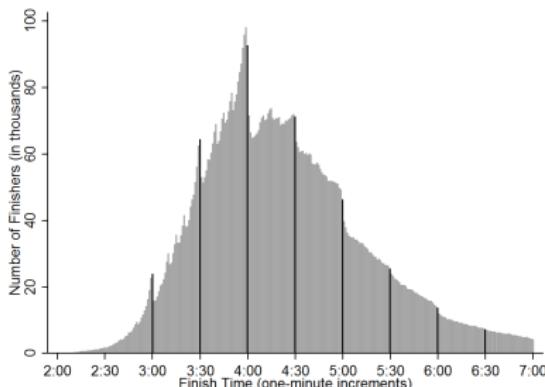


Figure: Panel C is density of income when there is no pre-announcement and no manipulation. Panel D is the density of income when there is pre-announcement and manipulation. From McCrary (2008).

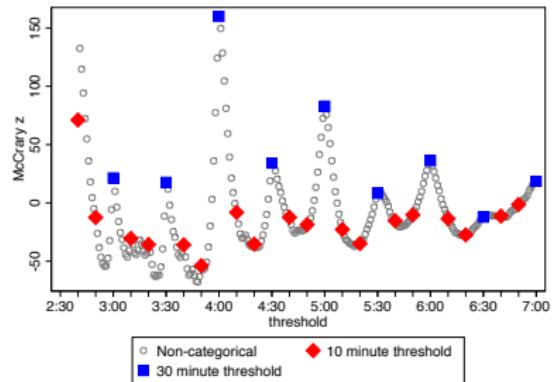
Visualizing manipulation

Figure 2: Distribution of marathon finishing times ($n = 9,378,546$)



NOTE: The dark bars highlight the density in the minute bin just prior to each 30 minute threshold.

Figure 3: Running McCrary z -statistic



NOTE: The McCrary test is run at each minute threshold from 2:40 to 7:00 to test whether there is a significant discontinuity in the density function at that threshold.

Figure: Figures 2 and 3 from Eric Allen, Patricia Dechow, Devin Pope and George Wu's (2013)
"Reference-Dependent Preferences: Evidence from Marathon Runners".

http://faculty.chicagobooth.edu/devin.pope/research/pdf/Website_Marathons.pdf

More examples of testing for manipulation: Lee (2008) Inc incumbency Effect

- David Lee's (2008) "Randomized Experiments from non-random selection in US House Elections", *Journal of Econometrics* analyzes the incumbency effect using Democratic incumbents for US congressional elections.
 - A large political science literature on the "incumbency advantage" – having won an election once helps you win subsequent elections.
 - Empirical challenge: how to separate incumbency advantage from selection – (i.e., candidates win multiple elections because they are better)
 - Identification: Inc incumbency is assigned to a candidate discontinuously at 50% voteshare under two-party democracy with majority rule.
- Lee (2008) analyzes the probability of winning the election in year $t + 1$ by comparing candidates who just won to candidates who just lost the election in year t .

Density test

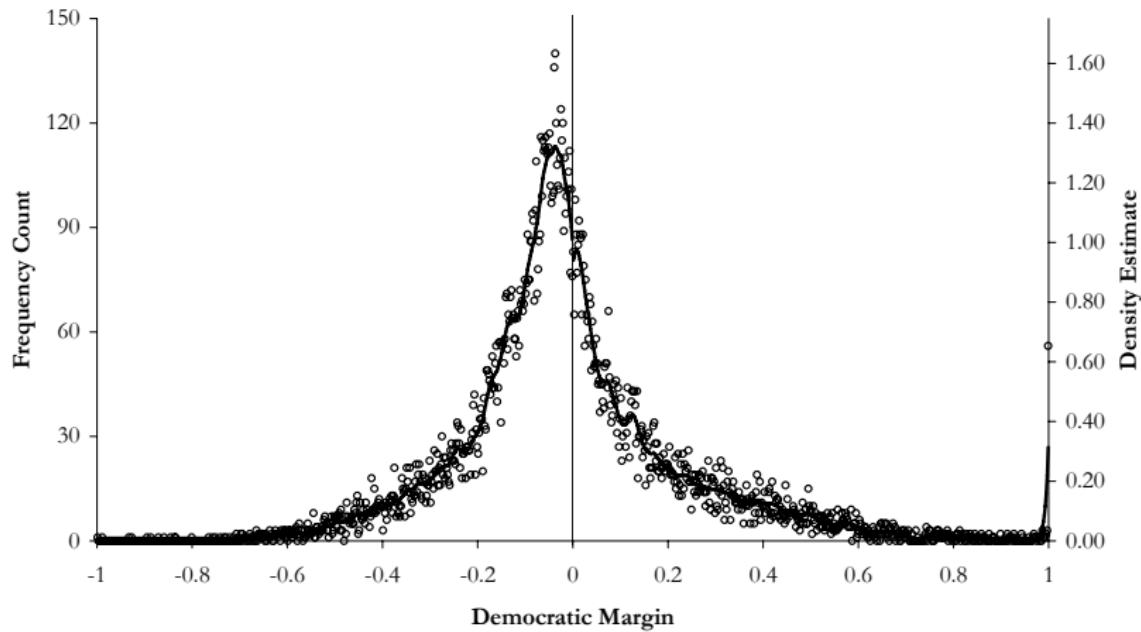


Figure: Democratic vote share relative to cutoff: popular elections to the House of Representatives, 1900-1990 (McCrary 2008).

More evidence of manipulation

Contrast this with roll call voting in the US House of Representatives

- Coordination is expected because these are repeated games, votes are public records, and side payments are possible in the form of future votes
- Bills around the cutoff are more likely to be passed than not. Seems like a good candidate for RDD
- Fails McCrary Density Test; **cannot** use RDD because policy decisions are not quasi-randomly assigned around cutoff

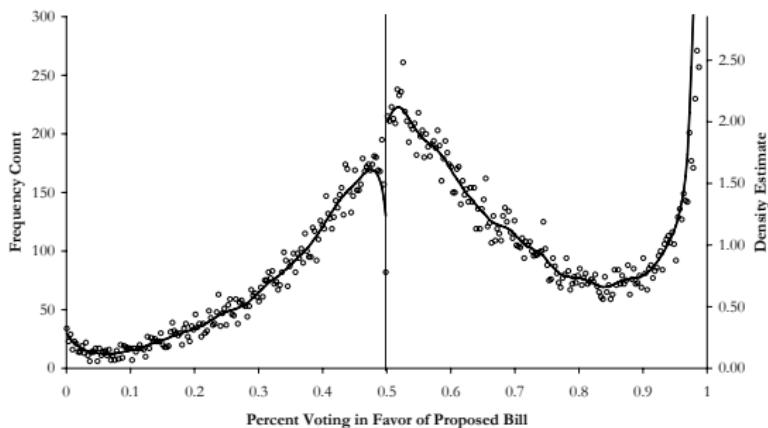


Figure: Percent voting Yeay: Roll Call Votes, US House of Representatives, 1857-2004
(McCrary 2008)

Test 2: Balance test on covariates

- This is a type of placebo test. For RDD to be valid in your study, you do not want to observe a discontinuity around the cutoff, X_0 , for average values of covariates that should **not** be affected by the treatment – e.g., pretreatment characteristics
- Question: What does a jump in the average values of pre-treatment characteristics have to do with the continuous (smoothness) assumption?
 - Choose other pre-treatment covariates, Z and do a similar graphical plot as you did for Y
 - You do **not** want to see a jump around the cutoff, X_0
 - A formal balance test involves the same procedure used to estimate the treatment effect, only use Z instead of Y as a LHS variable
 - Can combine test on multiple covariates into a single test statistic using seemingly unrelated regression (SUR) or a single “stacked” regression

Visualizing Placebo Test

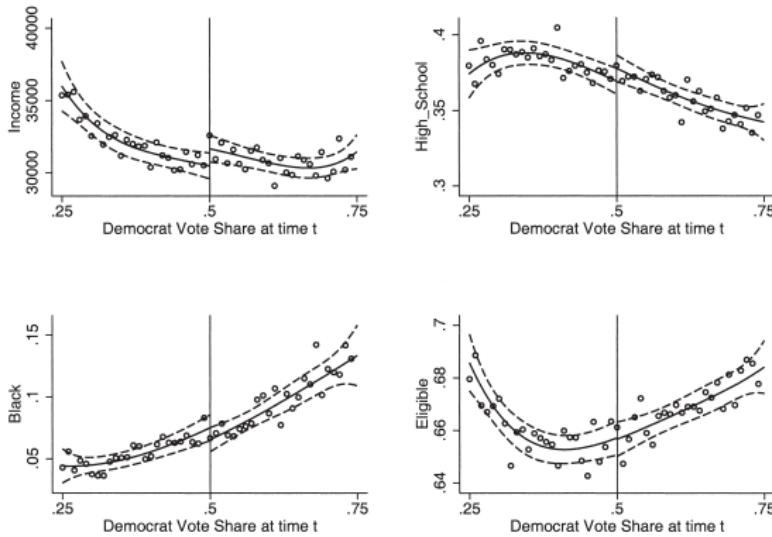


Figure: Figure 3 from Lee, Moretti and Butler (2004), "Do Voters Affect or Elect Policies?" *Quarterly Journal of Economics*. Panels refer to (top left to bottom right) the following district characteristics: real income, percentage with high-school degree, percentage black, percentage eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democratic vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line represents the 95 percent confidence interval.

Test 3: Jumps at non-discontinuous points

- Imbens and Lemieux (2008) suggest to look at one side of the discontinuity (e.g., $X < X_0$), take the median value of the running variable in that section, and pretend it was a discontinuity, X'_0
- Then test whether in reality there is a discontinuity at X'_0 . You do **not** want to find anything.
- Another kind of placebo test. Similar to synthetic control falsification exercises where you look for non-effects in non-treatment periods. (See Abadie, Diamond and Hainmueller 2014)

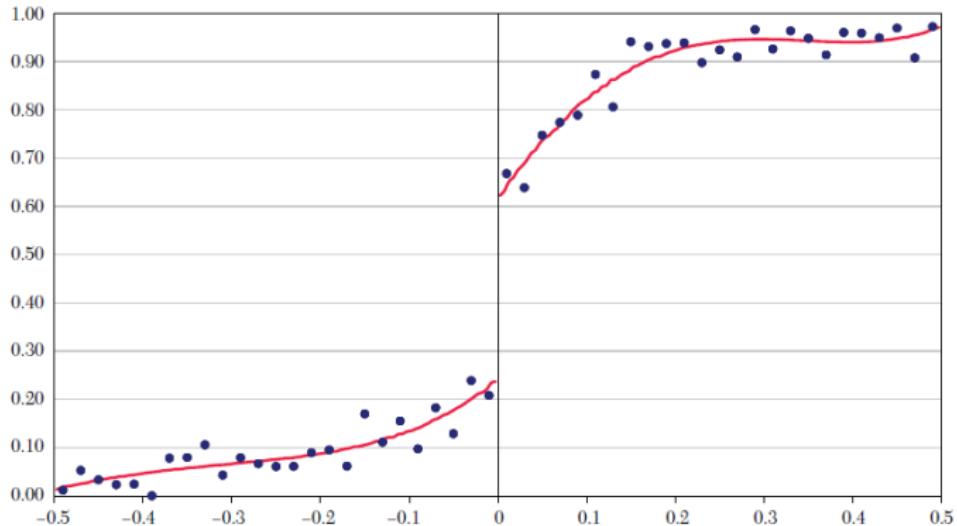
RDD graphs

A graphical analysis should be an integral part of any RDD.

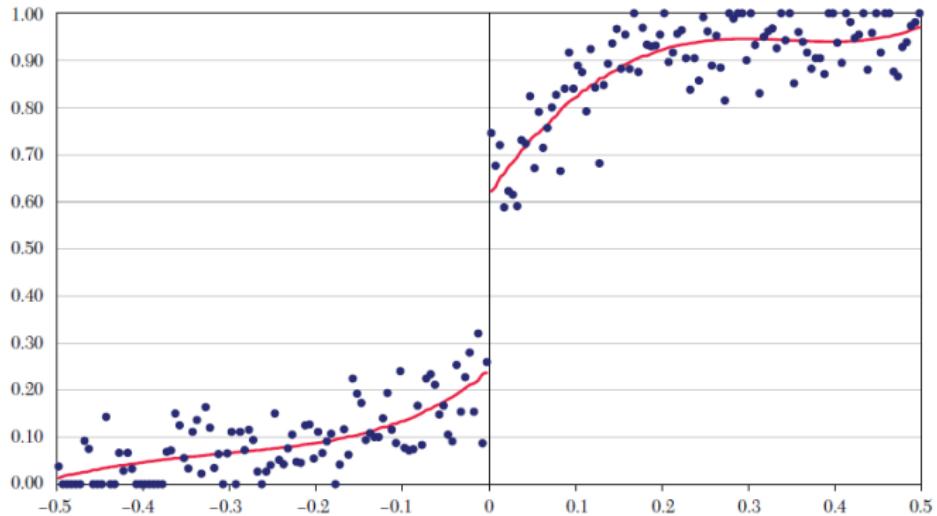
① Outcome by running variable, (X_i):

- The standard graph showing the discontinuity in the outcome variable
- Construct bins and average the outcome within bins on both sides of the cutoff
- You should also look at different bin sizes when constructing these graphs (see Lee and Lemieux, 2010, for details)
- Plot the running variables, X_i , on the horizontal axis and the average of Y_i for each bin on the vertical axis
- You may also want to plot a relatively flexible regression line on top of the bin means
- Inspect whether there is a discontinuity at x_0
- Inspect whether there are other unexpected discontinuities

Example: Outcomes by Running Variables



Example: Outcomes by Running Variables with smaller bins



More RDD Graphs

② Probability of treatment by running variable if fuzzy RDD

- In a fuzzy RDD, you also want to see that the treatment variable jumps at x_0
- This tells you whether you have a first stage

③ Covariates by a running variable

- Construct a similar graph to the one before but using a covariate as the “outcome”
- There should be **no jump** in other covariates at the discontinuity, x_0 .
- If the covariates jump at the discontinuity, one would doubt the identifying assumption

Example: Covariates by Running Variable

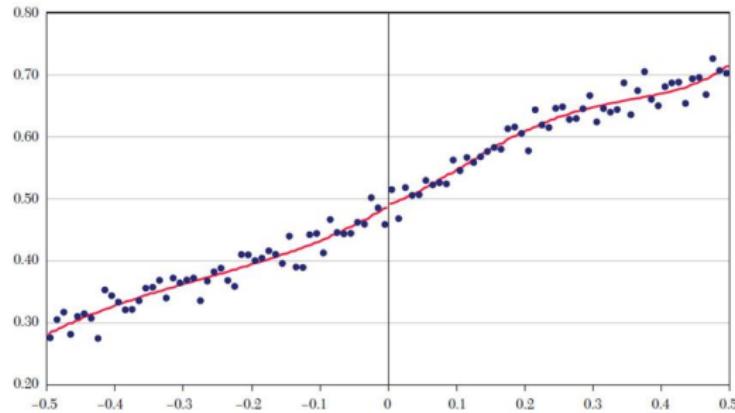


Figure 17. Discontinuity in Baseline Covariate (Share of Vote in Prior Election)

More RDD graphs!

④ The density of the running variable

- One should plot the number of observations in each bin.
- This plot allows to investigate whether there is a discontinuity in the distribution of the running variable at the threshold
- This would suggest that people can manipulate the running variable around the threshold.
- This is an indirect test of the identifying assumption that each individual has imprecise control over the assignment variable

Density of the running variable

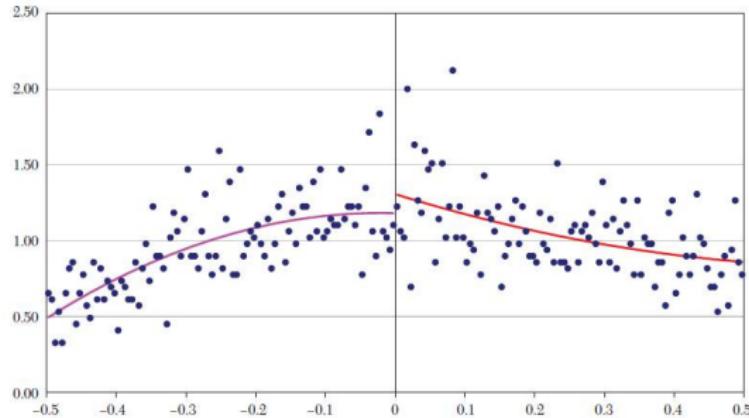


Figure 16. Density of the Forcing Variable (Vote Share in Previous Election)

“Do Voters Affect or Elect Policies?”

by Lee, Moretti and Butler (2004)

How do voters affect policy? There are two fundamentally different views of the role of elections in a representative democracy.

- ① Convergence:** Heterogenous voter ideology forces each candidates to moderate their positions (e.g., median voter theorem).

“Competition for votes can force even the most partisan Republicans and Democrats to moderate their policy choices. In the extreme case, competition may be so strong that it leads to ‘full policy convergence’: opposing parties are forced to adopt identical policies” (Lee, Moretti, and Butler 2004; Downs 1957).

- ② Divergence:** When partisan politicians cannot credibly commit to certain policies, then convergence is undermined. The result can be full policy divergence. Divergence is when the winning candidate, after taking office, simply pursues his most-preferred policy. In this case, voters fail to compel candidates to reach any kind of policy compromise.

Simplified model

- **Candidates:** R and D are candidates in a Congressional race. Policy space is unidimensional where D 's and R 's per-period policy preferences are quadratic loss functions, $u(I)$ and $v(I)$ and I is the policy variable

D 's bliss point: $I^* = c (> 0)$

$$\begin{aligned}\max_I u(I) &= -\frac{1}{2}(I - c)^2 \\ \frac{\partial u(I)}{\partial I} &= 2 \times -\frac{1}{2}(I - c) = 0 \\ I^* &= c (> 0)\end{aligned}$$

R 's bliss point: $I^* = 0$

$$\begin{aligned}\max_I v(I) &= -\frac{1}{2}I^2 \\ \frac{\partial v(I)}{\partial I} &= -\frac{1}{2} \times 2I = 0 \\ I^* &= 0\end{aligned}$$

- **Voters:** *Ex ante*, voters expect D (R) to choose policy x^e (y^e), and expect D to win with probability $P(x^e, y^e)$
 - When $x^e > y^e$ then $\frac{\partial P}{\partial x^e} > 0$, $\frac{\partial P}{\partial y^e} < 0$. Democrats could gain more votes by moderating the policy position.
 - P^* represents the underlying popularity of D party, or put differently, the probability D would win if $x = c$ and $y = 0$.
- if D (R) wins, policy x (y) will be implemented; rational expectations equilibrium assumed ($x = x^e$; $y = y^e$), game repeats in periods t and $t + 1$ where a period includes the election and the Congressional session.

Multiple Nash equilibria

1 Partial/Complete Convergence: Voters affect policies

- Key result: $\frac{\partial x^*}{\partial P^*} > 0$ and $\frac{\partial y^*}{\partial P^*} > 0$
- If we did a “helicopter drop” of more Democrats in the district to exogenously increase P^* and this resulted in candidates changing their policy positions, then $\frac{\partial x^*}{\partial P^*} > 0$ and $\frac{\partial y^*}{\partial P^*} > 0$

2 Complete divergence: Voters elect politicians with fixed policies

- Key result: $\frac{\partial x^*}{\partial P^*} = \frac{\partial y^*}{\partial P^*} = 0$
- An exogenous shock to P^* (e.g., the helicopter drop of Democrats) does *nothing* to equilibrium policies. Voters *elect* politicians' fixed policies.

Estimation

- Potential roll-call voting record outcomes of the representative in the district following election t is

$$RC_t = D_t x_t + (1 - D_t) y_t$$

where D_t is an indicator for whether D won the election. That is, only the winning candidate's policy is observed.

- This expression can be transformed into regression equations:

$$RC_t = \alpha_0 + \pi_0 P_t^* + \pi_1 D_t + \varepsilon_t \quad (95)$$

$$RC_{t+1} = \beta_0 + \pi_0 P_{t+1}^* + \pi_1 D_{t+1} + \varepsilon_{t+1} \quad (96)$$

where α_0 and β_0 are constants

- They parameterized the derivatives from the prior slide as π_0 , the coefficient on the probability, P . See their Appendix 1, p. 849 for details.
- This also allows an independent effect of the party, π_1

Estimation (cont.)

- Equation 198 cannot be directly estimated because we don't observe P^* . But suppose we could randomize D_t . Then D_t would be independent of P_t^* and ε_t . Then taking conditional expectations with respect to D_t we get:

$$\underbrace{E[RC_{t+1}|D_t = 1] - E[RC_{t+1}|D_t = 0]}_{\text{Observable}} = \pi_0[P_{t+1}^{*D} - P_{t+1}^{*R}] + \underbrace{\pi_1[P_{t+1}^D - P_{t+1}^R]}_{\text{Observable}} = \underbrace{\gamma}_{\text{Total effect of initial win on future roll call votes}} \quad (97)$$

$$\underbrace{E[RC_t|D_t = 1] - E[RC_t|D_t = 0]}_{\text{Observable}} = \pi_1 \quad (98)$$

$$\underbrace{E[D_{t+1}|D_t = 1] - E[D_{t+1}|D_t = 0]}_{\text{Observable}} = P_{t+1}^D - P_{t+1}^R \quad (99)$$

Why this works

- The “elect” component is $\pi_1[P_{t+1}^D - P_{t+1}^R]$ and it’s estimated as the difference in mean voting records between the parties at time t
- The fraction of districts won by Democrats in $t + 1$ is an estimate of $[P_{t+1}^D - P_{t+1}^R]$
- Because we can estimate the total effect, γ , of a Democrat victory in t on RC_{t+1} , we can net out the elect component to implicitly get the “affect” component
- Random assignment of D_t is crucial. Without it, equation 206 would reflect π_1 and selection (i.e., that Dem districts have more liberal bliss points).

Data and RDD Jargon

- Two main datasets: liberal voting score from the Americans for Democratic Action (ADA) linked with House of Representatives election results for 1946-1995
 - Authors use the ADA score for all US House Representatives from 1946 to 1995 as their voting record index
 - For each Congress, ADA chooses about twenty high-profile roll-call votes and creates an index varying 0 and 100 for each Representative of the House. Higher scores correspond to a more “liberal” voting record.
- RDD Jargon
 - The running variable is `voteshare` which is the share of all votes that went to a Democrat. ADA scores are linked to election returns data during that period.
 - They use exogenous variation in Democratic wins to check whether convergence or divergence is correct.
 - Discontinuity in the running variable occurs at `voteshare= 0.5`. When `voteshare > 0.5`, the Democratic candidate wins.
- Download STATA do file from my website: http://business.baylor.edu/scott_cunningham/teaching/causalinf/lmb2004.txt, save it on your computer, and rename the extension to say `lmb2004.do`. Open it in your STATA do-editor.

Statistical results

TABLE I
RESULTS BASED ON ADA SCORES—CLOSE ELECTIONS SAMPLE

Variable	Total effect			Elect component	Affect component
	γ	$\pi_1 (P_{t+1}^D - P_{t+1}^R)$	$\pi_1 [(P_{t+1}^D - P_{t+1}^R)]$	$\pi_0 [P_{t+1}^{SD} - P_{t+1}^{SR}]$	
	ADA_{t+1}	ADA_t	DEM_{t+1}	(col. (2) π_1 (col. (3))	(col. (1)) – (col. (4))
	(1)	(2)	(3)	(4)	(5)
Estimated gap	21.2 (1.9)	47.6 (1.3)	0.48 (0.02)	22.84 (2.2)	-1.64 (2.0)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample size is 915.

Figure: Lee, Moretti, and Butler 2004, Table 1.

```
. ** Tables - using OLS
. * Table 1: Results Based on ADA Scores
. reg score lagdemocrat if lagdemvoteshare>.48 & lagdemvoteshare<.52, cluster(id)
```

```
Linear regression                                         Number of obs =      915
<snip>
```

score	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lagdemocrat	21.28387	1.951234	10.91	0.000	17.45445	25.11329
_cons	31.19576	1.333983	23.39	0.000	28.57773	33.81378

```
. reg score democrat if lagdemvoteshare>.48 & lagdemvoteshare<.52, cluster(id)
```

```
Linear regression                                         Number of obs =      915
<snip>
```

score	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
democrat	47.7056	1.356011	35.18	0.000	45.04434	50.36686
_cons	18.7469	.8432428	22.23	0.000	17.09198	20.40182

```
. reg democrat lagdemocrat if lagdemvoteshare>.48 & lagdemvoteshare<.52, cluster(id)
```

```
Linear regression                                         Number of obs =      915
<snip>
```

democrat	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
lagdemocrat	.4843287	.0289322	16.74	0.000	.4275475	.5411099
_cons	.2417582	.0200939	12.03	0.000	.2023227	.2811938

Bandwidth, bias and efficiency

Relaxing bandwidth requirement is about trading off bias and efficiency

. * Table 1 (continued): Examine variations in model specification
. reg score democrat , cluster(id2)

Linear regression

Number of obs = 13588

<snip>

(Std. Err. adjusted for 505 clusters in id2)

	Robust				
score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
democrat	40.76266	1.495659	27.25	0.000	37.82416 43.70115
_cons	17.5756	.8225887	21.37	0.000	15.95947 19.19172

. reg score democrat if lagdemvoteshare>.48 & lagdemvoteshare<.52, cluster(id2)

Linear regression

Number of obs = 915

<snip>

(Std. Err. adjusted for 250 clusters in id2)

	Robust				
score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
<hr/>					
democrat	47.7056	2.043658	23.34	0.000	43.68054 51.73066
_cons	18.7469	1.38312	13.55	0.000	16.02279 21.47101

Control for the running variable?

```
. * Center vote share at 0.5 to improve interpretation in interaction models
. gen      demvoteshare_c = demvoteshare - 0.5
(11 missing values generated)

. * Next, control for the (centered) running variable as a linear control variable.
. * This is the simplest RDD.
. reg      score democrat demvoteshare_c, cluster(id2)
```

Linear regression Number of obs = 13577
<snip> (Std. Err. adjusted for 505 clusters in id2)

	Robust				
score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
democrat	58.50236	1.555847	37.60	0.000	55.44561 61.5591
demvoteshare_c	-48.93761	4.441693	-11.02	0.000	-57.66412 -40.21109
_cons	11.03413	.9602657	11.49	0.000	9.147518 12.92075

```
. * Modeling the linearity such that slopes can differ above vs. below discontinuity
. xi: reg score i.democrat*demvoteshare_c, cluster(id2)
i.democrat      _Idemocrat_0-1      (naturally coded; _Idemocrat_0 omitted)
i.demov~t*demv~c _IdemXdemvo_#      (coded as above)
```

Linear regression Number of obs = 13577
<snip> (Std. Err. adjusted for 505 clusters in id2)

	Robust				
score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
_Idemocrat_1	55.43136	1.448568	38.27	0.000	52.58538 58.27734
demvoteshare_c	-5.682785	5.939863	-0.96	0.339	-17.35273 5.987156
_IdemXdemvo_1	-55.15188	8.236231	-6.70	0.000	-71.33346 -38.97031
_cons	16.81598	.9050024	18.58	0.000	15.03794 18.59403

Other bandwidths

```
* Use +/- 0.1 from cutoff (i.e., .4 to .6)
xi: reg score i.democrat*demoveshare_c if demoveshare>.40 & demoveshare<.60, cluster(id2)
i.democrat _Idemocrat_0-1      (naturally coded; _Idemocrat_0 omitted)
i.demo*t*demov_c _IdemXdemvo_#  (coded as above)
```

```
Linear regression                                         Number of obs = 4632
<snip>                                                 (Std. Err. adjusted for 428 clusters in id2)
```

Robust						
score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Idemocrat_1	47.15915	1.810413	26.05	0.000	43.60072	50.71758
demoveshare_c	9.421594	18.33431	0.51	0.608	-26.61514	45.45832
_IdemXdemvo_1	-9.127629	31.8722	-0.29	0.775	-71.77357	53.51831
_cons	17.22656	1.239458	13.90	0.000	14.79036	19.66276

```
* Use +/- 0.05
xi: reg score i.democrat*demoveshare_c if demoveshare>.45 & demoveshare<.55, cluster(id2)
i.democrat _Idemocrat_0-1      (naturally coded; _Idemocrat_0 omitted)
i.demo*t*demov_c _IdemXdemvo_#  (coded as above)
```

```
Linear regression                                         Number of obs = 2387
<snip>                                                 (Std. Err. adjusted for 367 clusters in id2)
```

Robust						
score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Idemocrat_1	46.77845	2.491464	18.78	0.000	41.87906	51.67783
demoveshare_c	54.82604	50.12314	1.09	0.275	-43.73946	153.3915
_IdemXdemvo_1	-91.1152	81.05893	-1.12	0.262	-250.5149	68.28449
_cons	18.09713	1.693516	10.69	0.000	14.76689	21.42737

```
* Use +/- 0.01
xi: reg score i.democrat*demoveshare_c if demoveshare>.49 & demoveshare<.51, cluster(id2)
i.democrat _Idemocrat_0-1      (naturally coded; _Idemocrat_0 omitted)
i.demo*t*demov_c _IdemXdemvo_#  (coded as above)
```

```
Linear regression                                         Number of obs = 441
<snip>                                                 (Std. Err. adjusted for 160 clusters in id2)
```

Robust						
score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_Idemocrat_1	45.27375	5.813497	7.79	0.000	33.79212	56.75539
demoveshare_c	55.75797	716.0422	0.08	0.938	-1358.423	1469.939
_IdemXdemvo_1	-88.56438	955.1282	-0.09	0.926	-1974.939	1797.81
_cons	20.60166	4.514799	4.56	0.000	11.68495	29.51837

No voteshare control vs. polynomials

```
. * Narrow bandwidth, no vote share control (this is comparing means close to cutoff)
. reg score democrat if demvoteshare>.49 & demvoteshare<.51, cluster(id2)
```

Linear regression

Number of obs = 441
F(1, 159) = 244.90
Prob > F = 0.0000
R-squared = 0.5525
Root MSE = 20.456

(Std. Err. adjusted for 160 clusters in id2)

score	Robust				
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
democrat	45.3931	2.900628	15.65	0.000	39.66438 51.12183
_cons	20.31812	2.026848	10.02	0.000	16.3151 24.32114

. * What about a polynomial of degree 5?

```
. gen demvoteshare2=demvoteshare^2
```

(11 missing values generated)

```
. gen demvoteshare3=demvoteshare^3
```

(11 missing values generated)

```
. gen demvoteshare4=demvoteshare^4
```

(11 missing values generated)

```
. gen demvoteshare5=demvoteshare^5
```

(11 missing values generated)

```
. reg score demvoteshare demvoteshare2 demvoteshare3 demvoteshare4 demvoteshare5 demvoteshare5 democrat, cluster(id2)
```

note: demvoteshare5 omitted because of collinearity

Linear regression

Number of obs = 13577

<snip>

(Std. Err. adjusted for 505 clusters in id2)

score	Robust				
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
demvoteshare	-19.30281	71.97002	-0.27	0.789	-160.701 122.0954
demvoteshare2	424.0805	543.3939	0.78	0.436	-643.5157 1491.677
demvoteshare3	-1746.451	1475.699	-1.18	0.237	-4645.731 1152.828
demvoteshare4	2559.367	1666.599	1.54	0.125	-714.9694 5833.704
demvoteshare5	-1248.399	664.662	-1.88	0.061	-2554.249 57.45029
demvoteshare5	0	(omitted)			
democrat	49.13039	1.764906	27.84	0.000	45.66291 52.59787
_cons	16.36452	3.07522	5.32	0.000	10.32269 22.40635

Polynomial with interactions

```
. * Center the running variable and use polynomials and interactions to model
* the nonlinearities below and above discontinuity
.gen x.c=demvoteeshare-0.5
(11 missing values generated)

.gen x2.c=x.c^2
(11 missing values generated)

.gen x3.c=x.c^3
(11 missing values generated)

.gen x4.c=x.c^4
(11 missing values generated)

.gen x5.c=x.c^5
(11 missing values generated)

reg score i.democrat##(c.x.c c.x2.c c.x3.c c.x4.c c.x5.c)
```

Source	SS	df	MS	Number of obs	=	13577
				F(11, 13565)	=	1058.32
Model	6677033.81	11	607003.073	Prob > F	=	0.0000
Residual	7780253.69	13565	573.553534	R-squared	=	0.4618
				Adj R-squared	=	0.4614
Total	14457287.5	13576	1064.91511	Root MSE	=	23.949

score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
1.democrat	47.73325	2.042906	23.37	0.000	43.72887 51.73763
x.c	-28.79765	74.09167	-0.39	0.698	-174.0276 116.4323
x2.c	-1138.232	1144.497	-0.99	0.320	-3381.605 1105.141
x3.c	-10681.29	7137.315	-1.50	0.135	-24671.42 3308.839
x4.c	-33490.23	18844.92	-1.78	0.076	-70428.88 3448.424
x5.c	-32873.77	17212.08	-1.91	0.056	-66611.83 864.302
democrat#c.x.c					
1	-5.793828	97.79088	-0.06	0.953	-197.4775 185.8899
democrat#c.x2.c					
1	1768.225	1433.252	1.23	0.217	-1041.149 4577.599
democrat#c.x3.c					
1	6279.346	8553.95	0.73	0.463	-10487.58 23046.28
democrat#c.x4.c					
1	47111.44	21834.51	2.16	0.031	4312.767 89910.11
democrat#c.x5.c					
1	17786.85	19488.24	0.91	0.361	-20412.81 55986.51
_cons	17.05847	1.469785	11.61	0.000	14.17749 19.93946

More polynomial and bandwidth regressions

```
. * We could limit the regressions to a window using a 2nd degree polynomial
. * interacted with the treatment variable
. reg score i.democrat##(c.x_c c.x2_c) if (demvoteshare>0.4 & demvoteshare<0.6)
```

Source	SS	df	MS	Number of obs	=	4632
Model	2622762.02	5	524552.404	F(5, 4626)	=	1153.29
Residual	2104043.2	4626	454.829918	Prob > F	=	0.0000
Total	4726805.22	4631	1020.6878	R-squared	=	0.5549
				Adj R-squared	=	0.5544
				Root MSE	=	21.327

score	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
1.democrat	45.9283	1.892566	24.27	0.000	42.21797	49.63863
x_c	38.63988	60.77525	0.64	0.525	-80.5086	157.7884
x2_c	295.1723	594.3159	0.50	0.619	-869.9704	1460.315
democrat#c.x_c						
1	6.507415	88.51418	0.07	0.941	-167.0226	180.0374
democrat#c.x2_c						
1	-744.0247	862.0435	-0.86	0.388	-2434.041	945.9916
_cons	17.71198	1.310861	13.51	0.000	15.14207	20.28189

Nonparametric estimation

- Hahn, Todd and Van der Klaauw (2001) clarified assumptions about RDD (i.e., continuity in conditional expectation regression functions)
- Also framed estimation as a non-parametric problem and emphasized using local polynomial regressions
- Nonparametric methods mean a lot of different things to different people in statistics.
- In RDD context, the idea is to estimate a model that doesn't assume a functional form for the relationship between Y (outcome variable) and X (running variable)
- That model would be something general like

$$Y = f(X) + \varepsilon$$

- A very basic method is to calculate $E[Y]$ for each bin on X – like a histogram.
- STATA has an option to do this called `cmogram` and it has a lot of useful options. We can recreate Figures I, IIA and IIB using it

Graphical estimate of equation 205

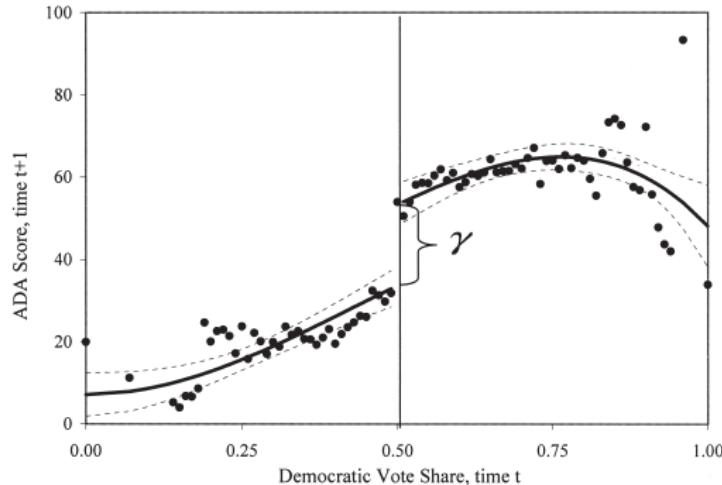


FIGURE I

Total Effect of Initial Win on Future ADA Scores: γ

This figure plots ADA scores after the election at time $t + 1$ against the Democrat vote share, time t . Each circle is the average ADA score within 0.01 intervals of the Democrat vote share. Solid lines are fitted values from fourth-order polynomial regressions on either side of the discontinuity. Dotted lines are pointwise 95 percent confidence intervals. The discontinuity gap estimates

$$\gamma = \underbrace{\pi_0(P_{t+1}^{*D} - P_{t+1}^{*R})}_{\text{"Affect"}} + \underbrace{\pi_1(P_{t+1}^{*D} - P_{t+1}^{*R})}_{\text{"Elect"}}$$

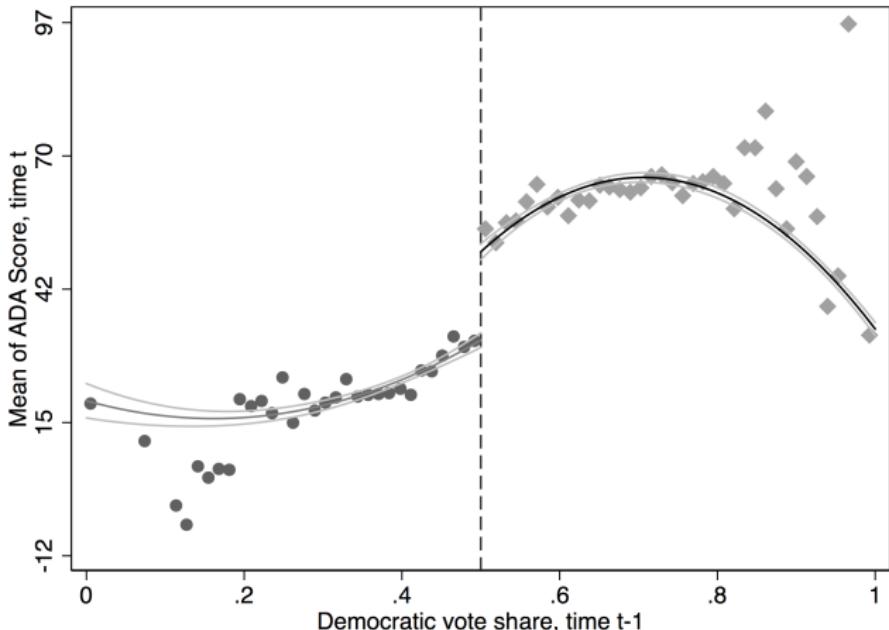
Figure: Lee, Moretti, and Butler 2004, Figure I. $\gamma \approx 20$

```
. cmogram score lagdemvoteshare , cut(0.5) scatter line(0.5) qfitci
```

Plotting mean of score, conditional on lagdemvoteshare.

n = 13577

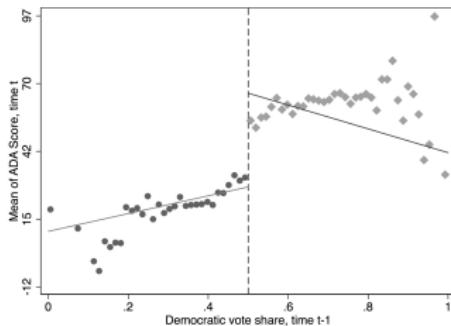
```
Bin #1: [0,.0135101361854656] (n = 181) (mean = 18.90983436977007)
Bin #2: (.0135101361854656,.0270202723709312] (n = 0) (mean = .)
Bin #3: (.0270202723709312,.040530408563968] (n = 0) (mean = .)
Bin #4: (.040530408563968,.0540405447418624] (n = 0) (mean = .)
Bin #5: (.0540405447418624,.067550680927328] (n = 0) (mean = .)
Bin #6: (.067550680927328,.0810608171127936] (n = 2) (mean = 11.2350001335144)
<snip>
Bin #36: (.9605263157894751,.9736842105263173] (n = 2) (mean = 96.58000183105469)
Bin #37: (.9736842105263173,.9868421052631594] (n = 0) (mean = .)
Bin #38: (.9868421052631594,1] (n = 1646) (mean = 33.014483642933)
```



Linear and lowess fits

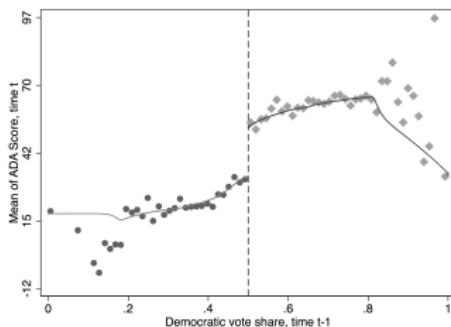
```
. cmogram score lagdemvoteshare , cut(0.5) scatter line(0.5) lfit
```

Plotting mean of score, conditional on lagdemvoteshare.



```
. cmogram score lagdemvoteshare , cut(0.5) scatter line(0.5) lowess
```

Plotting mean of score, conditional on lagdemvoteshare.



Graphical estimate of equation 206

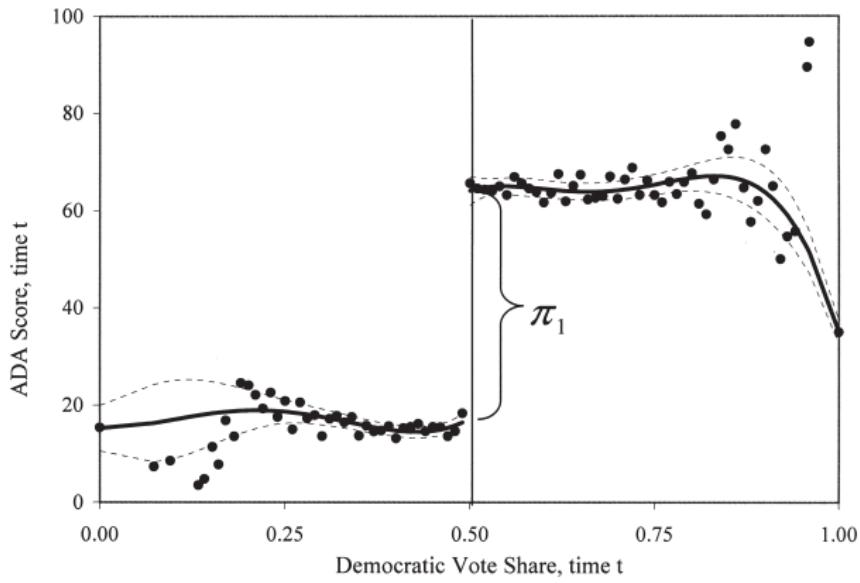


FIGURE IIa
Effect of Party Affiliation: π_1

Figure: Lee, Moretti, and Butler 2004, Figure IIa. $\pi_1 \approx 45$

* Lowess
cmogram score demvoteshare, cut(.5) scatter line(.5) lowess

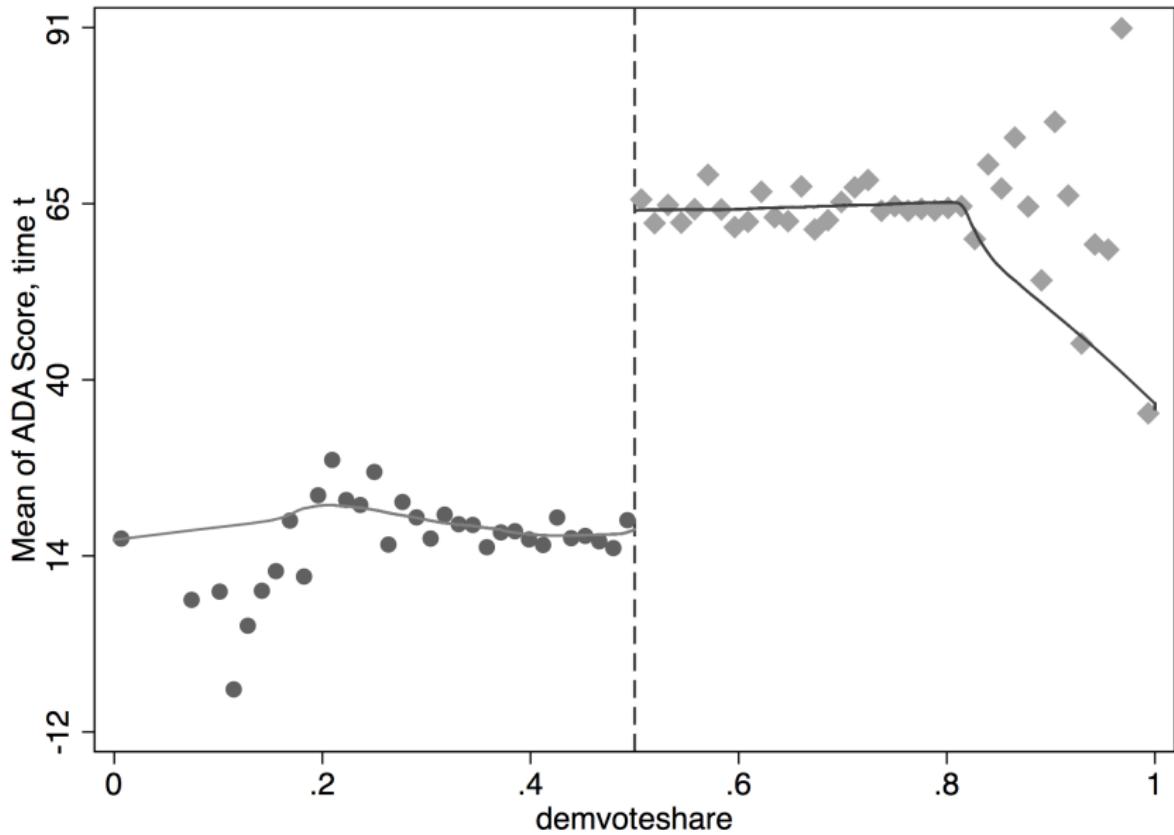


Figure: Cubic feet with confidence intervals

Graphical estimate of equation 207

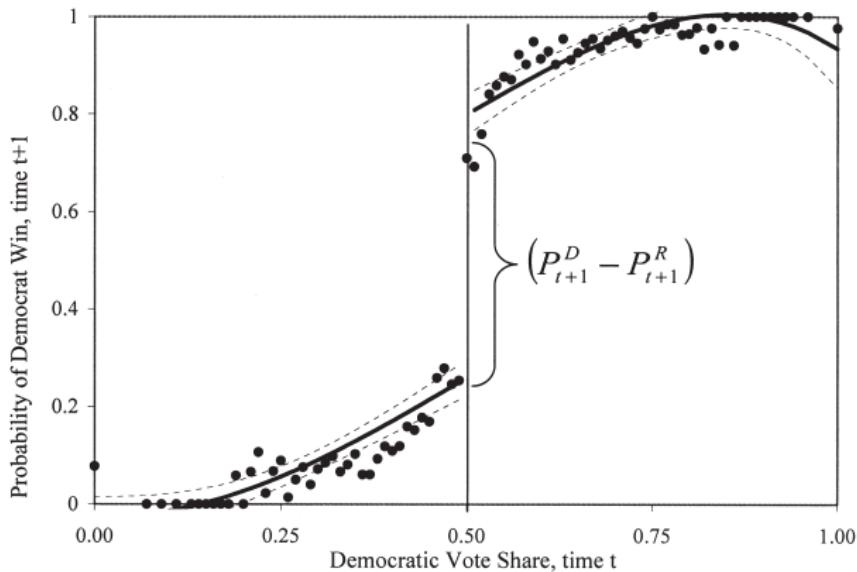


FIGURE IIb
Effect of Initial Win on Winning Next Election: $(P_{t+1}^D - P_{t+1}^R)$

Figure: Lee, Moretti, and Butler 2004, Figure IIb. $(P_{t+1}^D - P_{t+1}^R) \approx 0.50$

Kernel weighted local polynomial regression

- Hahn, Todd and Van der Klaauw (2001) showed that the one-sided kernel estimation (such as lowess) may have poor properties because the point of interest is at a boundary (i.e., the discontinuity), called the “boundary problem”
- They proposed to use “local linear nonparametric regressions” instead
- STATA’s poly estimates kernel-weighted local polynomial regressions. Think of it as a weighted regression restricted to a window like we’ve been doing (hence “local”) where the kernel provides the weights
- A rectangular kernel would give the same result as $E[Y]$ at a given bin on X . The triangular kernel gives more importance to observations close to the center.
- This method will be sensitive to how large the bandwidth (window) you choose

```

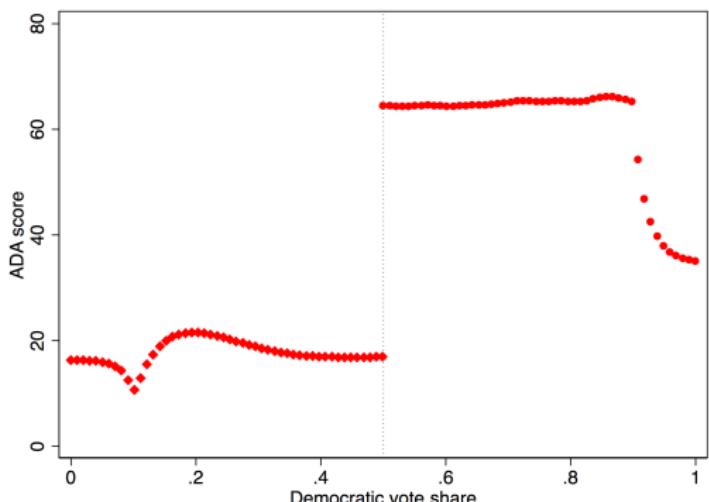
. * Note kernel-weighted local polynomial regression is a smoothing method.
. lpoly score demvoteshare if democrat == 0, nograph kernel(triangle) gen(x0 sdem0) ///
>           bwidth(0.1)

. lpoly score demvoteshare if democrat == 1, nograph kernel(triangle) gen(x1 sdem1) ///
>           bwidth(0.1)

. scatter sdm1 x1, color(red) msize(small) || scatter sdm0 x0, msize(small) color(red) ///
>           xline(0.5,lstyle(dot)) legend(off) xttitle("Democratic vote share") yttitle("ADA score")

. * Next, let's get the treatment effect at the cutoff where demvoteshare=0.5
. capture drop sdm0 sdm1
. gen forat=0.5 in 1
. lpoly score demvoteshare if democrat==0, nograph kernel(triangle) gen(sdm0) at(forat) bwidth(0.1)
. lpoly score demvoteshare if democrat==1, nograph kernel(triangle) gen(sdm1) at(forat) bwidth(0.1)
. gen late=sdm1 - sdm0
. list sdm1 sdm0 late in 1/1
+-----+
| sdm1     sdm0     late |
|-----|
1. | 64.395204  16.908821  47.48639 |
+-----+

```

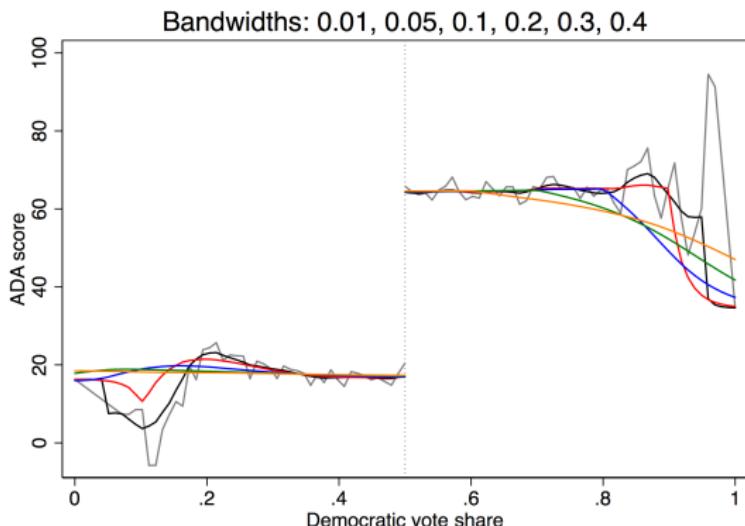


```

. * What happens when we change the bandwidth? Use 0.01, 0.05, 0.2, 0.3, 0.4
. capture drop smoothdem0* smoothdem1* x0* x1*
. local co 0
. foreach i in 0.01 0.05 0.1 0.20 0.30 0.40
2.    local co = `co' +1
3.    lpoly score demvoteshare if democrat == 0, nograph kernel(triangle) gen(x0`co' smoothdem0`co') ///
>        bwidth(`i')
4.    lpoly score demvoteshare if democrat == 1, nograph kernel(triangle) gen(x1`co' smoothdem1`co') ///
>        bwidth(`i')
5.

. line smoothdem01 x01, ms(1) color(gray) sort || line smoothdem11 x11, sort color(gray) || ///
>    line smoothdem02 x02, color(black) sort || line smoothdem12 x12, sort color(black) || ///
>    line smoothdem03 x03, color(red) sort || line smoothdem13 x13, sort color(red) || ///
>    line smoothdem04 x04, color(blue) sort || line smoothdem14 x14, sort color(blue) || ///
>    line smoothdem05 x05, color(green) sort || line smoothdem15 x15, sort color(green) || ///
>    line smoothdem06 x06, color(orange) sort || line smoothdem16 x16, sort color(orange) ||
>    xline(0.5,lstyle(dot)) legend(off) xtitle("Democratic vote share") ytitle("ADA score") ||
>    title("Bandwidths: 0.01, 0.05, 0.1, 0.2, 0.3, 0.4")

```



- Several methods for choosing the optimal bandwidth (window), but it's always a trade off between bias and variance
- In practical applications, you want to check for balance around that window
- Standard error of the treatment effects can be bootstrapped but there are also other alternatives
- You could add other variables to nonparametric methods.
- Calonico, Cattaneo and Titiunik (2013b) propose local-polynomial regression discontinuity estimators with robust confidence intervals
- STATA ado package and R package are both called `rdrobust`

```
. rdrobust score demvoteshare, c(0.5) all bwselect(IK)
Preparing data.
Computing Bandwidth Selectors.
Computing Variance-Covariance Matrix.
Computing RD Estimates.
Estimation Completed.
```

Sharp RD Estimates using Local Polynomial Regression.

Cutoff c = .5 Left of c Right of c			Number of obs = 13577
-----+-----			
Number of obs	3535	3318	NN Matches = 3
Order Loc. Poly. (p)	1	1	BW Type = IK
Order Bias (q)	2	2	Kernel Type = Triangular
BW Loc. Poly. (h)	0.152	0.152	
BW Bias (b)	0.191	0.191	
rho (h/b)	0.795	0.795	

Outcome: score. Running Variable: demvoteshare.

Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Conventional	47.171	.98058	48.1046	0.000	45.2488 49.0926
Robust	-	-	36.4839	0.000	44.0965 49.1034

All Estimates.

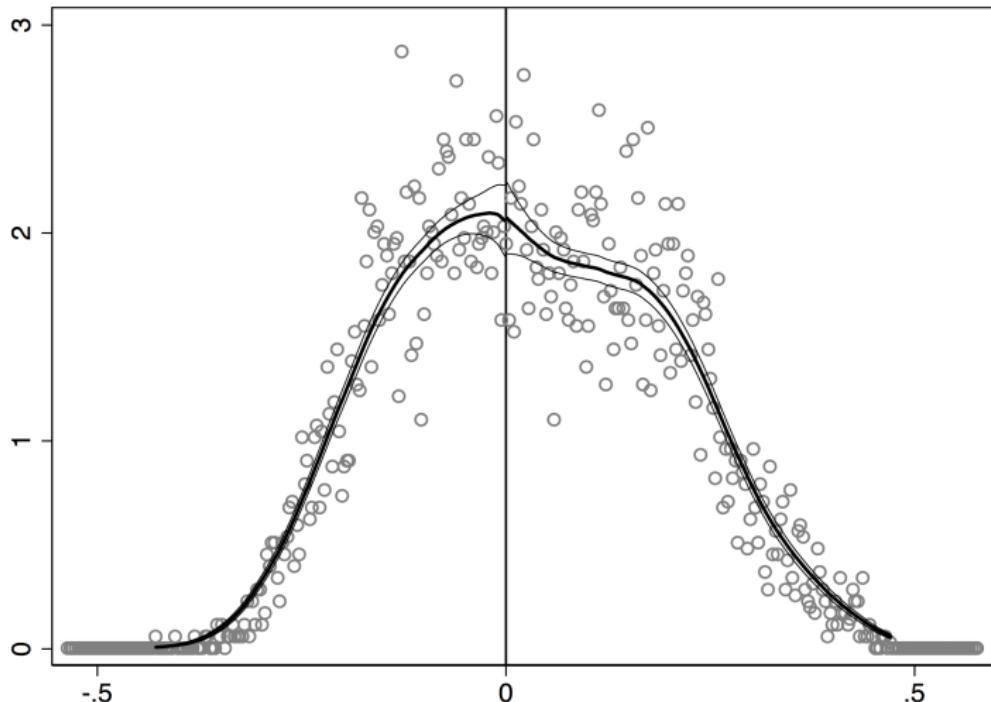
Method	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
Conventional	47.171	.98058	48.1046	0.000	45.2488 49.0926
Bias-Corrected	46.6	.98058	47.5226	0.000	44.678 48.5219
Robust	46.6	1.2773	36.4839	0.000	44.0965 49.1034

McCrary Density Test

```
• McCrary (2008) density test to check for manipulation of the running variable (DCdensity)
.DCdensity demovoteshare_c if (demovoteshare_c>=0.5 & demovoteshare_c<0.5), breakpoint(0) generate(Xj Yj r0 fhat se_fhat)
Using default bin size calculation, bin size = .003047982
Using default bandwidth calculation, bandwidth = .104944836

Discontinuity estimate (log difference in height): .011195629
(.061618519)

Performing LLR smoothing.
296 iterations will be performed
```



“Pioneers take the arrows in the back”
– Jim West during a faculty meeting

Lee (2008) is important for the econometric theory, but this particular application turned out to be *wrong*:

“Contrary to the assumptions of RD, we show that bare winners and bare losers in US House elections (1942-2008) differ markedly on pretreatment covariates. Bare winners possess large ex ante financial, experience, and incumbency advantages over their opponents and are usually the candidates predicted to win by Congressional Quarterly—’s pre-election ratings. Covariate imbalance actually worsens in the closest House election. National partisan tides help explain these patterns. We present evidence suggesting that sorting in close House elections is due mainly to activities on or before Election Day rather than post election recounts or other manipulation. The sorting is so strong that it is impossible to achieve covariate balance between matched treated and control observations, making covariate adjustment a dubious enterprise. Although RD is problematic for postwar House elections, this example does highlight the design’s advantages over alternatives: RD’s assumptions are clear and weaker than model-based alternatives, and their implications are empirically testable.” (Caughey and Sekhon, 2011).

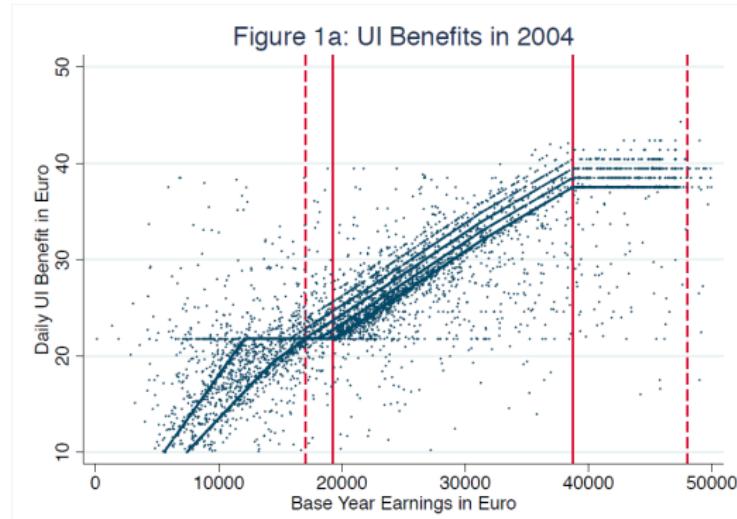
Regression Kink Design

- Card, Lee, Pei and Weber (2012) introduces a variant of the RDD which they call regression kink design (RKD)
- They essentially use a kink in some policy rule to identify the causal effect of the policy
- Instead of a jump in the outcome you now expect a jump in the first derivative

Unemployment benefits in Austria

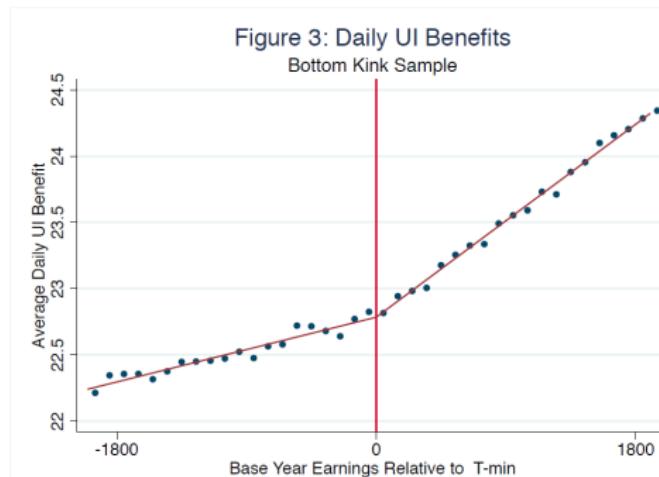
- They apply their design to answer the question whether the level of unemployment benefits affects the length of unemployment in Austria
- Unemployment benefits are based on income in a base period
- The benefit formula for unemployment exhibits 2 kinks
 - There is a minimum benefit level that isn't binding for people with low earnings
 - Then benefits are 55% of the earnings in the base period
 - There is a maximum benefit level that is adjusted every year
- People with dependents get small supplements (which is the reason why one can distinguish five "solid" lines in the following graph)
- Not everyone receives benefits that correspond one to one to the formula because of mistakes in the administrative data

Base Year Earnings and Unemployment Benefits



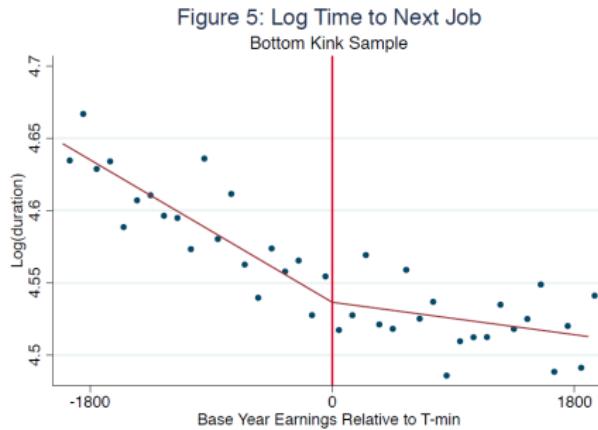
- The graph shows unemployment benefits (vertical axis) as a function of pre-unemployment earnings (horizontal axis)

Base Year Earnings and Benefits for Single Individuals



- Bin-size: 100 euros
- For single individuals UI benefits are flat below the cutoff. The relationship is still upward sloping because of family benefits.

Time to Next Job for Single Individuals



- People with higher base earnings have less trouble finding a job (negative slope)
- There is a kink: the relationship becomes shallower once benefits increase more.

Selection on unobservables

- Recall the identifying assumption when estimating causal effects using a conditioning strategy

$$\text{Independence: } (Y^0, Y^1) \perp\!\!\!\perp D$$

Use methods like propensity score matching and condition on $p(x)$, or linear regression

- What if treatment assignment to units is based on some unobserved variable, u , which is correlated with the outcome
 - Conditioning strategies are invalid
 - Selection on unobservable methods, though, may work
- Natural experiments may be useful

Natural experiments

- Natural experiments are neither an estimator or an experiment (natural or otherwise). Rather, it is simply an event that occurs *naturally* which causes exogenous variation in some treatment variable of interest
- An attempt to find in the world some rare circumstance such that a consequential treatment was handed to some people and denied to others haphazardly
- “The word *natural* has various connotations, but a *natural experiment* is a *wild experiment* not a *wholesome experiment*; nature in the way that a tiger is natural, not in the way that oatmeal is natural.” (Rosenbaum 2005)
- Haphazard variation in treatment is *not* like the variation induced by a randomized experiment

Natural experiments

- Naive matching – “people who look comparable are comparable”. **False**.
 - Matching can create matched samples of units that look similar on some x
 - Matching cannot create a matched sample who are similar on u (unobservables), though
- Definition of naive: someone who believes something because it's convenient to believe it.

Broad Street Pump

- John Snow is believed to be the bastard son of Edward Stark but is actually a Targaryan
- But before that John Snow, there was another John Snow – a practicing anesthesiologist in 19th century London
- This Snow is considered the father of epidemiology
- He is famous for providing convincing evidence that cholera was a waterborne disease
- He lived from 1813-1858

19th century science

- Microscopes were around but had horrible resolution
- Most human pathogens couldn't be seen
- Isolating these microorganisms wouldn't occur for half a century
- Two views of what caused disease:
 - Majority view - "Miasmas". Minute, inanimate poison particles in the air
 - Minority view - "infection theory".

Cholera background

- Cholera arrives in the early 1800s and exhibits “epidemic waves”.
- Cholera attacked victims suddenly, was usually fatal, and symptoms included vomiting and acute diarrhea
- Snow observed the clinical course of the disease and made the following conjecture
 - The active agent was a living organism that entered the body, got into the alimentary canal with food or drink, multiplied in the body, and generated some poison that caused the body to expel water
 - The organism passed out of the body with these evacuations, entered the water supply and infected new victims
 - The process repeated itself, growing rapidly through the common water supply, causing an epidemic

History

- There were three main epidemics in London
 - 1831-1832
 - 1848-1849 (>15,000 deaths)
 - 1853-1854 (>30,000 deaths)

Background

- Snow is an early advocate for the infection theory – cholera is being spread person-to-person through some unknown mechanism
- His earlier evidence was based on years of observations including
 - Cholera transmission tended to follow human commerce
 - A sailor on a ship from a cholera-free country who arrived at a cholera-stricken port would only get sick after landing or taking on supplies
 - Cholera hit the poor communities the worst, who also lived in the most crowded housing with the worst hygiene
- Facts were difficult to reconcile with the miasma theory, but are consistent with infection theory

More evidence

- Snow identifies Patient Zero: the first case of an early epidemic
 - “a seaman named John Harnold, who had arrived by the *Elbe* steamer from Hamburg, where the disease was prevailing”
- The second case? John Harnold's roommate

More evidence

- Snow studied two apartment buildings. The first was heavily hit with cholera but the second wasn't.
 - He found the water supply in the first building was contaminated by runoff from privies but the water supply in the second was cleaner
- Earlier water supply studies
 - In the London of the 1800s, there were many different water companies serving different areas of the city
 - Some were served by more than one company
 - Several took their water from the Thames, which was heavily polluted by sewage
 - The service areas of such companies had much higher rates of cholera
 - The Chelsea water company was an exception, but it had an exceptionally good filtration system

Broad Street Water Pump

- 1849: Lambeth water company moves the intake point upstream along the Thames, above the main sewage discharge point (i.e., purer water)
- 1849: Southwark and Vauxhall water company left their intake point downstream from where the sewage discharged (gross) (i.e., infected water)
- Comparisons of data on cholera deaths from the 1853-1854 year showed the epidemic hit the Southwark and Vauxhall services areas harder, and largely spared the Lambeth areas

Snow's Table IX

	No. houses	Cholera deaths	Deaths / 10,000 houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

Shoe leather research

"As a piece of statistical technology, [Snow's Table IX] is by no means remarkable. But the story it tells is very persuasive. The force of the argument results from the clarity of the prior reasoning, the bringing together of many different lines of evidence, and the amount of shoe leather Snow was willing to use to get the data."

"Snow did some brilliant detective work on nonexperimental data. What is impressive is not the statistical technique but the handling of the scientific issues. He made steady progress from shrewd observation through case studies to analyze ecological data. In the end, he found and analyzed a natural experiment."

Introduction to instrumental variables

- Pearl argues that there are three ways to estimate causal effects: the backdoor criterion ("selection on observables"), instrumental variables, and the front door criterion.
- Now we move into the instrumental variables research design
- Illustrious history – discovered by Philip Wright and published as an appendix in his 1928 book
- One of the most powerful methods for identifying causal effects in the social sciences

- Sometimes, the researcher *can* find the flag. That is, the researcher knows of a variable (Z) that actually *is* randomly assigned and that affects fertility decisions. Such a variable is called an “instrument”.
- Example: Angrist and Evans (1998), “Children and their parents’ labor supply” *American Economic Review*,
 - Z is a dummy variable indicating whether the first two children born were of the same gender
 - Many parents have a preference for having at least one child of each gender
 - Consider a couple whose first two kids were both boys; they will often have a third, hoping to have a girl
 - Consider a couple whose first two kids were girls; they will often have a third, hoping for a boy
 - Consider a couple with one boy and one girl; they will often not have a third kid
 - The gender of your kids is arguably randomly assigned (maybe not exactly, but close enough)

- “No causation without manipulation” (Holland, 1985). If you want to use IV, then ask:

What moves around the covariate of interest that might be plausibly viewed as random?

- “Focusing on the selection process”
 - How was the covariate of interest selected?
 - Do researchers select randomly, as in drawing balls from an urn?
 - Or do subjects choose the covariate partially based on random factors and partially based on non-random factors?
 - If the random factors underlying the decision can be *observed by the researcher*, then you can use IV to estimate the effect of the covariate of interest on the outcome
- Angrist and Evans (1988) example:
 - Families with at least two kids are the subject population
 - The covariate of interest is the number of kids (numkids).
 - There are several outcomes of interest measuring labor market decisions of the couple, such as whether the mother worked for pay in the last year (workforpay).
- Once you have identified such a variable, begin to think about what data sets might have information on an outcome of interest, the covariate of interest, and the instrument you have put your finger on.

- In a pinch, you can even get by with two different data sets, one of which has information on the outcome and the instrument, and the other of which has information on the covariate of interest and the instrument.
- This is known as “Two sample IV” because there are two *samples* involved, rather than the traditional one sample.
- Once we define what IV is measuring carefully, you will see why this works.

- Instrumental variables strategies formalize *untainted inference*, which is the inference drawn by an intelligent layperson with no particular training or background in statistics.
- Example of what I mean by “untainted inference”:
 - The researchers tell a layperson that the gender of a woman’s first two children is predictive of her labor force attachment. In particular, women whose first two children are of the same gender work substantially less than women whose first two children are of different genders
 - On its face, this is a puzzling fact – without further information, it is hard to see why the gender of your children would be so predictive of labor market participation
 - The researchers additionally point out that women whose first two children are of the same gender are more likely to have additional children than women whose first two children are of different genders
 - The layperson then wonders whether the labor market differences are due *solely* to the differences in the number of kids the woman has

Traditional and Contemporary IV Pedagogy

We want to learn the IV framework in two iterations:

- ① Constant treatment effects (i.e., β is constant across all individual units)
 - Constant treatment effects is the traditional econometric pedagogy when first learning instrumental variables, and is not based explicitly on the potential outcomes model or notation
 - Constant treatment effects is identical to assuming that $ATE = ATT = ATU$ because constant treatment effects assumes $\beta_i = \beta_{-i} = \beta$ for all units
- ② Heterogeneous treatment effects (i.e., β_i varies across individual units)
 - This is the “modern IV pedagogy”, and you may not have learned it in econometrics if only because potential outcomes is not ordinarily the basis of most first year econometrics sequences
 - Heterogeneous treatment effects means that the $ATE \neq ATT \neq ATU$ because β_i differs across the population
 - This is equivalent to assuming the coefficient, β_i , is a random variable that varies across the population
 - Heterogenous treatment effects is based on work by Angrist, Imbens and Rubin (1996) and Imbens and Angrist (1994) which introduced the “local average treatment effect” (LATE) concept

When should you think of using instrumental variables?

- Instrumental variables methods are typically used to address the following kinds of problems encountered in OLS regressions:
 - 1 Omitted variable bias
 - 2 Measurement error
 - 3 Simultaneity bias, or “reverse causality”

Omitted Variable Bias (Angrist and Pischke, 2009)

- Labor economists have been studying the returns to schooling a long time – typically some version of a “Mincer regression”:

$$Y_i = \alpha + \rho S_i + \gamma A_i + \nu_i$$

Y_i = log of earnings

S_i = schooling measured in years

A_i = individual ability

- Typically the econometrician cannot observe A_i ; for instance, the CPS tells us nothing about adult respondents' family background, intelligence, or motivation.
- What are the consequences of leaving ability out of the regression? Suppose you estimated this short regression instead:

$$Y_i = \alpha + \rho S_i + \eta_i$$

where $\eta_i = \gamma A_i + \nu_i$; α , ρ , and γ are population regression coefficients; S_i is correlated with η_i through A_i only; and ν_i is a regression residual uncorrelated with all regressors by definition.

Derivation of Ability Bias

- Suppressing the i subscripts, the OLS estimator for ρ is:

$$\hat{\rho} = \frac{\text{Cov}(Y, S)}{\text{Var}(S)} = \frac{E[YS] - E[Y]E[S]}{\text{Var}(S)}$$

- Plugging in the true model for Y , we get:

$$\begin{aligned}\hat{\rho} &= \frac{\text{Cov}[(\alpha + \rho S + \gamma A + \nu), S]}{\text{Var}(S)} \\ &= \frac{E[(\alpha S + S^2 \rho + \gamma S A + \nu S)] - E(S)E[\alpha + \rho S + \gamma A + \nu]}{\text{Var}(S)} \\ &= \frac{\rho E(S^2) - \rho E(S)^2 + \gamma E(AS) - \gamma E(S)E(A) + E(\nu S) - E(S)E(\nu)}{\text{Var}(S)} \\ &= \rho + \gamma \frac{\text{Cov}(AS)}{\text{Var}(S)}\end{aligned}$$

- If $\gamma > 0$ and $\text{Cov}(A, S) > 0$ the coefficient on schooling in the shortened regression (without controlling for A) would be upward biased

How can IV be used to obtain unbiased estimates?

- Suppose there exists a variable, Z_i , that is correlated with S_i .
- We can estimate ρ with this variable, Z :

$$\begin{aligned}
 \text{Cov}(Y, Z) &= \text{Cov}(\alpha + \rho S + \gamma A + \nu, Z) \\
 &= E[(\alpha + \rho S + \gamma A + \nu), Z] - E[\alpha + \rho S + \gamma A + \nu]E[Z] \\
 &= \{\alpha E(Z) - \alpha E(Z)\} + \rho \{E(SZ) - E(S)E(Z)\} \\
 &\quad + \gamma \{E(AZ) - E(A)E(Z)\} + E(\nu Z) - E(\nu)E(Z) \\
 \text{Cov}(Y, Z) &= \rho \text{Cov}(S, Z) + \gamma \text{Cov}(A, Z) + \text{Cov}(\nu, Z)
 \end{aligned}$$

- What conditions must hold for a valid instrumental variable?
 - $\text{Cov}(S, Z) \neq 0$ – “first stage” exists. S and Z are correlated
 - $\text{Cov}(A, Z) = \text{Cov}(\nu, Z) = 0$ – “exclusion restriction”. Z is orthogonal to the factors in η , such as unobserved ability or the structural disturbance term, η
- Assuming the first stage exists and that the exclusion restriction holds, then we can estimate ρ with ρ_{IV} :

$$\rho_{IV} = \frac{\text{Cov}(Y, Z)}{\text{Cov}(S, Z)} = \rho$$

IV is Consistent if IV Assumptions are Satisfied

- The IV estimator is consistent if the IV assumptions are satisfied. Substitute true model for Y :

$$\begin{aligned}
 \rho_{IV} &= \frac{\text{Cov}([\alpha + \rho S + \gamma A + \nu], Z)}{\text{Cov}(S, Z)} \\
 &= \rho \frac{\text{Cov}([S], Z)}{\text{Cov}(S, Z)} + \gamma \frac{\text{Cov}([A], Z)}{\text{Cov}(S, Z)} + \frac{\text{Cov}([\nu], Z)}{\text{Cov}(S, Z)} \\
 &= \rho + \gamma \frac{\text{Cov}(\eta, Z)}{\text{Cov}(S, Z)}
 \end{aligned}$$

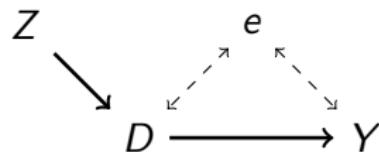
- Taking the plim:

$$\text{plim } \hat{\rho}_{IV} = \rho$$

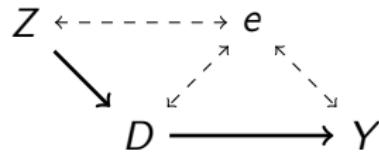
because $\text{Cov}([A], Z) = 0$ and $\text{Cov}([\nu], Z) = 0$ due to the exclusion restriction, and $\text{Cov}(S, Z) \neq 0$ (due to the first stage)

- But, if Z is *not* independent of η (either correlated with A or ν), *and* if the correlation between S and Z is “weak”, then the second term blows up. We will return to this later when we discuss the problem of “weak instruments”.

In which DAG is Z a valid instrument for D?



(a)



(b)

Reviewing some of the IV Jargon

- Causal model. Sometimes called the structural model:

$$Y_i = \alpha + \rho S_i + \eta_i$$

- First-stage regression. Gets the name because of two-stage least squares:

$$S_i = \alpha + \rho Z_i + \zeta_i$$

- Second-stage regression. Notice the fitted values, \hat{S} :

$$Y_i = \alpha + \rho \hat{S}_i + \nu_i$$

- Reduced form. Notice this is just a regression of Y onto the instrument:

$$Y_i = \alpha + \pi Z_i + \varepsilon_i$$

Two-stages least squares

- Suppose you have a sample of data on Y , X , and Z . For each observation i we assume the data are generated according to

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \varepsilon_i; \\ X_i &= \gamma + \delta Z_i + \nu_i \end{aligned}$$

where $\text{Cov}(Z, \varepsilon) = 0$ and $\delta \neq 0$.

- Plug in covariance, and using the result that $\sum_{i=1}^n (x_i - \bar{x}) = 0$, write out the IV estimator:

$$\widehat{\beta}_{IV} = \frac{\text{Cov}(Z, Y)}{\text{Cov}(Z, X)} = \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})(X_i - \bar{X})} = \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})Y_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})X_i}$$

- Substitute the causal model definition of Y to get:

$$\begin{aligned} \widehat{\beta}_{IV} &= \frac{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})\{\alpha + \beta X_i + \varepsilon_i\}}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})X_i} \\ &= \beta + \frac{\frac{1}{n} (Z_i - \bar{Z})\varepsilon_i}{\frac{1}{n} \sum_{i=1}^n (Z_i - \bar{Z})X_i} \\ &= \beta + \text{"small if } n \text{ is large"} \end{aligned}$$

Two-stages least squares

- Note β_{IV} is ratio of “reduced form” (π) to “first stage” coefficient (δ):

$$\hat{\beta}_{IV} = \frac{Cov(Z, Y)}{Cov(Z, X)} = \frac{\frac{Cov(Z, Y)}{Var(Z)}}{\frac{Cov(Z, X)}{Var(Z)}} = \frac{\hat{\pi}}{\hat{\delta}}$$

- Rewrite $\hat{\delta}$ as

$$\hat{\delta} = \frac{Cov(Z, X)}{Var(Z)} \Leftrightarrow Cov(Z, X) = \hat{\delta} Var(Z) \quad (100)$$

- Then rewrite β_{IV}

$$\begin{aligned} \widehat{\beta}_{IV} &= \frac{Cov(Z, Y)}{Cov(Z, X)} = \frac{\hat{\delta} Cov(Z, Y)}{\hat{\delta} Cov(Z, X)} = \frac{\hat{\delta} Cov(Z, Y)}{\hat{\delta}^2 Var(Z)} \\ &= \frac{Cov(\hat{\delta} Z, Y)}{Var(\hat{\delta} Z)} \end{aligned} \quad (101)$$

Two-stage least squares

- Recall $X = \gamma + \delta Z + \nu$; $\widehat{\beta}_{IV} = \frac{\text{Cov}(\widehat{\delta}Z, Y)}{\text{Var}(\widehat{\delta}Z)}$ and let $\widehat{X} = \widehat{\gamma} + \widehat{\delta}Z$.
- Then the two-stage least squares (2SLS) estimator is

$$\widehat{\beta}_{IV} = \frac{\text{Cov}(\widehat{\delta}Z, Y)}{\text{Var}(\widehat{\delta}Z)} = \frac{\text{Cov}(\widehat{X}, Y)}{\text{Var}(\widehat{X})}$$

Proof.

We will show that $\widehat{\delta}\text{Cov}(Y, Z) = \text{Cov}(\widehat{X}, Y)$. I will leave it to you to show that $\text{Var}(\widehat{\delta}Z) = \text{Var}(\widehat{X})$

$$\begin{aligned}\text{Cov}(\widehat{X}, Y) &= E[\widehat{X}Y] - E[\widehat{X}]E[Y] \\ &= E(Y[\widehat{\gamma} + \widehat{\delta}Z]) - E(Y)E(\widehat{\gamma} + \widehat{\delta}Z) \\ &= \widehat{\gamma}E(Y) + \widehat{\delta}E(YZ) - \widehat{\gamma}E(Y) - \widehat{\delta}E(Y)E(Z) \\ &= \widehat{\delta}[E(YZ) - E(Y)E(Z)] \\ \text{Cov}(\widehat{X}, Y) &= \widehat{\delta}\text{Cov}(Y, Z)\end{aligned}$$



- The 2SLS estimator replaces X with the fitted values of X (i.e., \widehat{X}) from the first stage regression of X onto Z and all other covariates.

Intuition of 2SLS

- I've said that learning about instrumental variables through the “intuition” of two-stage least squares is valuable, but what do I mean exactly?
- 2SLS emphasizes the use of the fitted values of the endogenous regressor – what has that transformation done?
- By using the fitted values of the endogenous regressor from the first stage regression, our regression now uses *only* the exogenous variation in the regressor due to the instrumental variable itself
- We recover exogenous variation in other words
- ... but think about it – that variation was there before, but was just a subset of all the variation in the regressor
- Instrumental variables therefore reduces the variation in the data, but that variation which is left is *exogenous*

Implementing 2SLS in STATA

- In a sample of data, you could get the reduced form and first stage coefficients manually by the following two regression commands in STATA:

```
. reg Y Z
```

```
. reg X Z
```

- While it is always a good idea to run these two regressions, don't compute your IV estimate this way
 - Example: It is often the case that a pattern of missing data will differ between Y and X ; in such a case, the usual procedure of "casewise deletion" is to keep the subsample with non-missing data on Y , X , and Z .
 - But the reduced form and first stage regressions would be estimated off of different sub-samples if you used the two step method above
 - The standard errors from the second stage regression are also wrong
- Best practice is to use your built-in procedure (which also gives standard errors):

```
. ivregress Y (X=Z)
```

Implementing 2SLS in STATA

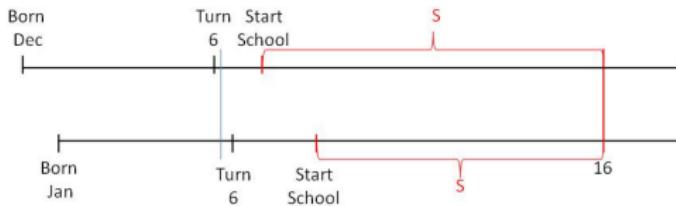
- You can also estimate 2SLS using the auxiliary regression approach we just covered:

```
. reg X Z  
. predict Xhat  
. reg Y Xhat
```

- For the same reasons that you shouldn't actually implement 2SLS manually using the ratio of the reduced form and first stage coefficients, you shouldn't manually use the auxiliary regression approach because, again, the standard errors are incorrect, and any complex missing patterns may leave you with different samples
- This “two stage least squares” interpretation of IV – called an *interpretation*, because it is not the actual suggested procedure – is useful for understanding what IV does, but stick with ivregress

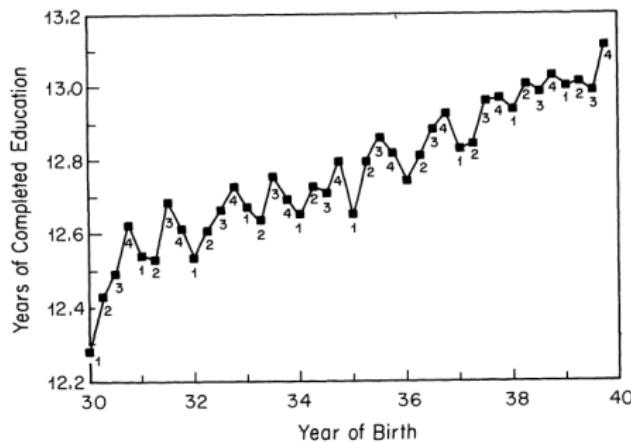
Instrument for Education using Compulsory Schooling Laws

- In practice, it is often difficult to find convincing instruments – usually because potential instruments don't satisfy the exclusion restriction
- In the returns to education literature, Angrist and Krueger (1991) had a very influential study where they used quarter of birth as an instrumental variable for schooling
- In the US, you could drop out of school once you turned 16
- "School districts typically require a student to have turned age six by January 1 of the year in which he or she enters school" (Angrist and Krueger 1991, p. 980)
- Children have different ages when they start school, though, and this creates different lengths of schooling at the time they turn 16 (potential drop out age):



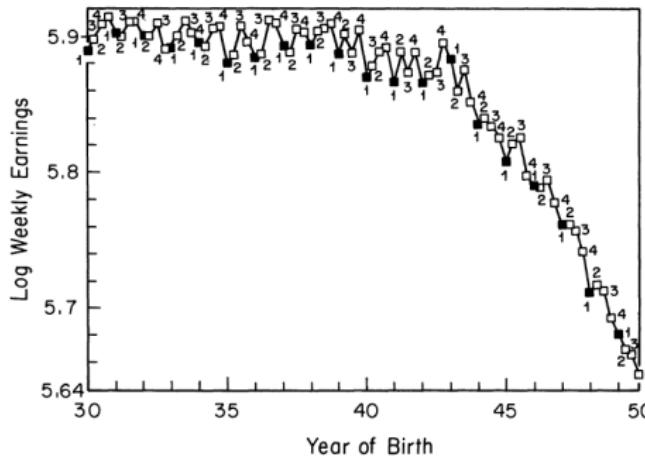
First Stage

- Men born earlier in the year have lower schooling. This indicates that there is a first stage.



Reduced Form

- Do differences in schooling due to different quarter of birth translate into different earnings?



Two Stage Least Squares model

- The first stage regression is:

$$S_i = X\pi_{10} + \pi_{11}Z_i + \eta_{1i}$$

- The reduced form regression is:

$$Y_i = X\pi_{20} + \pi_{21}Z_i + \eta_{2i}$$

- The covariate adjusted IV estimator is the sample analog of the ratio, $\frac{\pi_{21}}{\pi_{11}}$
- Again, how was this estimator calculated?
 - ① Obtain the first stage fitted values:

$$\hat{S}_i = X\hat{\pi}_{10} + \hat{\pi}_{11}Z_i$$

where $\hat{\pi}_{1j}$ for $j = 1, 2$ are OLS estimates of the first stage regression

- ② Plug the first stage fitted values into the “second-stage equation” to then estimate

$$Y_i = \alpha X + \hat{S}_i\rho + \text{error}$$

Two Stage Least Squares

- But as we note, they don't actually manually do this – I remind you of this because the 2SLS intuition is very useful. 2SLS only retains the variation in S generated by the quasi-experimental variation, which we hope is exogenous
- Angrist and Krueger use more than one instrumental variable to instrument for schooling: they include a dummy for each quarter of birth. Their estimated first-stage regression is therefore:

$$S_i = X\pi_{10} + Z_{1i}\pi_{11} + Z_{2i}\pi_{12} + Z_{3i}\pi_{13} + \eta_1$$

- The second stage is the same as before, but the fitted values are from the new first stage

First stage regression results

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect ^a			F-test ^b [P-value]
			I	II	III	
Total years of education	1930–1939	12.79	−0.124 (0.017)	−0.086 (0.017)	−0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	−0.085 (0.012)	−0.035 (0.012)	−0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	−0.019 (0.002)	−0.020 (0.002)	−0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	−0.015 (0.001)	−0.012 (0.001)	−0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	−0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	−0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]
	1940–1949	0.30	−0.003 (0.002)	0.004 (0.002)	0.000 (0.002)	5.0 [0.0018]

First stage regression results

- Quarter of birth is a strong predictor of total years of education

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect ^a			F-test ^b [P-value]
			I	II	III	
Total years of education	1930–1939	12.79	−0.124 (0.017)	−0.086 (0.017)	−0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	−0.085 (0.012)	−0.035 (0.012)	−0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	−0.019 (0.002)	−0.020 (0.002)	−0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	−0.015 (0.001)	−0.012 (0.001)	−0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	−0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	−0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]
	1940–1949	0.30	−0.003 (0.002)	0.004 (0.002)	0.000 (0.002)	5.0 [0.0018]

First stage regression results: Placebos

Completed master's degree	1930–1939	0.09	−0.001	0.002	−0.001	1.7	
	1940–1949	0.11	0.000	0.004	0.001	3.9	
Completed doctoral degree	1930–1939	0.03	0.002	0.003	0.000	2.9	
	1940–1949	0.04	−0.002	0.001	−0.001	4.3	

a. Standard errors are in parentheses. An $MA(+2, -2)$ trend term was subtracted from each dependent variable. The data set contains men from the 1980 Census, 5 percent Public Use Sample. Sample size is 312,718 for 1930–1939 cohort and is 457,181 for 1940–1949 cohort.

b. F -statistic is for a test of the hypothesis that the quarter-of-birth dummies jointly have no effect.

IV Estimates Birth Cohorts 20-29, 1980 Census

Independent variable	(1) OLS	(2) TSLS
Years of education	0.0711 (0.0003)	0.0891 (0.0161)
Race (1 = black)	—	—
SMSA (1 = center city)	—	—
Married (1 = married)	—	—
9 Year-of-birth dummies	Yes	Yes
8 Region-of-residence dummies	No	No
Age	—	—
Age-squared	—	—
χ^2 [dof]	—	25.4 [29]

IV Estimates - including some covariates

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0711 (0.0003)	0.0891 (0.0161)	0.0711 (0.0003)	0.0760 (0.0290)
Race (1 = black)	—	—	—	—
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
Age	—	—	-0.0772 (0.0621)	-0.0801 (0.0645)
Age-squared	—	—	0.0008 (0.0007)	0.0008 (0.0007)
χ^2 [dof]	—	25.4 [29]	—	23.1 [27]

Wald estimator

- They also present an alternative to 2SLS called the Wald estimator – both are versions of instrumental variables
- Recall that 2SLS uses the predicted values from a first stage regression – but we showed that the 2SLS method was equivalent to $\frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)}$
- The Wald estimator simply calculates the return to education as the ratio of the difference in earnings by quarter of birth to the difference in years of education by quarter of birth – it's a version of the above
- Formally, $IV_{Wald} = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)}$

TABLE III
PANEL A: WALD ESTIMATES FOR 1970 CENSUS—MEN BORN 1920–1929^a

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) – (2)
ln (wkly. wage)	5.1484	5.1574	-0.00898 (0.00301)
Education	11.3996	11.5252	-0.1256 (0.0155)
Wald est. of return to education			0.0715 (0.0219)
OLS return to education ^b			0.0801 (0.0004)

Panel B: Wald Estimates for 1980 Census—Men Born 1930–1939

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) – (2)
ln (wkly. wage)	5.8916	5.9027	-0.01110 (0.00274)
Education	12.6881	12.7969	-0.1088 (0.0132)
Wald est. of return to education			0.1020 (0.0239)
OLS return to education			0.0709 (0.0003)

IV Estimates - more covariates and interacting quarter of birth

- They also include specifications where they use 30 (quarter of birth \times year) dummy variables and 150 (quarter of birth \times state) dummies as instrumental variables
 - What's the intuition here? The effect of quarter of birth may vary by birth year or by state
- It reduced the standard errors, but that comes at a cost of potentially having a weak instruments problem

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS
Years of education	0.0673 (0.0003)	0.0928 (0.0093)	0.0673 (0.0003)	0.0907 (0.0107)
Race (1 = black)	—	—	—	—
SMSA (1 = center city)	—	—	—	—
Married (1 = married)	—	—	—	—
9 Year-of-birth dummies	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No
50 State-of-birth dummies	Yes	Yes	Yes	Yes
Age	—	—	-0.0757 (0.0617)	-0.0880 (0.0624)
Age-squared	—	—	0.0008 (0.0007)	0.0009 (0.0007)

plain

Mechanism

- In addition to log weekly wage, they examined the impact of compulsory schooling on log annual salary and weeks worked
- The main impact of compulsory schooling is on the log weekly wage – not on weeks worked

Weak Instruments

- As we mentioned earlier, IV is consistent but biased
- For a long time, researchers were not attentive to this subtle difference and didn't care much about the small sample bias of IV
- But in the early 1990s, a number of papers highlighted that IV can be *severely* biased – in particular, when instruments have only a weak correlation with the endogenous variable of interest and when many instruments are used to instrument for one endogenous variable (i.e., there are many overidentifying restrictions).
- In the worst case, if the instruments are so weak that there is no first stage, then the 2SLS sampling distribution is centered on the probability limit of OLS

Weak instruments and bias towards OLS

- Let's consider a model with a single endogenous regressor and a simple constant treatment effect
- The causal model of interest is:

$$y = \beta x + \nu$$

- The matrix of instrumental variables is Z with the first stage equation:

$$x = \mathbf{Z}'\pi + \eta$$

- If ν_i and η_i are correlated, estimating the first equation by OLS would lead to biased results, wherein the OLS bias is:

$$E[\beta_{OLS} - \beta] = \frac{Cov(\nu, x)}{Var(x)}$$

- If ν_i and η_i are correlated the OLS bias is therefore: $\frac{\sigma_{\nu\eta}}{\sigma_x^2}$

Weak instruments and bias towards OLS

- It can be shown that the bias of 2SLS is approximately:

$$E[\widehat{\beta_{2SLS}} - \beta] \approx \frac{\sigma_{\nu\eta}}{\sigma_{\eta}^2} \frac{1}{F + 1}$$

where F is the population analogue of the F -statistic for the joint significance of the instruments in the first stage regression. See Angrist and Pischke pp. 206-208 for a derivation.

- If the first stage is weak (i.e., $F \rightarrow 0$), then the bias of 2SLS approaches $\frac{\sigma_{\nu\eta}}{\sigma_{\eta}^2}$
- This is the same as the OLS bias as for $\pi = 0$ in the second equation on the earlier slide (i.e., there is no first stage relationship between Z and D) $\sigma_x^2 = \sigma_{\eta}^2$ and therefore the OLS bias $\frac{\sigma_{\nu\eta}}{\sigma_x^2}$ becomes $\frac{\sigma_{\nu\eta}}{\sigma_{\eta}^2}$.
- But if the first stage is very strong ($F \rightarrow \infty$) then the IV bias goes to 0.

Weak Instruments - Adding More Instruments

- Adding more weak instruments will increase the bias of 2SLS
 - By adding further instruments without predictive power, the first stage F -statistic goes toward zero and the bias increases
- If the model is “just identified” – mean the same number of instrumental variables as there are endogenous covariates – weak instrument bias is less of a problem
 - See Angrist and Pischke, p. 209 where they write that IV is “approximately biased” – this is only true if the first stage is not zero
 - See http://econ.lse.ac.uk/staff/spischke/mhe/josh/solon_justid_April14.pdf
- Bound, Jaeger and Baker (1995) highlighted this problem for the Angrist and Krueger study. AK present findings from using different sets of instruments
 - ① Quarter of birth dummies → 3 instruments
 - ② Quarter of birth dummies + (quarter of birth) \times (year of birth) + (quarter of birth) \times (state of birth) → 180 instruments

Adding instruments in Angrist and Krueger

	(1) OLS	(2) IV	(3) OLS	(4) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)
<i>F</i> (excluded instruments)		13.486		4.747
Partial <i>R</i> ² (excluded instruments, $\times 100$)		.012		.043
<i>F</i> (overidentification)		.932		.775
<i>Age Control Variables</i>				
Age, Age ²	x	x		
9 Year of birth dummies			x	x
<i>Excluded Instruments</i>				
Quarter of birth		x		x
Quarter of birth \times year of birth			x	
Number of excluded instruments		3		30

- Adding more weak instruments reduced the first stage *F*-statistic and moves the coefficient towards the OLS coefficient

Adding instruments in Angrist and Krueger

	(1) OLS	(2) IV
Coefficient	.063 (.000)	.083 (.009)
<i>F</i> (excluded instruments)	2.428	
Partial <i>R</i> ² (excluded instruments, $\times 100$)	.133	
<i>F</i> (overidentification)	.919	
<i>Age Control Variables</i>		
Age, Age ²		
9 Year of birth dummies	x	x
<i>Excluded Instruments</i>		
Quarter of birth	x	
Quarter of birth \times year of birth	x	
Quarter of birth \times state of birth	x	
Number of excluded instruments	180	

- Adding more weak instruments reduced the first stage *F*-statistic and moves the coefficient towards the OLS coefficient

What can you do if you have weak instruments?

- With weak instruments, you have the following options:
 - ① Use a just identified model with your strongest IV
 - ② Use a limited information maximum likelihood estimator (LIML). This is approximately median unbiased for over identified constant effects models. It provides the same asymptotic distribution as 2SLS (under constant effects) but provides a finite-sample bias reduction. (LIML is programmed for STATA in the `ivregress` command.)
 - ③ Find stronger instruments.

Practical Tips for IV Papers

- ① Report the first stage
 - Does it make sense?
 - Do the coefficients have the right magnitude and sign?
- ② Report the F -statistic on the excluded instrument(s).
 - Stock, Wright and Yogo (2002) suggest that F -statistics > 10 indicate that you do not have a weak instrument problem – this is not a proof, but more like a rule of thumb
 - If you have more than one endogenous regressor for which you want to instrument, reporting the first stage F -statistic is not enough (because 1 instrument could affect both endogenous variables and the other could have no effect – the model would be under identified). In that case, you want to report the Cragg-Donald EV statistic.

Practical Tips for IV Papers

- ③ If you have many IVs, pick your best instrument and report the just identified model (weak instrument problem is much less problematic)
- ④ Check over identified 2SLS models with LIML
- ⑤ Look at the reduced form
 - The reduced form is estimated with OLS and is therefore unbiased
 - If you can't see the causal relationship of interest in the reduced form, it is probably not there

Angrist (1990) Veteran Draft Lottery

- Angrist (1990) uses the Vietnam draft lottery as an instrumental variable for military service
- In the 1960s and 1970s, young American men were drafted for military service to serve in Vietnam
- Concerns about the fairness of the conscription policy led to the introduction of a draft *lottery* in 1970
- From 1970 to 1972, random sequence numbers were randomly assigned to each birth date in cohorts of 19-year-olds
- Men with lottery numbers below a cutoff were drafted; in other words
 - Higher numbers were *less* likely to be drafted;
 - Lower numbers were *more* likely to be drafted.
- The draft did not perfectly determine military service:
 - Many draft-eligible men were exempt for health and other reasons
 - Exempt men would sometimes volunteer

Summary of Findings on Vietnam Draft Lottery

- ① First stage results: Having a low lottery number (i.e., being eligible for the draft) increased veteran status by about 16 percentage points (the mean of veteran status was 27 percent)
- ② Second stage results: Serving in the army lowers earnings by between \$2,050 and \$2,741 per year

IV with Heterogenous Treatment Effects

- Up to this point, we only considered models where the causal effect was the same for all individuals (i.e., homogenous treatment effects where $Y_i^1 - Y_i^0 = \delta$ for all i units)
- Let's now try to understand what instrumental variables estimation is measuring if treatment effects are *heterogenous* (i.e., $Y_i^1 - Y_i^0 = \delta_i$ which varies across the population)
- Why do we care?
 - We care about internal validity: Does the design successfully uncover causal effects for the population that we are studying?
 - We care about external validity: Does the study's results inform us about different populations?

IV with Heterogenous Treatment Effects

- Similar potential outcomes notation and terminology:
 - Causal chain is $Z_i \rightarrow D_i \rightarrow Y_i$
 - $Y_i(D_i = 0, Z_i = 1)$ is represented as $Y_i(0, 1)$
- “Potential treatment status” versus “observed treatment status”
 - $D_i^1 = i$ ’s treatment status when $Z_i = 1$
 - $D_i^0 = i$ ’s treatment status when $Z_i = 0$
 - Observed treatment status switching equation:

$$\begin{aligned} D_i &= D_i^0 + (D_i^1 - D_i^0)Z_i \\ &= \pi_0 + \pi_1 Z_i + \zeta_i \end{aligned}$$

$$\begin{aligned} \pi_{0i} &= E[D_i^0] \\ \pi_{1i} &= (D_i^1 - D_i^0) \text{ is the heterogenous causal effect of the IV on } D_i. \\ E[\pi_{1i}] &= \text{The average causal effect of } Z_i \text{ on } D_i \end{aligned}$$

Identifying assumptions under heterogenous treatment effects

- ① Stable Unit Treatment Value Assumption (SUTVA)
- ② Random Assignment
- ③ Exclusion Restriction
- ④ Nonzero First Stage
- ⑤ Monotonicity

Stable Unit Treatment Value Assumption (SUTVA)

Stable Unit Treatment Value Assumption (SUTVA)

If $Z_i = Z'_i$, then $D_i(\mathbf{Z}) = D_i(\mathbf{Z}')$

If $Z_i = Z'_i$ and $D_i = D'_i$, then $Y_i(\mathbf{D}, \mathbf{Z}) = Y_i(\mathbf{D}', \mathbf{Z}')$

Interpretation Potential outcomes for each person i are unrelated to the treatment status of other individuals.

Example Veteran status of person at risk of being drafted is not affected by the draft status of others at risk of being drafted.

Implication Rewrite $Y_i(\mathbf{D}, \mathbf{Z})$ as $Y_i(D_i, Z_i)$ and $D_i(\mathbf{Z})$ as $D_i(Z_i)$.

Definition 1: Intention-to-treat effects

Causal effect of Z on D is $D_i^1 - D_i^0$

Causal effect of Z on Y is $Y_i(D_i^1, 1) - Y_i(D_i^0, 0)$

Independence assumption

Independence assumption (e.g., "as good as random assignment")

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

Interpretation The IV is independent of the vector of potential outcomes and potential treatment assignments (i.e. "as good as randomly assigned")

The independence assumption is sufficient for a causal interpretation of the reduced form:

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= E[Y_i(D_i^1, 1)|Z_i = 1] - E[Y_i(D_i^0, 0)|Z_i = 0] \\ &= E[Y_i(D_i^1, 1)] - E[Y_i(D_i^0, 0)] \end{aligned}$$

Independence means that the first stage measures the causal effect of Z_i on D_i :

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

Example Vietnam conscription for military service was based on randomly generated draft lottery numbers. The assignment of draft lottery number was independent of potential earnings or potential military service – as good as random.

Exclusion Restriction

Exclusion Restriction

$$Y(D, Z) = Y(D, Z') \text{ for all } Z, Z', \text{ and for all } D$$

Interpretation Any effect of Z on Y must be via the effect of Z on D . In other words, $Y_i(D_i, Z_i)$ is a function of D only. Or formally:

$$Y_i(D_i, 0) = Y_i(D_i, 1) \text{ for } D = 0, 1$$

Example In the Vietnam draft lottery example, an individual's earnings potential as a veteran or a non-veteran are assumed to be the same regardless of draft eligibility status. The exclusion restriction would be violated if low lottery numbers may have affected schooling (e.g., to avoid the draft). If this was the case, the lottery number would be correlated with earnings for at least two cases:

- ① through its effect on military service
- ② through its effect on educational attainment

Implication Random lottery numbers (independence) does not imply that the exclusion restriction is satisfied

Exclusion restriction

- Use the exclusion restriction to define potential outcomes indexed solely against treatment status:

$$\begin{aligned}Y_i^1 &= Y_i(1, 1) = Y_i(1, 0) \\Y_i^0 &= Y_i(0, 1) = Y_i(0, 0)\end{aligned}$$

- Rewrite the switching equation:

$$\begin{aligned}Y_i &= Y_i(0, Z_i) + [Y_i(1, Z_i) - Y_i(0, Z_i)]D_i \\Y_i &= Y_i^0 + [Y_i^1 - Y_i^0]D_i\end{aligned}$$

- Random coefficients notation for this is:

$$Y_i = \alpha_0 + \rho_i D_i;$$

with $\alpha_0 = E[Y_i^0]$ and $\rho_i = Y_i^1 - Y_i^0$

First stage

Nonzero Average Causal Effect of Z on D

$$E[D_i^1 - D_i^0] \neq 0$$

Interpretation Z has to have some statistically significant effect on the average probability of treatment

Example Having a low lottery number increases the average probability of service.

Monotonicity

Monotonicity

Either $\pi_{1i} \geq 0$ for all i or $\pi_{1i} \leq 0$ for all $i = 1, \dots, N$

Interpretation Recall that $\pi + 1i$ is the reduced form causal effect of the instrumental variable on an individual i 's treatment status. Monotonicity requires that the instrumental variable (weakly) operate in the same direction on all individual units. In other words, while the instrument may have no effect on some people, all those who are affected are affected *in the same direction* (i.e., positively or negatively, but not both).

Draft example While draft eligibility may have had no effect on the probability of military service for some, monotonicity means the draft lottery either shifted people into service, or it shifted people out of service – but it did not do both.

Schooling example In the quarter of birth example for schooling, this assumption may not be satisfied (see Barua and Lang 2009). Being born in the 4th quarter (which typically increases schooling) may have reduced schooling for some because their school enrollment was held back by their parents

Implication Without monotonicity, IV estimators are not guaranteed to estimate a weighted average of the underlying causal effects of the affected group, $Y_i^1 - Y_i^0$.

Local average treatment effect

If all 1-5 assumptions are satisfied, then IV estimates the **local average treatment effect (LATE)** of D on Y :

$$\delta_{IV,LATE} = \frac{\text{Effect of } Z \text{ on } Y}{\text{Effect of } Z \text{ on } D}$$

Instrumental variables (IV) estimand:

$$\begin{aligned}\delta_{IV,LATE} &= \frac{E[Y_i(D_i^1, 1) - Y_i(D_i^0, 0)]}{E[D_i^1 - D_i^0]} \\ &= E[(Y_i^1 - Y_i^0) | D_i^1 - D_i^0 = 1]\end{aligned}$$

Local Average Treatment Effect

- The LATE parameters is the average causal effect of D on Y for those whose treatment status was changed by the instrument, Z
- Vietnam draft lottery example: IV estimates the average effect of military service on earnings for the subpopulation who enrolled in military service because of the draft but would not have served otherwise.
 - In other words, LATE would not tell us what the causal effect of military service was for volunteers or those who were exempted from military service for medical reasons
- We have reviewed the properties of IV with heterogenous treatment effects using a very simple dummy endogenous variable, dummy IV, and no additional controls example.
 - The intuition of LATE generalizes to most cases where we have continuous endogenous variables and instruments, and additional control variables.

More IV Jargon!

- The LATE framework partitions any population with an instrument into potentially 4 groups:
 - ❶ Compliers: The subpopulation with $D_i^1 = 1$ and $D_i^0 = 0$. Their treatment status is affected by the instrument in the “correct direction”.
 - ❷ Always takers: The subpopulation with $D_i^1 = D_i^0 = 1$. They always take the treatment independently of Z .
 - ❸ Never takers: The subpopulation with $D_i^1 = D_i^0 = 0$. They never take the treatment independently of Z .
 - ❹ Defiers: The subpopulation with $D_i^1 = 0$ and $D_i^0 = 1$. Their treatment status is affected by the instrument in the “wrong direction”.
- (You'll notice that increasingly our terms are borrowing from the medical literature where the treatment is taking a pill. Same here.)

Never-Takers

$$D_i^1 - D_i^0 = 0$$
$$Y_i(0, 1) - Y_i(0, 0) = 0$$

By **Exclusion Restriction**, causal effect of Z on Y is zero.

Defier

$$D_i^1 - D_i^0 = -1$$
$$Y_i(0, 1) - Y_i(1, 0) = Y_i(0) - Y_i(1)$$

By **Monotonicity**, no one in this group

Complier

$$D_i^1 - D_i^0 = 1$$
$$Y_i(1, 1) - Y_i(0, 0) = Y_i(1) - Y_i(0)$$

Average Treatment Effect among
Compliers

Always-taker

$$D_i^1 - D_i^0 = 0$$
$$Y_i(1, 1) - Y_i(1, 0) = 0$$

By **Exclusion Restriction**, causal effect of Z on Y is zero.

Monotonicity Ensures that there are no defiers

- Monotonicity ensures that there are no defiers
- Why is it important to not have defiers?
 - If there were defiers, effects on compliers could be (partly) canceled out by opposite effects on defiers
 - One could then observe a reduced form which is close to zero even though treatment effects are positive for everyone (but the compliers are pushed in one direction by the instrument and the defiers in the other direction)

What Does IV (Not) Estimate?

- As outlined above, with all 5 assumptions satisfied, IV estimates the average treatment effect for *compliers*
 - Contrast this with the traditional IV pedagogy with constant treatment effects (i.e., $\delta_i = \delta$ for all i units).
 - Question: What does IV estimate when treatment effects are assumed to be constant?
- Without further assumptions (e.g., constant causal effects), LATE is not informative about effects on never-takers or always-takers because the instrument does not affect their treatment status
- So what? Well, it matters because in most applications, we would be mostly interested in estimating the average treatment effect on the whole population:

$$ATE = E[Y_i^1 - Y_i^0]$$

- But that's not possible usually with IV

Sensitivity to assumptions: exclusion restriction

Example Someone at risk of draft (low lottery number) changes education plans to retain draft deferments and avoid conscription.

Implication Increased bias to IV estimand through two channels:

- Average direct effect of Z on Y for compliers
- Average direct effect of Z on Y for noncompliers multiplied by odds of being a non-complier

Severity Depends on:

- Odds of noncompliance (smaller \rightarrow less bias)
- “Strength” of instrument (stronger \rightarrow less bias)
- Effect of the alternative channel on Y

Sensitivity to assumptions: Monotonicity violations

Example Someone who would have volunteered for Army when not at risk of draft (high lottery number) chooses to avoid military service when at risk of being drafted (low lottery number)

Implication Bias to IV estimand (multiplication of 2 terms):

- Proportion defiers relative to compliers
- Difference in average causal effects of D on Y for compliers and defiers

Severity Depends on:

- Proportion of defiers (small \rightarrow less bias)
- "Strength" of instrument (stronger \rightarrow less bias)
- Variation in effect of D on Y (less \rightarrow less bias)

Summarizing

- The potential outcomes framework gives a more subtle interpretation of what IV is measuring
 - In the constant coefficients world (i.e., traditional pedagogy), IV measures δ which is “the” causal effect of D_i on Y_i , and assumed to be the same for all i units
 - In the random coefficients world, IV measures instead an *average* of heterogeneous causal effects across a particular population – $E[\delta_i]$ for some group of i units
 - IV, therefore, measures the *local average treatment effect* or LATE parameter, which is the average of causal effects across the subpopulation of *compliers*, or those units whose covariate of interest, D_i , is influenced by the instrument.
- Angrist and Evans (1996) example: not every woman whose first two kids share gender will go on to have a third kid.
 - Under heterogeneous treatment effects, Angrist and Evans (1996) identify the causal effect of the gender composition of the first two kids on labor supply
 - Remark: That is not the same thing as identifying the causal effect of children on labor supply; the former is a LATE whereas the latter might be better described as an ATE
- *Ex post* this is probably obvious, but *ex ante* this was a real breakthrough (see Angrist, Imbens and Rubin 1996; Imbens and Angrist 1994)

Other potentially interesting treatment effects

- Another effect which we may be potentially interested in estimating is the familiar estimand, the average treatment effect on the treatment group (ATT)
- Remark: LATE is *not* the same as ATT, though.

$$\underbrace{E[Y_i^1 - Y_i^0 | D_i = 1]}_{\text{ATT}} = \underbrace{E[Y_i^1 - Y_i^0 | D_i^0 = 1]}_{\text{Effect on always takers}} P[D_i^0 = 1 | D_i = 1] + \underbrace{E[Y_i^1 - Y_i^0 | D_i^1 > D_i^0]}_{\text{Effect on compliers}} P[D_i^1 > D_i^0, Z_i = 1 | D_i = 1]$$

- The average treatment effect on the treated, ATT, is a weighted average of the effects on always-takers and compliers.
- If there are no always takers we can, however, estimate ATT which is equal to LATE in that case.

Discussions and questions

- When might we *not* be interested in the local average treatment effect?
 - Romneycare in Massachusetts examples: If the compliance rate or treatment effects differ in the community than during some quasi-experimental expansion, are we more interested in LATE, ATT or ATE?
 - What might be other examples of this in economics? In education?
 - This has a similar flavor to methodological questions regarding a study's "external" vs. "internal validity"
- What do we make of the fact that LATE is defined for an *unobservable sub-population* (i.e., can't label all units in the population as compliers or noncompliers)?
- What do we make of the fact that IV identification is based on a set of untestable assumptions?
 - For example: colonial settler mortality (Z) influences economic development (Y) *only through* Z 's association with human capital accumulation rather than institutions, D (Acemoglu, Johnson and Robinson, 2001).

IV in Randomized Trials

- The use of IV methods may be helpful when evaluating a randomized trial
- In many randomized trials, participation is nonetheless voluntary among those randomly assigned to treatment
- On the other hand, persons in the control group usually do not have access to treatment
 - only those who are particularly likely to benefit from treatment therefore will probably take up treatment which almost always leads to positive selection bias
 - if you just compare means between treated and untreated individuals using OLS, you will obtain biased treatment effects *even for the randomized trial* due to non-compliance
- Solution: instrument for treatment with whether you were offered treatment and estimate LATE

Dean Eckles (Facebook), Rene Kizilcec and Eytan Bakshy (2015). “Identifying Peer Effects in Social Networks with Peer Encouragement Designs”

Peer encouragement designs

- Dean Eckles, et al. (2015) randomly assign vertices in Facebook network to encouragement to behavior of interest and examine how this spills over to others
 - Indirectly affect behaviors of existing peers
- Eckles peer encouragement design allows them to estimate the effect of peer behaviors itself through instrumental variables – not just the encouragement itself
- An alternative view: assign peers to behaviors, see effect it has on the person.

Effects of receiving feedback

- Motivation – when an individual shares content in social media, what are the effects of receiving additional feedback like comments and “likes”?
 - Generating further conversation – ego replies to comments
 - In-kind peer effects in giving feedback (reciprocity)
 - Creating and sharing more content in the future
- How do these effects vary by prior ego behaviors?
- How can minor user interface changes affect user experiences and engagement? How should we make these tradeoffs?

Goals

- Goal 1: Estimate peer effects
 - How does a marginal peer adopting affect your adoption?
 - Ideal experiment: directly assign behaviors of existing peers
- Goal 2: Estimate effects of global treatment
 - What would happen if we gave everyone the treatment?
 - Average treatment effect (ATE) of global treatment Z_1 vs. Z_0

$$\delta(z_1, z_0) = \frac{1}{N} \sum_i E[Y_i(Z = z_1) - Y_i(Z = z_0)]$$

- Ideal experiment: *assign connected components to treatment*
- This is an active area of research (Eckles, Karrer, Ugander 2014; Ugander, et al 2013)

What are “Encouragement designs”?

- Randomly assign units to encouragement Z to a focal behavior D
 - Examples: randomly encourage students to study (Powers and Swinton 1984)
 - Assign to take a drug or not (but they may not take it)
- Formal analysis using potential outcomes model
 - Total effect of encouragement: intent-to-treat (ITT)

$$Y_i(Z_i = 1) - Y_i(Z_i = 0)$$

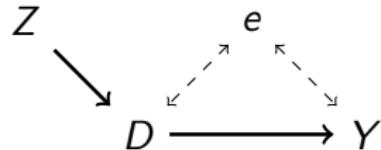
- Effect of behavior (i.e., effect of D on Y)

$$Y_i(D_i = 1) - Y_i(D_i = 0)$$

- Can we use Z to estimate the effect of D on Y ?

Encouragement designs

- The *best* instrumental variables are randomly assigned
- Find a variable Z that affects D but is otherwise unrelated to Y
- Use this exogenous variation in D to estimate the effect of D on Y



IV analysis

- Number of encouraged peers is an instrumental variable for peer effects
 - Complete mediation (i.e., “exclusion restriction”) – Peer encouragement only affects ego behavior via peer behavior
 - Even if effects are heterogenous, IV analysis of encouragement designs identifies the LATE for the complier subpopulation
 - Eckles, et al (2015) argue that this is likely an advantage over other instruments that aren’t encouragements

So what exactly is the instrument??

- Facebook claim as fact: Pre-expanding vs. not pre-expanding comment boxes (Z) increases feedback (D)

Expanded comment box



Eytan Bakshy

chicken pho | smitten kitchen
smittenkitchen.com

A home cooking weblog from a tiny kitchen in New York City. The place to find all of your new favorite things to cook.

Like · Comment · Share ·  4  2  1 · 2 hours ago · 

 Erica Stone, Michael Bernstein and 2 others like this.

 1 share

Eytan Bakshy "Goodbye, Jewish grandmother chicken noodle soup; we had a good run"
2 hours ago · Like

Write a comment... 

Higher interaction rate

Unexpanded comment box



Eytan Bakshy

chicken pho | smitten kitchen
smittenkitchen.com

A home cooking weblog from a tiny kitchen in New York City. The place to find all of your new favorite things to cook.

Like · Comment · Share ·  4  2  1 · 2 hours ago · 

Lower interaction rate

Binary encouragement designs

- Four types of people by potential outcomes
 - **Compliers** - treatment if encouraged, control if not
 - **Always-takers** - treatment whether encouraged or not
 - **Never-takers** - control whether encouraged or not
 - **Defiers** - control if encouraged, treatment if not encouraged
- Not all of these may exist in a particular study
- In a trial of a new drug or offering (removing) a new (existing) feature, there are neither always-takers nor defiers

Binary encouragement designs

	D=0	D=1
Z=0	Compliers and Never-takers	Defiers and Always-takers
Z=1	Defiers and Never-takers	Compliers and Always-takers

Heterogenous treatment effects

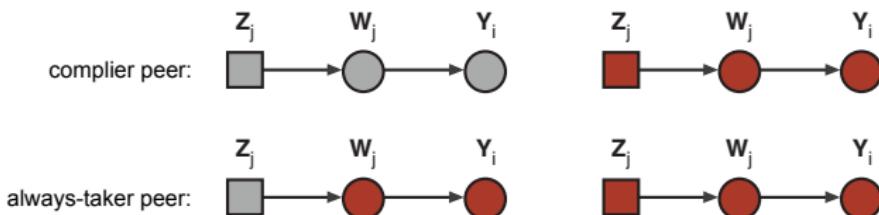
- *Monotonicity*: With probability 1, $D_i^z \geq D_i^{z'}$ for all $z \geq z'$ and all i
- Then local average treatment effect (LATE) is identified
- In binary Z, D case, LATE is the average treatment effect for the population of compliers
- We always have to ask ourselves: are we interested in the LATE?

Peer encouragement designs with a single behavior of interest

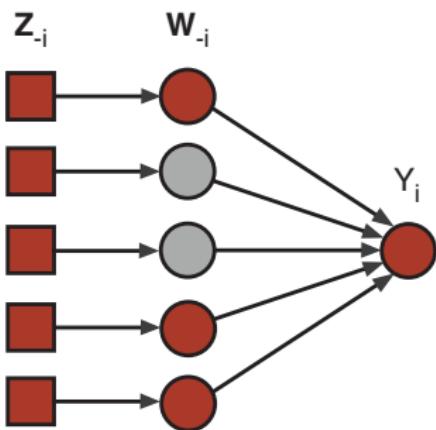
- Assign to encouragement to:
 - enroll in a retirement savings account (Duflo and Saez 2003)
 - post a thankful status update on Thanksgiving Day
- Summarize peer assignments (e.g., number of peers assigned)
- Summarize peer behaviors (e.g., number of adopter peers)
- Compute average ego behaviors as a function of these

Peer encouragement with dyads

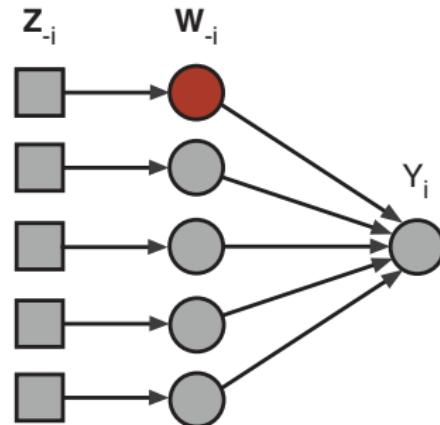
- 1 Randomly encourage j or not
- 2 Observe j 's behavior (endogenous treatment for i)
- 3 Observe i 's behavior (outcome)



Peer encouragement with multiple peers

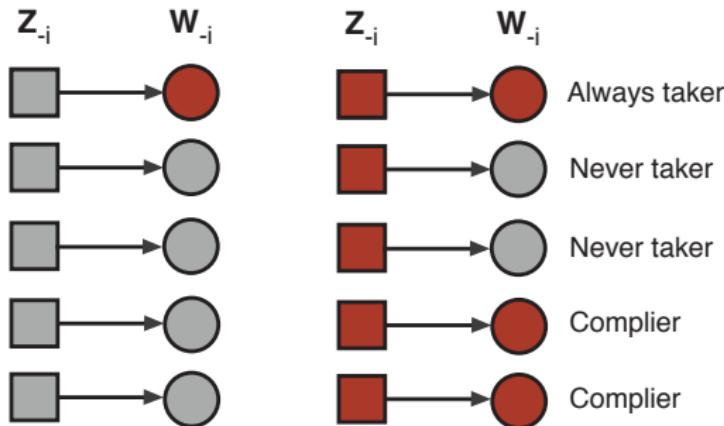


Encourage all peers



Encourage no peers

Noncompliance with multiple peers



Reduced form: Intent-to-treat (ITT)

- One option is forego the IV analysis and instead estimate the ITT
- In other words, analyze ego outcome Y as a function of the number of peers assigned Z
- What else?
 - Compute probabilities of assignment, $Pr(Z)$
 - Use inverse probability weighting to estimate average outcomes

Statistical inference for spillovers

- Other issue is statistical inference.
- Briefly: statistical inference in a network application needs to account for network autocorrelation in regressors (i.e., regressors are correlated within a network)
 - Number of peers treated has high auto-correlation
- “Randomization inference” – increasingly used. Possible to do exact testing for spillovers using new versions of conditional randomization inference (Athey, Barrios, Eckles and Imbens 2015)

Instrumental variables analysis

- Ego's outcome caused by own assignment D and peer behaviors Y
- The number of adopter peers is what matters – how many peers will be “encouraged”
- Eckles, et al (2015) suggest two-stage least squares or other IV methods (e.g., LIML)
- See Bramoullé et al (2009) “Identification of Peer Effects through Social Networks”, *Journal of Econometrics* for exact formulation of using “peers of peers IV”
- Exclusion restriction requires that peer of peer behavior or covariates only affect ego behavior through peer behavior

Results

- Unfortunately, empirical results are currently not being circulated from this study so this is from memory
- First stage: the effect of encouragement on feedback was very large. Receiving expanding boxes resulted in peers leaving more comments.
- Effects on behavior Y : the IV analysis showed that peers influence our behavior – we are more likely to leave new comments, reciprocal comments, likes, and even post new updates
- These effects were very large – in both stages

Lottery designs

- Another design is to use a randomized lottery as an instrumental variable for some treatment
- Examples might be randomized lottery for attending charter schools to study effect of charter schools on educational outcomes
- We'll discuss two papers from 2012 and 2014 evaluating a lottery-based expansion of Medicaid health insurance on Oregon on numerous health and financial outcomes

Overarching question

- What are the effects of expanding access to public health insurance for low income adults?
 - Magnitudes, and even the signs, associated with that question were uncertain
- Limited existing evidence
 - Institute of Medicine review of evidence was suggestive, but a lot of uncertainty
 - Observational studies are confounded by selection into health insurance
 - Quasi-experimental work often focuses on elderly and children
 - Only one randomized experiment in a developed country: the RAND health insurance experiment
 - 1970s experiment on a general population
 - Randomized cost-sharing, not coverage itself

The Oregon Health Insurance Experiment

- Setting: Oregon Health Plan Standard
 - Oregon's Medicaid expansion program for poor adults
 - Eligibility
 - Poor (<100% federal poverty line) adults 19-64
 - Not eligible for other programs
 - Uninsured > 6 months
 - Legal residents
 - Comprehensive coverage (no dental or vision)
 - Minimum cost-sharing
 - Similar to other states in payments, management
 - Closed to new enrollment in 2004

The Oregon Medicaid Experiment

- Lottery
 - Waiver to operate lottery
 - 5-week sign-up period, heavy advertising (January to February 2008)
 - Low barriers to sign up, no eligibility pre-screening
 - Limited information on list
 - Randomly drew 30,000 out of 85,000 on list (March-October 2008)
 - Those selected given chance to apply
 - Treatment at household level
 - Had to return application within 45 days
 - 60% applied; 50% of those deemed eligible → 10,000 enrollees

Oregon Health Insurance Experiment

- Evaluate effects of Medicaid using lottery as randomized controlled trial (RCT)
 - Intent-to-treat: Reduced form comparison of outcomes between treatment group (lottery selected individuals) and controls (not selected)
 - LATE: IV using lottery as instrument for insurance coverage
 - First stage: about a 25 percentage point increase in insurance coverage
 - Archived analysis plan
 - Massive data collect effort – primary and secondary
- Similar to ACA expansion but limits to generalizability
 - Partial equilibrium vs. General equilibrium
 - Mandate and external validity
 - Oregon vs. other states
 - Short vs. Long-run

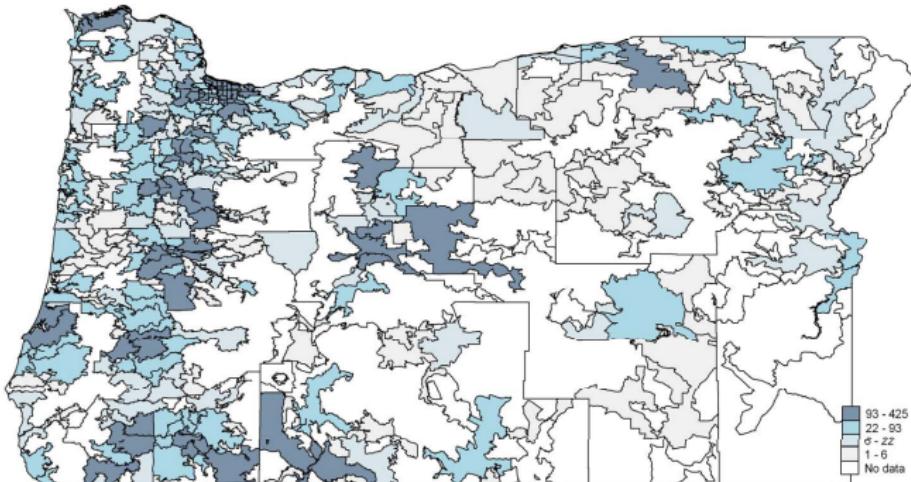
Examine Broad Range of Outcomes

- Costs: Health care utilization
 - Insurance increases resources (income) and lowers price, increasing utilization
 - But improved efficiency (and improved health), decreasing utilization ("offset")
 - Additional uncertainty when comparing Medicaid to no insurance
- Benefits I: Financial risk exposure
 - Insurance supposed to smooth consumption
 - But for very low income, is most care *de jure* or *de facto* free?
- Benefits II: Health
 - Expected to improve (via increased quantity / quality of care)
 - But could discourage health investments ("ex ante moral hazard")

Data

- Pre-randomization demographic information
 - From lottery sign-up
- State administrative records on Medicaid enrollment
 - Primary measure of first stage (i.e., insurance coverage)
- Outcomes
 - Administrative data (~16 months post-notification): Hospital discharge data, mortality, credit reports
 - Mail surveys (~15 months): some questions ask 6-month look-back; some ask current
 - In-person survey and measurements (~25 months): Detailed questionnaires, blood samples, blood pressure, body mass index

Lottery List Distribution Across Zip Codes



Empirical Framework

- Reduced form – estimates of the causal effect of lottery selection

$$Y_{ihj} = \beta_0 + \beta_1 \text{LOTTERY}_h + X_{ih}\beta_2 + V_{ih}\beta_3 + \varepsilon_{ihj}$$

- Validity of experimental design: randomization; balance on treatment and control
- Instrumental variables – effect of insurance coverage

$$\begin{aligned} \text{INSURANCE}_{ihj} &= \delta_0 + \delta_1 \text{LOTTERY}_{ih} + X_{ih}\delta_2 + V_{ih}\delta_3 + \mu_{ihj} \\ y_{ihj} &= \pi_0 + \pi_1 \text{INSURANCE}_{ih} + X_{ih}\pi_2 + V_{ih}\pi_3 + v_{ihj} \end{aligned}$$

- Effect of lottery on coverage: about 25 percentage points
- Additional assumption for causality: primary pathway
 - Could affect participation in other programs, but actually small
 - “Warm glow” of winning – especially early
- Analysis plan, multiple inference adjustment

Effect of lottery on coverage (first stage)

	Full sample		Credit subsample		Survey respondents	
	Control mean	Estimated FS	Control mean	Estimated FS	Control mean	Estimated FS
Ever on Medicaid	0.141 (0.004)	0.256 (0.004)	0.135 (0.004)	0.255 (0.004)	0.135 (0.007)	0.290 (0.007)
Ever on OHP Standard	0.027 (0.003)	0.264 (0.003)	0.028 (0.004)	0.264 (0.004)	0.026 (0.005)	0.302 (0.005)
# of Months on Medicaid	1.408 (0.045)	3.355 (0.045)	1.352 (0.055)	3.366 (0.055)	1.509 (0.055)	3.943 -0.09
On Medicaid, end of study period	0.106 (0.003)	0.148 (0.003)	0.101 (0.004)	0.151 (0.004)	0.105 (0.006)	0.189 (0.006)
Currently have any insurance (self report)					0.325 (0.008)	0.179 (0.008)
Currently have private ins. (self report)					0.128 (0.005)	-0.008 (0.005)
Currently on Medicaid (self report)					0.117 (0.006)	0.197 (0.006)
Currently on Medicaid					0.093 (0.006)	0.177 (0.006)

Amy Finkelstein, et al. (2012). “The Oregon Health Insurance Experiment: Evidence from the First Year”, Quarterly Journal of Economics, vol. 127, issue 3, August.

Effects of Medicaid

- Use primary and secondary data to gauge 1-year effects
 - Mail surveys: 70,000 surveys at baseline, 12 months
 - Administrative data
 - Medicaid enrollment records
 - Statewide Hospital discharge data, 2007-2010
 - Credit report data, 2007-2010
 - Mortality data, 2007-2010

Mail survey data

- **Fielding protocol**

- ~70,000 people, surveyed at baseline and 12 months later
- Basic protocol: three-stage male survey protocol, English/Spanish
- Intensive protocol on a 30% subsample included additional tracking, mailings, phone attempts (done to adjust for non-response bias)

- **Response rate**

- Effective response rate = 50%
- Non-response bias always possible, but response rate and pre-randomization measures in administrative data were balanced between treatment and control

Administrative data

- **Medicaid records**
 - Pre-randomization demographics from list
 - Enrollment records to assess “first stage” (how many of the selected got insurance coverage)
- **Hospital discharge data**
 - Probabilistically matched to list, de-identified at Oregon Health Plan
 - Includes dates and source of admissions, diagnoses, procedures, length of stay, hospital identifier
 - Includes years before and after randomization
- **Other data**
 - Mortality data from Oregon death records
 - Credit report data, probabilistically matched, de-identified

Sample

- 89,824 unique individuals on the waiting list
- Sample exclusions (based on pre-randomization data only)
 - Ineligible for OHP Standard (out of state address, age, etc.)
 - Individuals with institutional addresses on list
- Final sample: 79,922 individuals out of 66,385 households
 - 29,834 treated individuals (surveyed 29,589)
 - 40,088 control individuals (surveyed 28,816)

Sample characteristics

Variable	Mean	Variable	Mean
Panel A: Full sample			
% Female	0.56	Average Age	41
Panel B: Survey responders only			
<i>Demographics:</i>		<i>Health Status: Ever diagnosed with:</i>	
% White	0.82	Diabetes	0.18
% Black	0.04	Asthma	0.28
% Spanish/Hispanic/Latino	0.12	High Blood Pressure	0.40
% High school or less	0.67	Emphysema or Chronic Bronchitis	0.13
% don't currently work	0.55	Depression	0.56
<i>Determinants of eligibility:</i>			
Average hh income (2008)	13,050	% with any insurance	0.33
% below Federal poverty line	0.68	% with private insurance	0.13

Outcomes

- **Access and use of care**
 - Is access to care improved? Do the insured use more care? Is there a shift in the types of care being used?
 - Mail surveys and hospital discharge data
- **Financial strain**
 - How much does insurance protect against financial strain?
 - What are the out-of-pocket implications?
 - Mail surveys and credit reports
- **Health**
 - What are the short-term impacts on self-reported physical and mental health?
 - Mail surveys and vital statistics (mortality)

Effect of lottery on coverage (first stage)

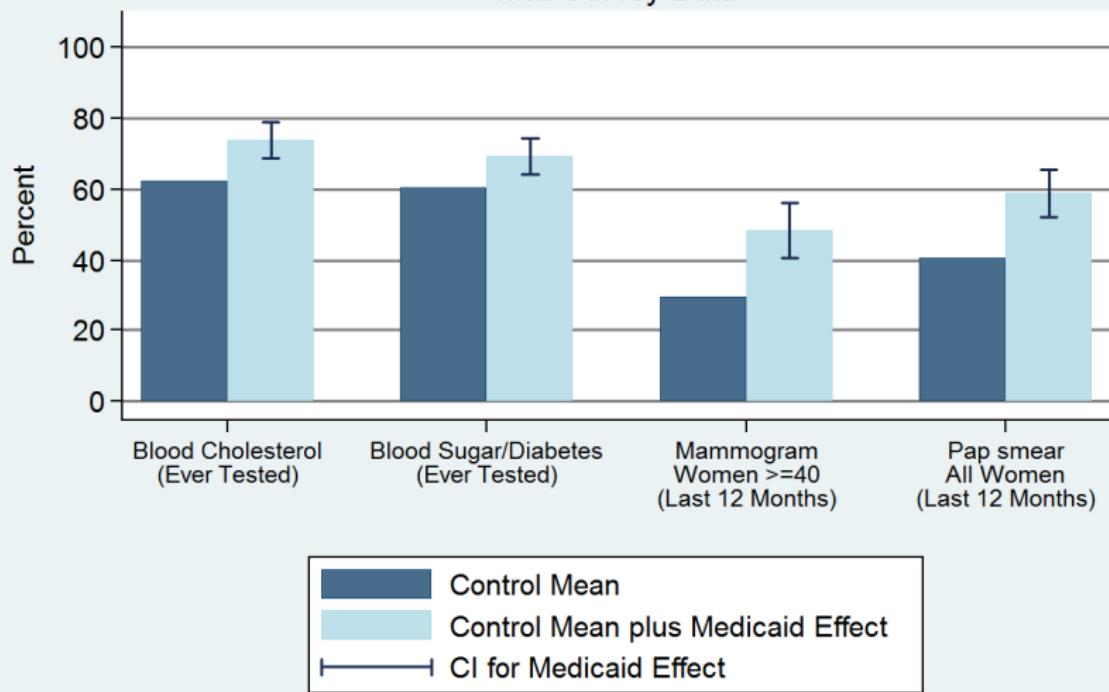
- Gaining insurance resulted in better access to care and higher satisfaction with care (conditional on actually getting care)

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Have a usual place of care	49.9%	+9.9%	+33.9%	.0001
Have a personal doctor	49.0%	+8.1%	+28.0%	.0001
Got all needed health care	68.4%	+6.9%	+23.9%	.0001
Got all needed prescriptions	76.5%	+5.6%	+19.5%	.0001
Satisfied with quality of care	70.8%	+4.3%	+14.2%	.001

SOURCE: Survey data

Preventive Care

Mail Survey Data



Effect of lottery on coverage (first stage)

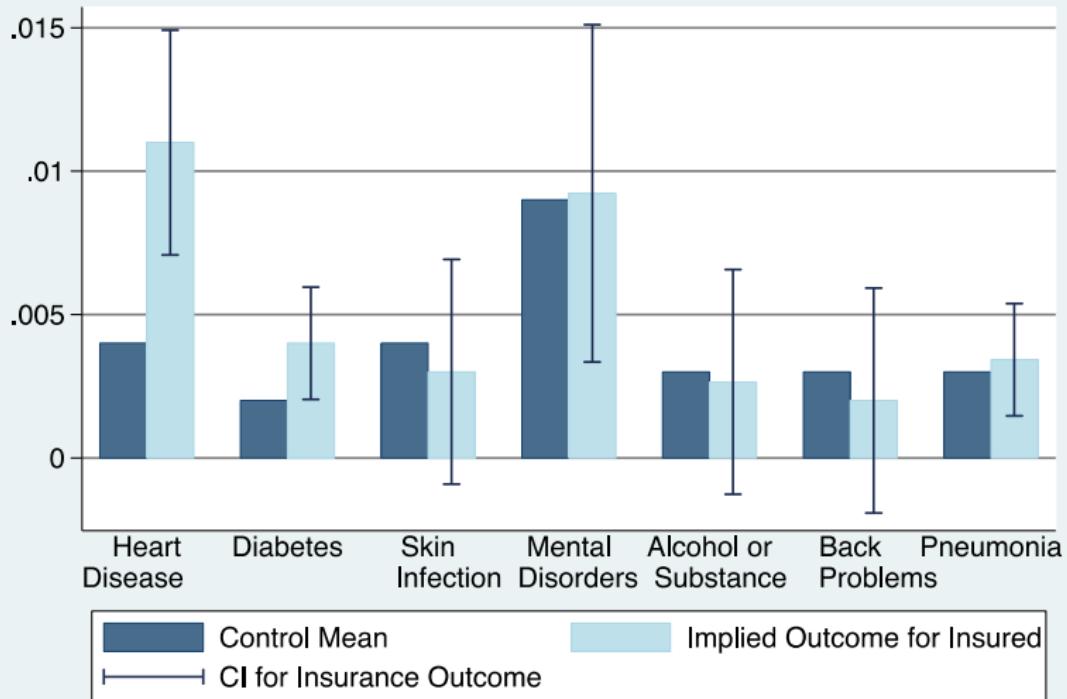
- Gaining insurance resulted in increased probability of hospital admissions, primarily driven by non-emergency department admissions

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Any hospital admission	6.7%	+.50%	+2.1%	.004
--Admits through ED	4.8%	+.2%	+.7%	.265
--Admits NOT through ED	2.9%	+.4%	+1.6%	.002

SOURCE: Hospital Discharge Data

Overall, this represents a 30% higher probability of admission, although admissions are still rare events

Hospital Utilization for Selected Conditions



Summary: Access and use of care

- Overall, utilization and costs went up relative to controls
 - 30% increase in probability of an inpatient admission
 - 35% increase in probability of an outpatient visit
 - 15% increase in probability of taking prescription medications
 - Total \$777 increase in average spending (a 25% increase)
- With this increased spending, those who gained insurance were
 - 35% more likely to get all needed care
 - 25% more likely to get all needed medications
 - Far more likely to follow preventive care guidelines, such as mammograms (60%) and PAP tests (45%)

Results: Financial Strain

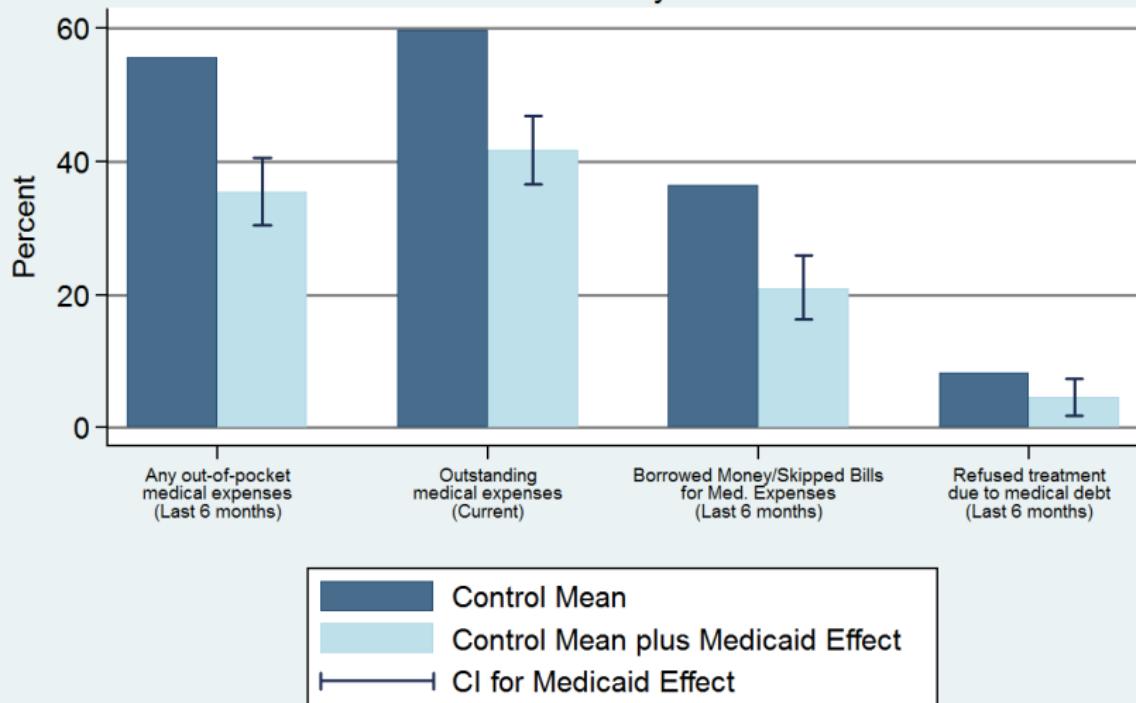
- Gaining insurance resulted in a reduced probability of having medical collections in credit reports, and in lower amounts owed

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Had a bankruptcy	1.4%	+0.2%	+0.9%	.358
Had a collection	50.0%	-1.2%	-4.8%	.013
--Medical collections	28.1%	-1.6%	-6.4%	.0001
--Non-medical collections	39.2%	-0.5	-1.8%	.455
\$ owed medical collections	\$1,999	-\$99	-\$390	.025

Source: Credit report data

Self-reported Financial Strain

Mail Survey Data



Summary: Financial Strain

- Overall, reductions in collections on credit reports were evident
 - 25% decrease in probability of a medical collection
 - Those with a collection owed significantly less
- Household financial strain related to medical costs was mitigated
 - Substantial reduction across all financial strain measures
 - Captures “informal channels” people use to make it work
- Implications for both patients and providers
 - Only 2% of bills sent to collections are ever paid

Results: Self-reported health

- Self-reported measures showed significant improvements one year after randomization

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Health good, v good, excellent	54.8%	+3.9%	+13.3%	.0001
Health stable or improving	71.4%	+3.3%	+11.3%	.0001
Depression screen NEGATIVE	67.1%	+2.3%	+7.8%	.003
CDC Healthy Days (physical)	21.86	+.381	+1.31	.018
CDC Healthy Days (mental)	18.73	+.603	+2.08	.003

Source: Survey data

Summary: Self-reported health

- Overall, big improvements in self-reported physical and mental health
 - 25% increase in probability of good, very good or excellent health
 - 10% decrease in probability of screening for depression
- Physical health measures open to several interpretations
 - Improvements consistent with findings of increased utilization, better access, and improved quality
 - BUT in their baseline surveys, results appeared shortly after coverage ($\sim 2/3$ rd magnitude of full result)
 - May suggest increase in *perception* of well-being rather than physical health
- Biomarker data can shed light on this issue

Discussion

- At 1 year, found increases in utilization, reductions in financial strain, and improvements in self-reported health
 - Medicaid expansion had benefits and costs – didn't "pay for itself"
 - Confirmed biases inherent in observational studies – would have estimated bigger increases in use and smaller improvements in outcomes
- Policy-makers may have different views on value of different aspects of improved well-being
 - "I have an incredible amount of fear because I don't know if the cancer has spread or not."
 - "A lot of times I wanted to rob a bank so I could pay for the medicine I was just so scared . . . People with cancer either have a good chance or no chance. In my case it's hard to recover from lung cancer but it's possible. Insurance took so long to kick in that I didn't think I would get it. Now there is a big bright light shining on me." (Anecdotes)
- Important to have broad evidence on multifaceted effects of Medicaid expansions

Baicker, Katherine, et al. (2014). “The Oregon Experiment – Effects of Medicaid on Clinical Outcomes”, The New England Journal of Medicine.

In-person data collection

- Questionnaire and health examination including
 - Survey questions
 - Anthropometric and blood pressure measurement
 - Dried blood spot collection
 - Catalog of all medications
- Fielded between September 2009 and December 2010
 - Average response ~25 months after lottery began
- Limited to Portland area: 20,745 person sample
- 12,229 interviews for effective response rate of 73%

Analytic approach

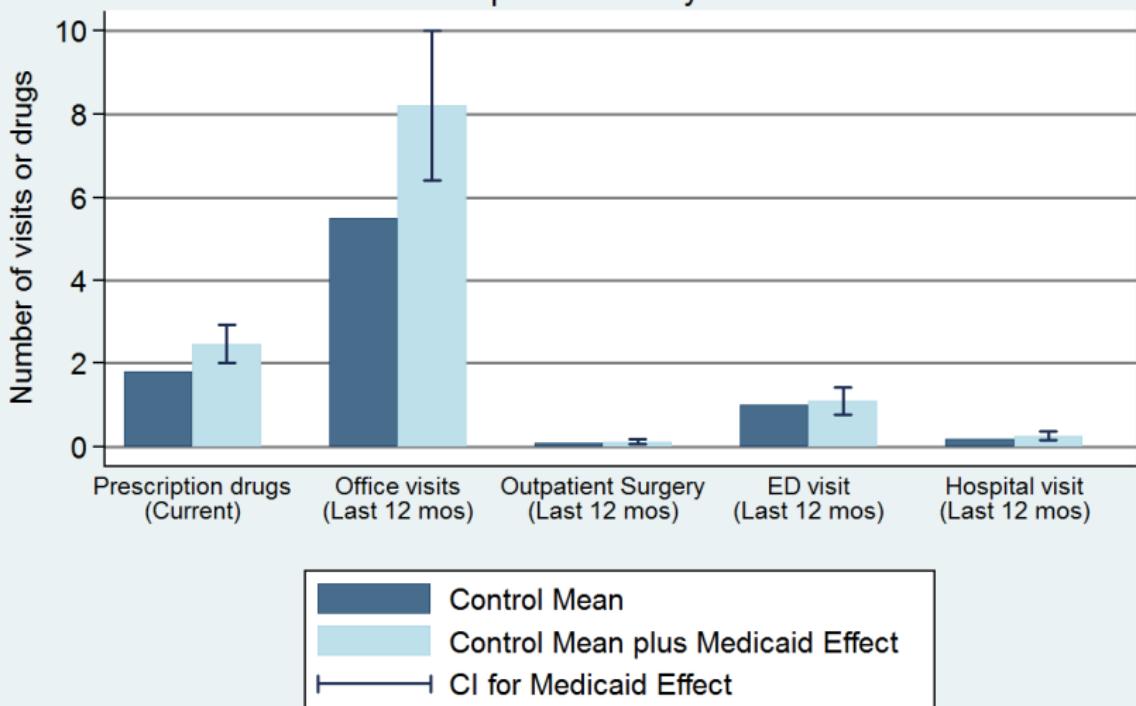
- Intent to treat effect of *lottery selection*
 - Comparing all selected with all not selected
 - Random treatment assignment
 - No differential selection for outcome measurement
- Local average treatment effect on *Medicaid coverage*
 - Using lottery selection as an instrument for coverage
 - ~24 percentage point increase in Medicaid enrollment
 - No change in private insurance (no crowd-out)
 - No effect of lottery except via Medicaid coverage
- Statistical inference is the same for both

Results

- ① *Health care use*
- ② Financial strain
- ③ Clinical health outcomes

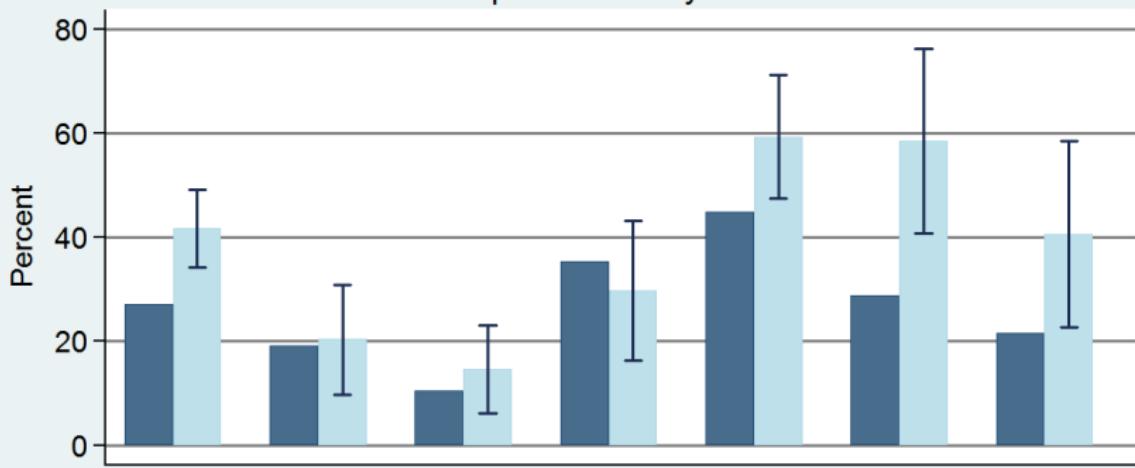
Health Care Utilization

Inperson Survey Data



Preventive Care (Last 12 Months)

Inperson Survey Data



Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

Health care use results

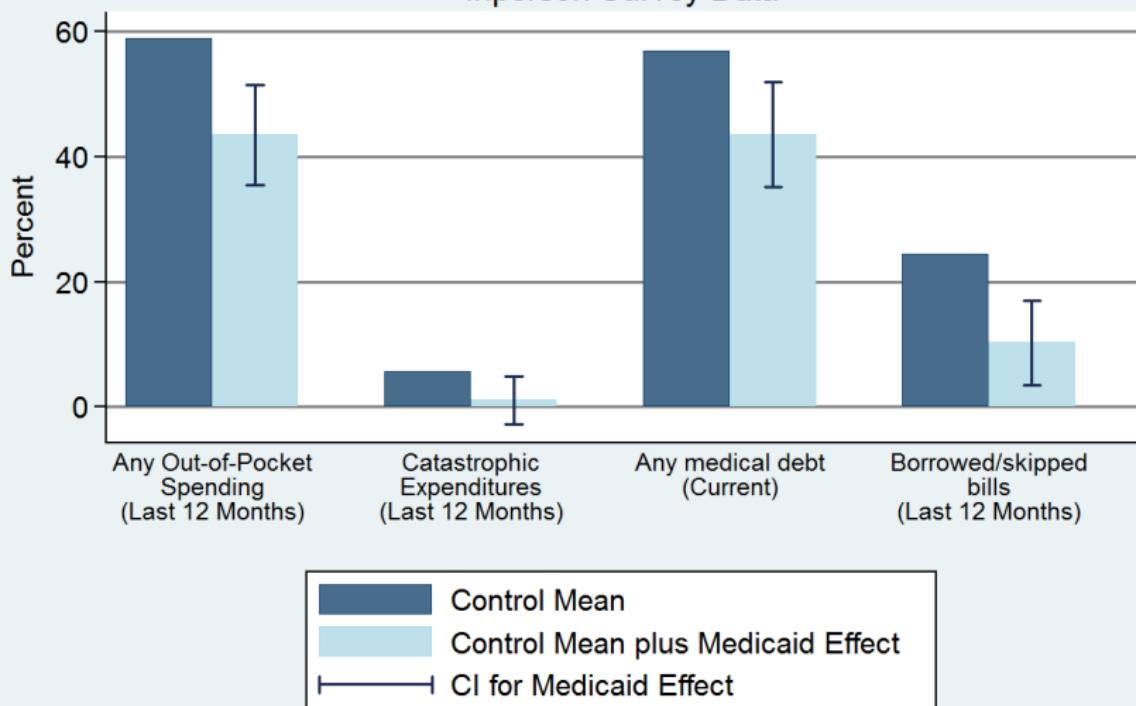
- Increases in use in various settings
 - Increases in probability and number of outpatient visits
 - Increases in probability and number of prescription drugs
 - No discernible change in hospital or ED use (imprecise)
- Increases in preventive care across range of services
- Increases in perceived access and quality
- Implied 35% increase in spending for insured

Results

- ① Health care use
- ② *Financial strain*
- ③ Clinical health outcomes

Financial Hardship

Inperson Survey Data



Financial Hardship Results

- Reduction in strain, out-of-pocket (OOP), money owed
 - Substantial reduction across measures
 - Elimination of catastrophic OOP health spending
- Implications for distribution of burden/benefits
 - Some borne by patients, some by providers
 - Non-financial burden of medical expenses and debt

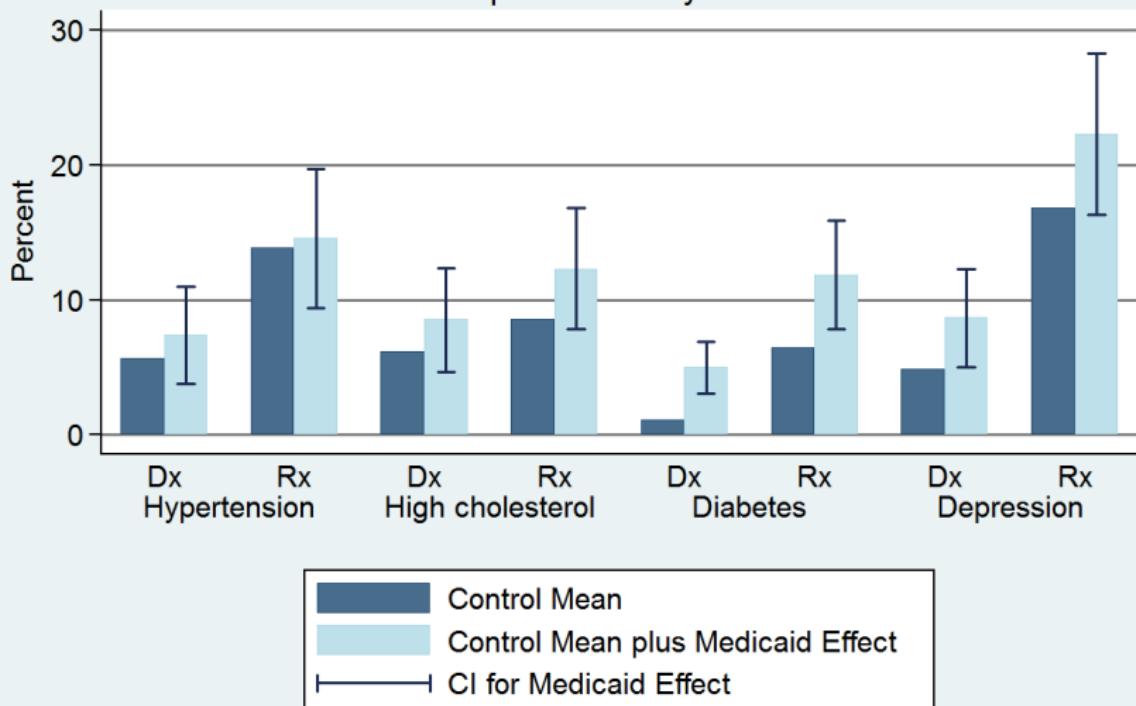
Results

- ① Health care use
- ② Financial strain
- ③ *Clinical health outcomes*

Focusing on specific conditions

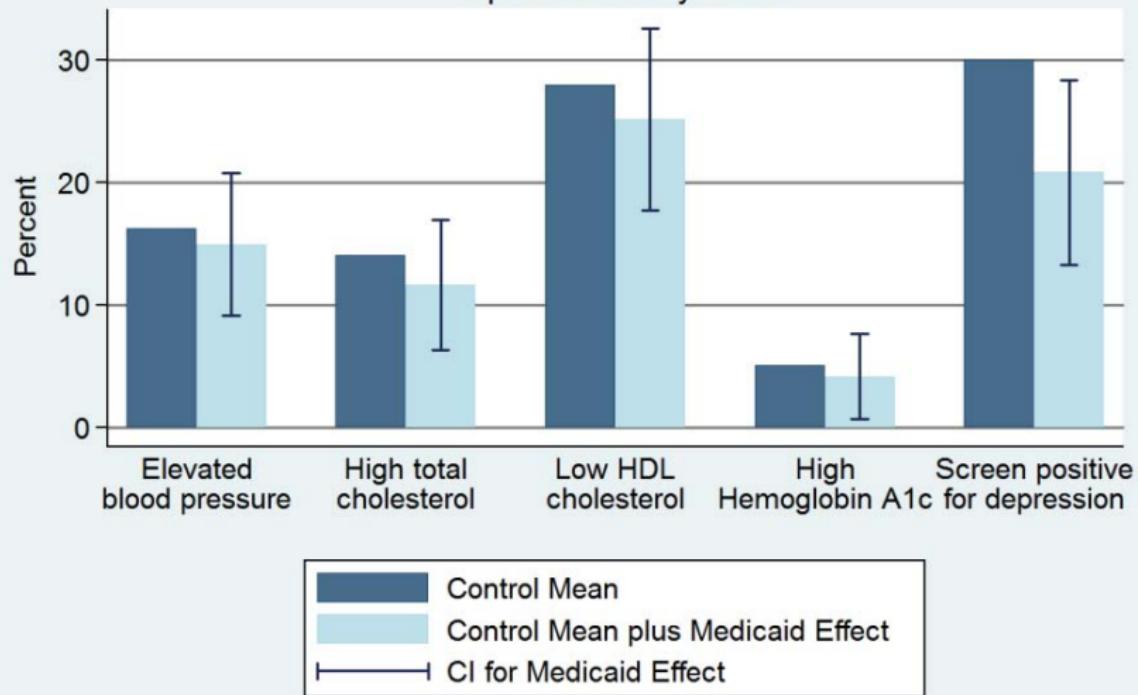
- Measured:
 - Blood pressure
 - Cholesterol levels
 - Glycated hemoglobin
 - Depression
- Reasons for selecting these:
 - Reasonably prevalent conditions
 - Clinically effective medications exist
 - Markers of longer term risk of cardiovascular disease
 - Can be measured by trained interviewers and lab tests
- A limited window into health status

Post-lottery Diagnosis (Dx) and Current Medication (Rx) Inperson Survey Data

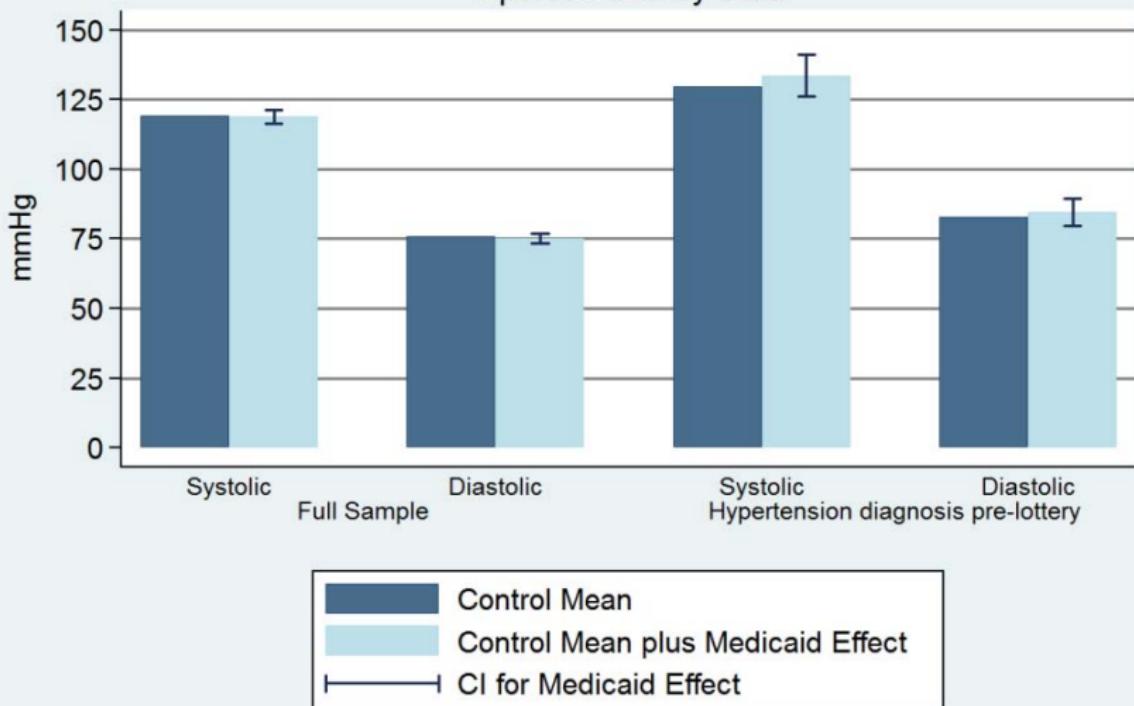


Current Clinical Measures

Inperson Survey Data

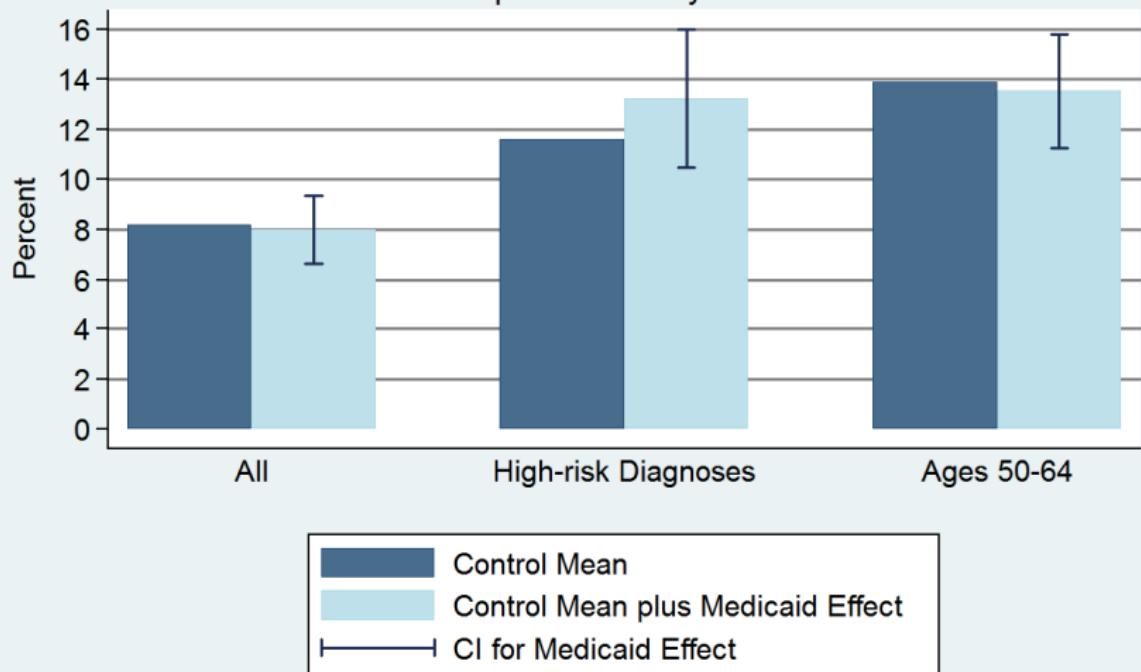


Blood Pressure Inperson Survey Data



Framingham Risk Scores

Inperson Survey Data



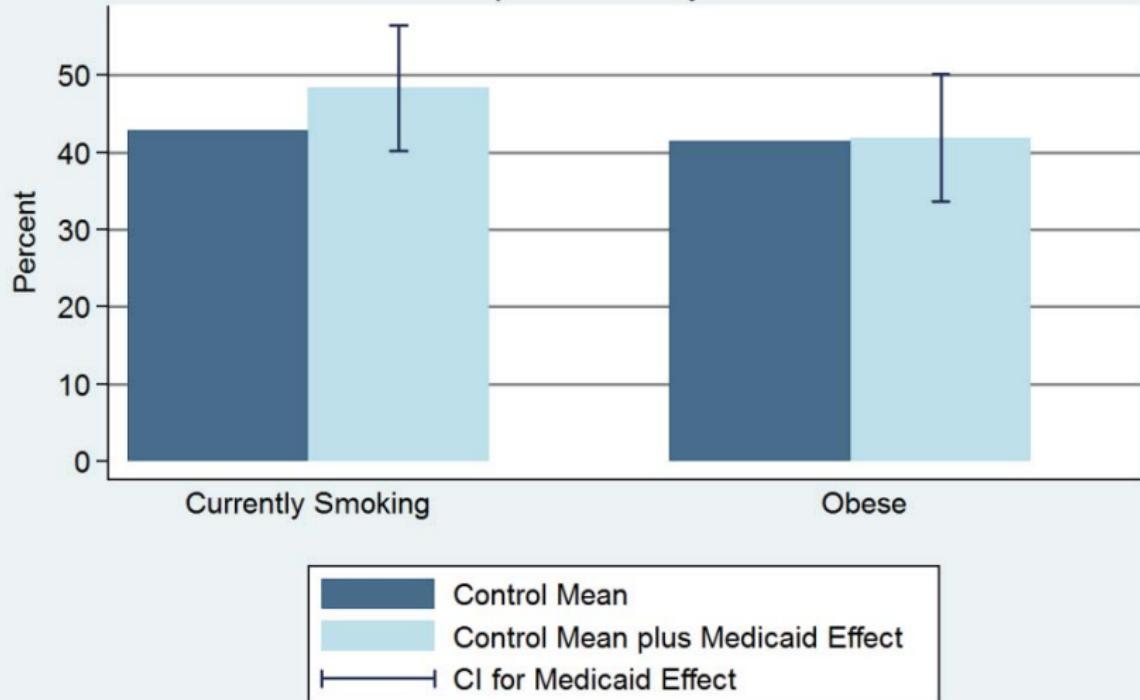
Framingham Risk Score gives the 10 year predicted risk of cardiovascular disease.

Results on specific conditions

- Large reductions in depression
 - Increases in diagnoses and medication
 - In-person estimate of –9 percentage points in being depressed
- Glycated hemoglobin
 - Increases in diagnosis and medication
 - No significant effect on HbA1c; wide confidence intervals
- Blood pressure and cholesterol
 - No significant effects on diagnosis or medication
 - No significant effects on outcomes
- Framingham risk score
 - No significant effect (in general or sub-populations)

Smoking and Obesity

Inperson Survey Data



Summary

- One to two years after expanded access to Medicaid:
 - Increases in health care use and associated costs
 - Increases in compliance with recommended preventive care
 - Improvements in quality and access
 - Reductions in financial strain
 - Improvements in self-reported health
 - Improvements in depression
 - No significant change in specific physical measures
- Sense of the relative magnitude of the effects
 - Use and access, financial benefits, general health, depression
 - Physical measures of specific chronic conditions

Extrapolation to Obamacare (ACA) Expansion

- Context quite relevant for health care reform:
 - States can choose to cover a similar population in planned 2014 Medicaid expansions (up to 138% of federal poverty line)
- But important caveats to bear in mind
 - Oregon and Portland vs. US generally
 - Voluntary enrollment vs. mandate
 - Partial vs. general equilibrium effects
 - Short-run (1-2 years) vs. medium or long run

Updating Priors based on Study's Findings

- “Medicaid is worthless or worse than no insurance”
 - Studies found increases in utilization and perceived access and quality
 - Reductions in financial strain, improvement in self-reported health
 - Improvement in depression
 - Can reject large declines in several physical measures
- “Health insurance expansion saves money”
 - In short run, studies showed increases in utilization and cost and no change in ED use
 - Increases in preventive care, improvements in self-reported health, improvements in depression

Conclusion

- Effects of expanding Medicaid likely to be manifold
 - Hard to establish with observational data and often misleading
- Expanding Medicaid generates both costs and benefits
 - Increased spending
 - Measurably improves *some* aspects of health but not others
 - Important caveats about generalizability
 - Weighing them depends on policy priorities
- Further research on alternative policies needed
 - Many steps in pathway between insurance and outcome
 - Role for innovation in insurance coverage
 - Complements to health care (e.g., social determinants)

Topics covered

- ① Review fixed effects regression models
- ② Differences-in-differences basics: Card and Krueger (1994)
- ③ Regression differences-in-differences
- ④ Synthetic control: Abadie, Diamond and Hainmueller (2010)
- ⑤ Combining differences-in-differences with IV: Waldinger (2010)

Panel Methods

- Panel data: we observe the same units (individuals, firms, countries, schools, etc.) over several time periods
- Often our outcome variable depends on unobserved factors which are also correlated with our explanatory variable of interest
- If these omitted variables are constant over time, we can use panel data estimators to consistently estimate the effect of our explanatory variable
- Main estimators for panel data:
 - Pooled OLS
 - Fixed effects estimator
 - Random effects estimator

Panel setup

- Let y and $x \equiv (x_1, x_2, \dots, x_k)$ be observable random variables and c be an unobservable random variable
- We are interested in the partial effects of variable x_j in the population regression function

$$E[y|x_1, x_2, \dots, x_k, c]$$

- We observe a sample of $i = 1, 2, \dots, N$ cross-sectional units for $t = 1, 2, \dots, T$ time periods (a balanced panel)
 - For each unit i , we denote the observable variables for all time periods as $\{(y_{it}, x_{it}) : t = 1, 2, \dots, T\}$
 - $x_{it} \equiv (x_{it1}, x_{it2}, \dots, x_{itk})$ is a $1 \times K$ vector
- Typically assume that cross-sectional units are i.i.d. draws from the population: $\{y_i, x_i, c_i\}_{i=1}^N \sim i.i.d.$ (cross-sectional independence)
 - $y_i \equiv (y_{i1}, y_{i2}, \dots, y_{iT})'$ and $x_i \equiv (x_{i1}, x_{i2}, \dots, x_{iT})$
 - Consider asymptotic properties with T fixed and $N \rightarrow \infty$

Panel setup

Single unit:

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{it} \\ \vdots \\ y_{iT} \end{pmatrix}_{T \times 1} \quad X_i = \begin{pmatrix} X_{i,1,1} & X_{i,1,2} & X_{i,1,j} & \dots & X_{i,1,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,t,1} & X_{i,t,2} & X_{i,t,j} & \dots & X_{i,t,K} \\ \vdots & \vdots & \vdots & & \vdots \\ X_{i,T,1} & X_{i,T,2} & X_{i,T,j} & \dots & X_{i,T,K} \end{pmatrix}_{T \times K}$$

Panel with all units:

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_N \end{pmatrix}_{NT \times 1} \quad X = \begin{pmatrix} X_1 \\ \vdots \\ X_i \\ \vdots \\ X_N \end{pmatrix}_{NT \times K}$$

Unobserved effects model: Farm output

- For a randomly drawn cross-sectional unit i , the model is given by

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- y_{it} : output of farm i in year t
- x_{it} : $1 \times K$ vector of variable inputs for farm i in year t , such as labor, fertilizer, etc. plus an intercept
- β : $K \times 1$ vector of marginal effects of variable inputs
- c_i : sum of all time-invariant inputs known to farmer i (but unobserved for the researcher), e.g., soil quality, managerial ability, etc.
 - often called the unobserved effect, unobserved heterogeneity, etc
- ε_{it} : time-varying unobserved inputs, such as rainfall, unknown to the farmer at the time the decision on the variable inputs x_{it} is made
 - often called the idiosyncratic error
- What happens when we regress y_{it} on x_{it} ?

Pooled OLS

- When we ignore the panel structure and regress y_{it} on x_{it} we get

$$y_{it} = x_{it}\beta + v_{it}; \quad t = 1, 2, \dots, T$$

with composite error $v_{it} \equiv c_i + \varepsilon_{it}$

- Main assumption to obtain consistent estimates for β is:
 - $E[v_{it}|x_{i1}, x_{i2}, \dots, x_{iT}] = E[v_{it}|x_{it}] = 0$ for $t = 1, 2, \dots, T$
 - x_{it} are strictly exogenous: the composite error v_{it} in each time period is uncorrelated with the past, current and future regressors
 - But: labour input x_{it} likely depends on soil quality c_i and so we have omitted variable bias and $\hat{\beta}$ is not consistent
 - No correlation between x_{it} and v_{it} implies no correlation between unobserved effect c_i and x_{it} for all t
 - Violations are common: whenever we omit a time-constant variable that is correlated with the regressors (heterogeneity bias)
 - Additional problem: v_{it} are serially correlated for same i since c_i is present in each t and thus pooled OLS standard errors are invalid

Unobserved effects model: program evaluation

- Program evaluation model:

$$y_{it} = \text{prog}_{it}\beta + c_i + \varepsilon_{it}; t = 1, 2, \dots, T$$

- y_{it} : log wage of individual i in year t
- prog_{it} : indicator coded 1 if individual i participants in training program at t and 0 otherwise
- β : effect of program
- c_i : sum of all time-invariant unobserved characteristics that affect wages, such as ability, etc.
- What happens when we regress y_{it} on prog_{it} ? $\hat{\beta}$ not consistent since prog_{it} is likely correlated with c_i (e.g., ability)
- Always ask: is there a time-constant unobserved variable (c_i) that is correlated with the regressors? If yes, then pooled OLS is problematic

Fixed effect regression

- Our unobserved effects model is:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; t = 1, 2, \dots, T$$

- If we have data on multiple time periods, we can think of c_i as **fixed effects** or “nuisance parameters” to be estimated
- OLS estimation with fixed effects yields

$$(\hat{\beta}, \hat{c}_1, \dots, \hat{c}_N) = \underset{b, m_1, \dots, m_N}{\operatorname{argmin}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it}b - m_i)^2$$

this amounts to including N farm dummies in regression of y_{it} on x_{it}

Derivation: fixed effects regression

$$(\hat{\beta}, \hat{c}_1, \dots, \hat{c}_N) = \underset{b, m_1, \dots, m_N}{\operatorname{argmin}} \sum_{i=1}^N \sum_{t=1}^T (y_{it} - x_{it}b - m_i)^2$$

The first-order conditions (FOC) for this minimization problem are:

$$\sum_{i=1}^N \sum_{t=1}^T x'_{it} (y_{it} - x_{it}\hat{\beta} - \hat{c}_i) = 0$$

and

$$\sum_{t=1}^T (y_{it} - x_{it}\hat{\beta} - \hat{c}_i) = 0$$

for $i = 1, \dots, N$.

Derivation: fixed effects regression

Therefore, for $i = 1, \dots, N$,

$$\hat{c}_i = \frac{1}{T} \sum_{t=1}^T (y_{it} - x_{it}\hat{\beta}) = \bar{y}_i - \bar{x}_i\hat{\beta},$$

where

$$\bar{x}_i \equiv \frac{1}{T} \sum_{t=1}^T x_{it}; \bar{y}_i \equiv \frac{1}{T} \sum_{t=1}^T y_{it}$$

Plug this result into the first FOC to obtain:

$$\hat{\beta} = \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)'(x_{it} - \bar{x}_i) \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T (x_{it} - \bar{x}_i)'(y_{it} - \bar{y}_i) \right)$$

$$\hat{\beta} = \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}' \ddot{x}_{it} \right)^{-1} \left(\sum_{i=1}^N \sum_{t=1}^T \ddot{x}_{it}' \ddot{y}_{it} \right)$$

with time-demeaned variables $\ddot{x}_{it} \equiv x_{it} - \bar{x}$, $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$

Fixed effects regression

Running a regression with the time-demeaned variables $\ddot{y}_{it} \equiv y_{it} - \bar{y}_i$ and $\ddot{x}_{it} \equiv x_{it} - \bar{x}$ is numerically equivalent to a regression of y_{it} on x_{it} and unit specific dummy variables.

Even better, the regression with the time demeaned variables is consistent for β even when $\text{Cov}[x_{it}, c_i] \neq 0$ because time-demeaning eliminates the unobserved effects

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}$$

$$\bar{y}_i = \bar{x}_i\beta + c_i + \bar{\varepsilon}_i$$

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x})\beta + (c_i - \bar{c}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

$$\ddot{y}_{it} = \ddot{x}_{it}\beta + \ddot{\varepsilon}_{it}$$

Fixed effects regression: main results

- Identification assumptions:

- $E[\varepsilon_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, c_i] = 0; t = 1, 2, \dots, T$
 - regressors are strictly exogenous conditional on the unobserved effect
 - allows x_{it} to be arbitrarily related to c_i
- $\text{rank}\left(\sum_{t=1}^T E[\ddot{x}'_{it} \ddot{x}_{it}]\right) = K$
 - regressors vary over time for at least some i and not collinear

- Fixed effects estimator

- Demean and regress \ddot{y}_{it} on \ddot{x}_{it} (need to correct degrees of freedom)
- Regress y_{it} on x_{it} and unit dummies (dummy variable regression)
- Regress y_{it} on x_{it} with canned fixed effects routine
 - STATA: `xtreg y x, fe i(PanelID)`

- Properties (under assumptions 1-2):

- $\hat{\beta}_{FE}$ is consistent: $\underset{N \rightarrow \infty}{\text{plim}} \hat{\beta}_{FE, N} = \beta$
- $\hat{\beta}_{FE}$ is unbiased conditional on \mathbf{X}

Fixed effects regression: main issues

- Inference:
 - Standard errors have to be “clustered” by panel unit (e.g., farm) to allow correlation in the ε_{it} ’s for the same i .
 - STATA: `xtreg , fe i(PanelID) cluster(PanelID)`
 - Yields valid inference as long as number of clusters is reasonably large
- Typically we care about β , but unit fixed effects c_i could be of interest
 - \hat{c}_i from dummy variable regression is unbiased but not consistent for c_i (based on fixed T and $N \rightarrow \infty$)
 - `xtreg , fe` routine demeans the data before running the regression and therefore does not estimate \hat{c}_i
 - intercept shows average \hat{c}_i across units
 - we can recover \hat{c}_i using $\hat{c}_i = \bar{y}_i - \bar{x}_i \hat{\beta}$
 - `predict c_i, u`

Example: Direct Democracy and Naturalizations

- Do minorities fare worse under direct democracy than under representative democracy?
- Hainmueller and Hangartner (2012) examine data on naturalization requests of immigrants in Switzerland, where municipalities vote on naturalization applications in:
 - referendums (direct democracy)
 - elected municipality councils (representative democracy)
- Annual panel data from 1,400 municipalities for the 1991-2009 period
 - y_{it} : naturalization rate =
$$\frac{\text{no. naturalizations}_{it}}{\text{eligible foreign population}_{i,t-1}}$$
 - x_{it} : 1 if municipality used representative democracy, 0 if municipality used direct democracy in year t

Naturalization Panel Data

```
. des muniID muni_name year nat_rate repdem
```

variable	name	storage	display	value	label	variable	label
	muniID	float	%8.0g		municipality code		
	muni_name	str43	%43s		municipality name		
	year	float	%ty		year		
	nat_rate	float	%9.0g		naturalization rate (percent)		
	repdem	float	%9.0g		1 representative democracy, 0 direct		

Panel Data Long Format

```
. list muniID muni_name year nat_rate repdem in 31/40
```

	muniID	muni_name	year	nat_rate	repdem
31.	2	Affoltern A.A.	2002	4.638365	0
32.	2	Affoltern A.A.	2003	4.844814	0
33.	2	Affoltern A.A.	2004	5.621302	0
34.	2	Affoltern A.A.	2005	4.387827	0
35.	2	Affoltern A.A.	2006	8.115358	1
36.	2	Affoltern A.A.	2007	7.067371	1
37.	2	Affoltern A.A.	2008	8.977719	1
38.	2	Affoltern A.A.	2009	6.119704	1
39.	3	Bonstetten	1991	.83333334	0
40.	3	Bonstetten	1992	.8403362	0

Pooled OLS

```
. reg nat_rate repdem , cl(muniID)
```

Linear regression

Number of obs = 4655
F(1, 244) = 130.04
Prob > F = 0.0000
R-squared = 0.0748
Root MSE = 3.98

(Std. Err. adjusted for 245 clusters in muniID)

nat_rate	Robust					[95% Conf. Interval]
	Coef.	Std. Err.	t	P> t		
repdem	2.503318	.2195202	11.40	0.000	2.070921	2.935714
_cons	2.222683	.10088	22.03	0.000	2.023976	2.421389

Decompose within and between variation

```
. tsset muniID year, yearly
  panel variable: muniID (strongly balanced)
  time variable: year, 1991 to 2009
  delta: 1 year

. xtsum nat_rate
```

Variable	Mean	Std. Dev.	Min	Max	Observations
nat_rate overall	2.938992	4.137305	0	24.13793	N = 4655
between		1.622939	0	7.567746	n = 245
within		3.807039	-3.711323	24.80134	T = 19

Time-demeaning for fixed effects: $y_{it} \rightarrow \ddot{y}_{it}$

```
. * get municipality means
. egen means_nat_rate = mean(nat_rate) , by(muniID)

. * compute deviations from means
. gen dm_nat_rate = nat_rate - means_nat_rate

. list muniID muni_name year nat_rate means_nat_rate dm_nat_rate in 20/40 ,ab(20)
```

	muniID	muni_name	year	nat_rate	means_nat_rate	dm_nat_rate
20.	2	Affoltern A.A.	1991	.2173913	3.595932	-3.37854
21.	2	Affoltern A.A.	1992	.9473684	3.595932	-2.648563
22.	2	Affoltern A.A.	1993	1.04712	3.595932	-2.548811
23.	2	Affoltern A.A.	1994	.8342023	3.595932	-2.761729
24.	2	Affoltern A.A.	1995	2.002002	3.595932	-1.59393
25.	2	Affoltern A.A.	1996	1.7769	3.595932	-1.819031
26.	2	Affoltern A.A.	1997	1.862745	3.595932	-1.733186
27.	2	Affoltern A.A.	1998	2.054155	3.595932	-1.541776
28.	2	Affoltern A.A.	1999	2.402135	3.595932	-1.193796

Fixed effects regression with demeaned data

```
. egen means_repdem = mean(repdem) , by(muniID)

. gen dm_repdem = repdem - means_repdem

.

. * regression with demeaned data
. reg dm_nat_rate dm_repdem , cl(muniID)
```

Linear regression

		Number of obs = 4655
		F(1, 244) = 265.18
		Prob > F = 0.0000
		R-squared = 0.1052
		Root MSE = 3.6017
		(Std. Err. adjusted for 245 clusters in muniID)

dm_nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
dm_repdem	3.0228	.1856244	16.28	0.000	2.657169	3.388431
_cons	6.65e-10	5.81e-09	0.11	0.909	-1.08e-08	1.21e-08

Fixed effects regression with canned routine

```
. xtreg nat_rate repdem , fe cl(muniID) i(muniID)

Fixed-effects (within) regression                         Number of obs      =     4655
Group variable: muniID                                Number of groups   =      245

R-sq:  within  =  0.1052                               Obs per group: min =       19
          between =  0.0005                             avg =      19.0
          overall =  0.0748                             max =       19

                                                F(1,244)           =    265.18
corr(u_i, Xb)  = -0.1373                           Prob > F        =  0.0000
```

(Std. Err. adjusted for 245 clusters in muniID)

nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
repdem	3.0228	.1856244	16.28	0.000	2.657169	3.388431
_cons	2.074036	.0531153	39.05	0.000	1.969413	2.178659
<hr/>						
sigma_u	1.7129711					
sigma_e	3.69998					
rho	.17650677	(fraction of variance due to u_i)				

Fixed effects regression with dummies

```
. reg nat_rate repdem i.muniID, cl(muniID)
```

Linear regression

Number of obs = 4655
F(0, 244) = .
Prob > F = .
R-squared = 0.2423
Root MSE = 3.7

(Std. Err. adjusted for 245 clusters in muniID)

nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
repdem	3.0228	.1906916	15.85	0.000	2.647188	3.398412
muniID						
2	1.367365	5.17e-14	2.6e+13	0.000	1.367365	1.367365
3	1.292252	5.17e-14	2.5e+13	0.000	1.292252	1.292252
9	1.284652	5.17e-14	2.5e+13	0.000	1.284652	1.284652
10	1.271783	5.17e-14	2.5e+13	0.000	1.271783	1.271783
13	.3265469	5.17e-14	6.3e+12	0.000	.3265469	.3265469

Applying fixed effects

- We can use fixed effects for other data structures to restrict comparisons to within unit variation
 - Matched pairs
 - Twin fixed effects to control for unobserved effects of family background
 - Cluster fixed effects in hierarchical data
 - School fixed effects to control for unobserved effects of school

Problems that even fixed effects do not solve

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, 2, \dots, T$$

- Where y_{it} is murder rate and x_{it} is police spending per capita
- What happens when we regress y on x and city fixed effects?
 - $\hat{\beta}_{FE}$ inconsistent unless strict exogeneity conditional on c_i holds
 - $E[\varepsilon_{it}|x_{i1}, x_{i2}, \dots, x_{iT}, c_i] = 0; t = 1, 2, \dots, T$
 - implies ε_{it} uncorrelated with past, current and future regressors
- Most common violations
 - ① Time-varying omitted variables
 - Economic boom leads to more police spending and less murders
 - Can include time-varying controls, but avoid post-treatment bias (i.e., collider)
 - ② Simultaneity
 - if city adjusts police based on past murder rate, then spending_{t+1} is correlated with ε_t (since higher ε_t leads to higher murder rate at t)
 - strictly exogenous x cannot react to what happens to y in the past or the future!
- Fixed effects do not obviate need for good research design!

Random Effects

- Reconsider our unobserved effects model:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- Cannot use the fixed effects regression to estimate effects of time-constant regressors in x_{it} (eg., soil quality, farm location, etc.)
 - Since fixed effect estimator allows c_i to be correlated with x_{it} , we cannot distinguish the effects of time-invariant regressors from the time-invariant unobserved effect c_i
- Need orthogonality assumption: $\text{Cov}[x_{it}, c_i] = 0$;
 $t = 1, \dots, T$
 - Strong assumption: Unobserved effects c_i are uncorrelated with each explanatory variable in x_{it} in each time period
 - For example if we include soil quality in x_{it} we have to assume it is uncorrelated with all other time-invariant inputs

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \dots, T$$

- 1 $E[\varepsilon_{it}|x_i, c_i] = 0; \quad t = 1, 2, \dots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
- 2 $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
- 3 rank $E[X_i'\Omega X_i] = K$: no collinearity among regressors
 - $\Omega = E[v_i v_i']$: the variance matrix of the composite error
 $v_{it} = c_i + \varepsilon_{it}$
- 4 We typically also assume that Ω takes a special form:
 - $E[\varepsilon_i \varepsilon_i' | x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
 - $E[c_i^2 | x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \dots, T$$

- ① $E[\varepsilon_{it}|x_i, c_i] = 0; \quad t = 1, 2, \dots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
- ② $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
- ③ $\text{rank } E[X_i' \Omega X_i] = K$: no collinearity among regressors
 - $\Omega = E[v_i v_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
- ④ We typically also assume that Ω takes a special form:
 - $E[\varepsilon_i \varepsilon_i' | x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
 - $E[c_i^2 | x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic

Assumption 4 implies $\Omega = E[v_i v_i' | x_i] = \begin{pmatrix} \sigma_c^2 + \sigma_\varepsilon^2 & \sigma_c^2 & \dots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_\varepsilon^2 & \dots & \vdots \\ \vdots & & \ddots & \sigma_c^2 \\ \sigma_c^2 & & \sigma_c^2 + \sigma_\varepsilon^2 & \end{pmatrix}_{T \times T}$

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \dots, T$$

- ① $E[\varepsilon_{it}|x_i, c_i] = 0; \quad t = 1, 2, \dots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
- ② $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
- ③ rank $E[X_i' \Omega X_i] = K$: no collinearity among regressors
 - $\Omega = E[v_i v_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
- ④ We typically also assume that Ω takes a special form:
 - $E[\varepsilon_i \varepsilon_i' | x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
 - $E[c_i^2 | x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic
- Given assumptions 1-3, pooled OLS is consistent, since composite error v_{it} is uncorrelated with x_{it} for all t
- However, pooled OLS ignores the serial correlation in v_{it}

Random effects assumptions

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}; \quad t = 1, \dots, T$$

- ① $E[\varepsilon_{it}|x_i, c_i] = 0; \quad t = 1, 2, \dots, T$: explanatory variables are strictly exogenous conditional on the unobserved effect
- ② $E[c_i|x_i] = E[c_i] = 0$: unobserved effects c_i are uncorrelated with regressors
- ③ rank $E[X_i' \Omega X_i] = K$: no collinearity among regressors
 - $\Omega = E[v_i v_i']$: the variance matrix of the composite error $v_{it} = c_i + \varepsilon_{it}$
- ④ We typically also assume that Ω takes a special form:
 - $E[\varepsilon_i \varepsilon_i' | x_i] = \sigma_\varepsilon^2 \mathbf{I}_T$: idiosyncratic errors are homoskedastic for all t and serially uncorrelated
 - $E[c_i^2 | x_i] = \sigma_c^2$: unobserved effect c_i is homoscedastic
- Random effects estimator $\widehat{\beta}_{RE}$ exploits this serial correlation in a generalized least squares (GLS) framework
 - $\widehat{\beta}_{RE}$ is consistent under assumption 1-3: $\underset{N \rightarrow \infty}{plim} \widehat{\beta}_{RE, N} = \beta$
 - $\widehat{\beta}_{RE}$ is asymptotically efficient given assumption 4 (in the class of estimators consistent under $E[v_i | x_i] = 0$)

Random effects estimator

- Consider the transformation parameter

$$\lambda = 1 - \left(\frac{\sigma_{\varepsilon}^2}{\sigma_{\varepsilon}^2 + T\sigma_c^2} \right)^{\frac{1}{2}} \text{ with } 0 \leq \lambda \leq 1$$

- $\sigma_{\varepsilon}^2 = \text{Var}[\varepsilon_{it}]$: variance of idiosyncratic error
- $\sigma_c^2 = \text{Var}(c_i)$: Variance of unobserved effect
- $\hat{\beta}_{RE}$ is equivalent to pooled OLS on:

$$\begin{aligned} y_{it} - \bar{y}_i &= (x_{it} - \lambda \bar{x}_i) \beta + (v_{it} - \lambda \bar{v}_i), \forall i, t \\ \tilde{y}_{it} &= \tilde{x}_{it} \beta + \tilde{v}_{it} \end{aligned}$$

- As $\lambda \rightarrow 1$, $\hat{\beta}_{RE} \rightarrow \hat{\beta}_{FE}$
- As $\lambda \rightarrow 0$, $\hat{\beta}_{RE} \rightarrow \hat{\beta}_{\text{Pooled OLS}}$
 - $\lambda \rightarrow 1$ as $T \rightarrow \infty$ or if variance of c_i is large relative to variance of ε_{it}
- λ can be estimated from data $\hat{\lambda} = 1 - (\hat{\sigma}_{\varepsilon}^2 / (\hat{\sigma}_{\varepsilon}^2 + T\hat{\sigma}_c^2))^{\frac{1}{2}}$
- Usually wise to cluster the standard errors since assumption 4 is strong

Random effects regression

```
. xtreg nat_rate repdem , re cl(muniID) i(muniID)

Random-effects GLS regression
Group variable: muniID

R-sq:  within = 0.1052
      between = 0.0005
      overall = 0.0748

Number of obs      =      4655
Number of groups  =      245

Obs per group: min =         19
                avg =      19.0
                max =         19

Wald chi2(1)      =     227.99
Prob > chi2       =     0.0000

corr(u_i, X)      = 0 (assumed)

(Std. Err. adjusted for 245 clusters in muniID)
```

nat_rate	Robust					[95% Conf. Interval]
	Coef.	Std. Err.	z	P> z		
repdem	2.859397	.1893742	15.10	0.000	2.48823	3.230564
_cons	2.120793	.0972959	21.80	0.000	1.930096	2.311489
sigma_u	1.3866768					
sigma_e	3.69998					
rho	.1231606	(fraction of variance due to u_i)				

Summary: Fixed effects, random effects, Pooled OLS

- Main assumptions
 - ① Regressors are strictly exogenous conditional on the time-invariant unobserved effects
 - ② Regressors are uncorrelated with the time-invariant unobserved effects
- Results
 - Fixed effects estimator is consistent given assumption 1, but rules out time-invariant regressors
 - Random effects estimators and pooled OLS are consistent under assumptions 1-2, and allow for time-invariant regressors
 - Given homoskedasticity assumptions (random effects assumption 4), the random effects estimator is asymptotically efficient
- Assumption 2 is strong so fixed effects are typically more credible
 - Often the main reason for using panel data is to rule out all time-invariant unobserved confounders!

Hausman test

	$\widehat{\beta}_{RE}$	$\widehat{\beta}_{FE}$
$H_0 : \text{Cov}[x_{it}, c_i] = 0$	Consistent and efficient	Consistent
$H_1 : \text{Cov}[x_{it}, c_i] \neq 0$	Inconsistent	Consistent

Then,

- Under H_0 , $\widehat{\beta}_{RE} - \widehat{\beta}_{FE}$ should be close to zero
- Under H_1 , $\widehat{\beta}_{RE} - \widehat{\beta}_{FE}$ should be different from zero
- It can be shown that in large samples, under H_0 , the test statistic

$$(\widehat{\beta}_{FE} - \widehat{\beta}_{RE})'(\widehat{\text{Var}}[\beta_{FE}] - \widehat{\text{Var}}[\beta_{RE}])^{-1}(\widehat{\beta}_{FE} - \widehat{\beta}_{RE}) \xrightarrow{d} \chi_k^2$$

where k is the number of time-varying regressors.

- We may reject the null hypothesis of “random effects” and stick with the less efficient, but consistent fixed effects specification

Random effects regression

```
. quietly: xtreg nat_rate repdem , fe i(muniID)  
. estimates store FE  
. quietly: xtreg nat_rate repdem , re i(muniID)  
. estimates store RE  
. hausman FE RE
```

	Coefficients			
	(b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
repdem	3.0228	2.859397	.1634027	.0304517

b = consistent under H_0 and H_a ; obtained from xtreg
B = inconsistent under H_a , efficient under H_0 ; obtained from xtreg

Test: H_0 : difference in coefficients not systematic

$$\begin{aligned} \text{chi2}(1) &= (b-B)' [(V_b-V_B)^{-1}] (b-B) \\ &= 28.79 \end{aligned}$$

Hausman test

- Hausman test does not test if the fixed effect model is correct; the test assumes that the fixed effects estimator is consistent!
- Conventional Hausman test assumes homoskedastic model and does not allow for clustering
- There are Haumsman like tests that allow for clustered standard errors

```
. * hausman test with clustering
. quietly: xtreg nat_rate repdem , re i(muniID) cl(muniID)

. xtoverid

Test of overidentifying restrictions: fixed vs random effects
Cross-section time-series model: xtreg re robust cluster(muniID)
Sargan-Hansen statistic  26.560  Chi-sq(1)      P-value = 0.0000
```

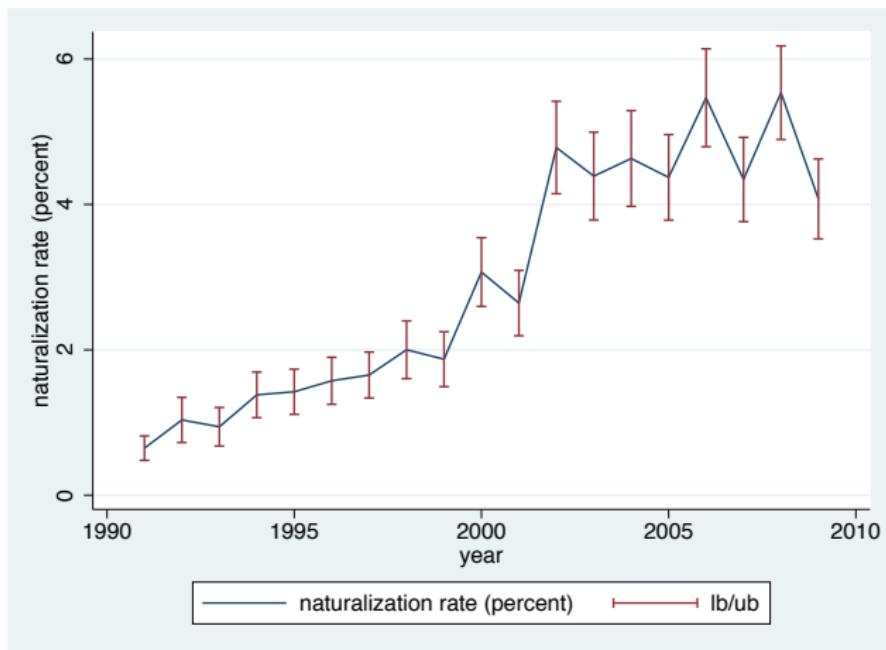
Adding Time Effects

- Reconsider our unobserved effects model:

$$y_{it} = x_{it}\beta + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

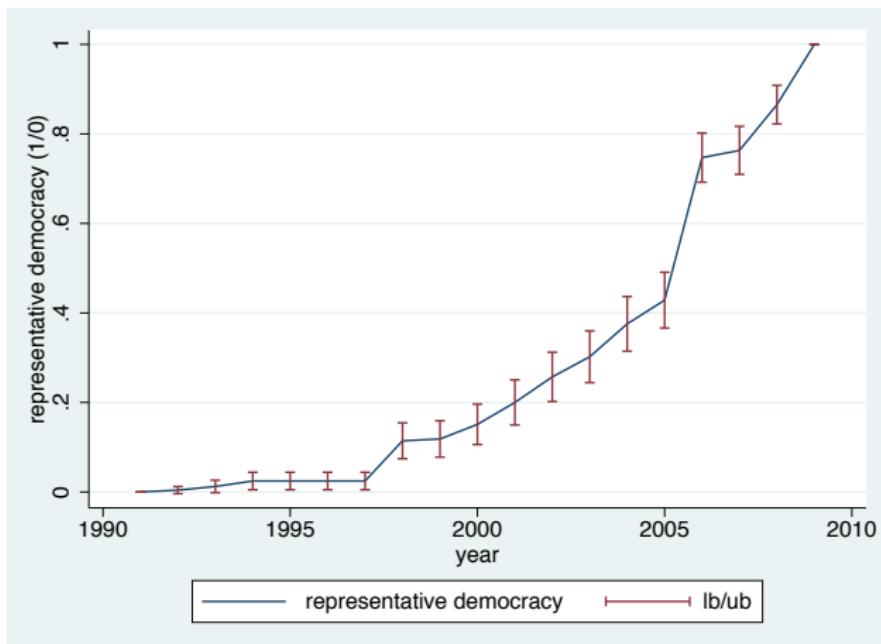
- Fixed effects assumption: $E[\varepsilon_{it}|x_i, c_i] = 0; t = 1, 2, \dots, T$: regressors are strictly exogenous conditional on the unobserved effect
- Typical violation: Common shocks that affect all units in the same way and are correlated with x_{it}
 - Trends in farming technology or climate affect productivity
 - Trends in immigration inflows affect naturalization rates
- We can allow for such common shocks by including time effects into the model

Random effects regression



xtgraph nat_rate

Random effects regression



xtgraph repdem

Fixed effects: adding time effects

- Linear time trend:

$$y_{it} = x_{it}\beta + c_i + t + \varepsilon_{it}; \quad t = 1, 2, \dots, T$$

- Linear time trend common to all units

- Time fixed effects:

$$y_{it} = x_{it}\beta + c_i + t_t + \varepsilon_{it}; \quad t = 1, 2, \dots, T$$

- Common shock in each time period
- Generalized difference-in-difference model

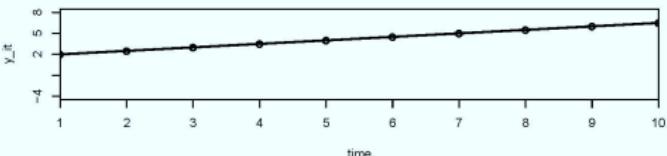
- Unit specific linear time trends:

$$y_{it} = x_{it}\beta + c_i + g_i \cdot t + t_t + \varepsilon_{it}; \quad t = 1, 2, \dots, T$$

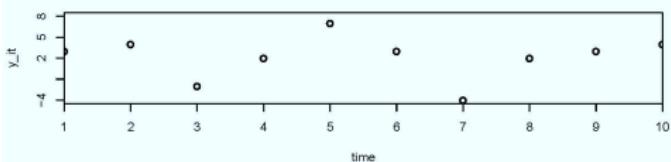
- Linear time trends that vary by unit

Modeling time effects

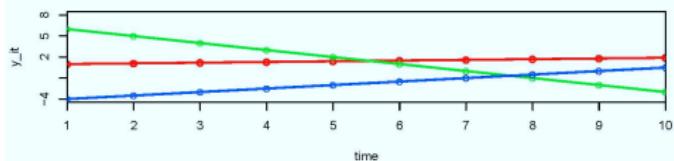
common linear time trend (t)



time fixed effects (t_{-i}) t_{-t}



unit specific linear time trends ($g_{i,t}$)



Fixed effects: adding time effects

```
. egen time = group(year)  
  
. list muniID muni_name year time in 20/40 ,ab(20)
```

	muniID	muni_name	year	time
20.	2	Affoltern A.A.	1991	1
21.	2	Affoltern A.A.	1992	2
22.	2	Affoltern A.A.	1993	3
23.	2	Affoltern A.A.	1994	4
24.	2	Affoltern A.A.	1995	5
25.	2	Affoltern A.A.	1996	6
26.	2	Affoltern A.A.	1997	7
27.	2	Affoltern A.A.	1998	8
28.	2	Affoltern A.A.	1999	9
29.	2	Affoltern A.A.	2000	10
30.	2	Affoltern A.A.	2001	11

Fixed effects: linear time trend

```
. xtreg nat_rate repdem time , fe cl(muniID) i(muniID)

Fixed-effects (within) regression                         Number of obs     =      4655
Group variable: muniID                                Number of groups  =       245

R-sq:  within  = 0.1604                                Obs per group: min =         19
          between = 0.0005                                avg =      19.0
          overall = 0.1350                                max =         19

                                                F(2,244)     =     247.57
corr(u_i, Xb)  = -0.0079                                Prob > F    =     0.0000

                                                (Std. Err. adjusted for 245 clusters in muniID)
```

nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
repdem	.8247928	.2590615	3.18	0.002	.3145106	1.335075
time	.2313692	.0171752	13.47	0.000	.1975386	.2651997
_cons	.3892908	.1309232	2.97	0.003	.1314069	.6471747
sigma_u	1.6271657					
sigma_e	3.584409					
rho	.17086519	(fraction of variance due to u_i)				

Fixed effects: year fixed effects

```
. xtreg nat_rate repdem i.time , fe cl(muniID) i(muniID)

Fixed-effects (within) regression                         Number of obs      =      4655
Group variable: muniID                                Number of groups    =       245

R-sq:  within  = 0.1885                                Obs per group: min =        19
          between = 0.0005                               avg =      19.0
          overall = 0.1575                               max =        19

                                                F(19,244)      =      31.48
corr(u_i, Xb)  = -0.0168                                Prob > F        =     0.0000
```

(Std. Err. adjusted for 245 clusters in muniID)

nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
repdem	1.203658	.3031499	3.97	0.000	.6065335	1.800783
time						
2	.3829173	.1723225	2.22	0.027	.0434879	.7223468
3	.2789777	.1514124	1.84	0.067	-.0192644	.5772198
4	.7034078	.167466	4.20	0.000	.3735443	1.033271

Fixed effects: unit specific time trends

```
. xtreg nat_rate repdem muniID#c.time i.time , fe cl(muniID) i(muniID)
note: 19.time omitted because of collinearity
```

```
Fixed-effects (within) regression                         Number of obs      =      4655
Group variable: muniID                               Number of groups   =       245
R-sq:  within  =  0.2650                               Obs per group: min =        19
          between =  0.5185                               avg =      19.0
          overall =  0.2864                               max =        19
                                                F(18,244)      =      .
corr(u_i, Xb)  = -0.3963                           Prob > F        =      .
```

(Std. Err. adjusted for 245 clusters in muniID)

nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
repdem	.9865241	.322868	3.06	0.002	.3505601	1.622488
muniID#c.time						
1	.333343	.024298	13.72	0.000	.2854823	.3812036
2	.2914274	.024298	11.99	0.000	.2435667	.339288
3	.248985	.024298	10.25	0.000	.2011244	.2968457

Unit specific time trends often eliminate “results”

TABLE 5.2.3
Estimated effects of labor regulation on the performance of firms
in Indian states

	(1)	(2)	(3)	(4)
Labor regulation (lagged)	-.186 (.064)	-.185 (.051)	-.104 (.039)	.0002 (.020)
Log development expenditure per capita		.240 (.128)	.184 (.119)	.241 (.106)
Log installed electricity capacity per capita		.089 (.061)	.082 (.054)	.023 (.033)
Log state population		.720 (.96)	0.310 (1.192)	-1.419 (2.326)
Congress majority			-.0009 (.01)	.020 (.010)
Hard left majority			-.050 (.017)	-.007 (.009)
Janata majority			.008 (.026)	-.020 (.033)
Regional majority			.006 (.009)	.026 (.023)
State-specific trends	No	No	No	Yes
Adjusted R^2	.93	.93	.94	.95

Notes: Adapted from Besley and Burgess (2004), table IV. The table reports regression DD estimates of the effects of labor regulation on productivity. The

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- Model recognizes the effect of change in x may occur with a lag
 - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
 - Consider **temporary increase** in x_{it} from level m to $m + 1$ at t which lasts only one period
 - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_t = (m + 1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_{t+1} = m\beta_0 + (m + 1)\beta_1 + m\beta_2 + c_i$
 - $y_{t+2} = m\beta_0 + m\beta_1 + (m + 1)\beta_2 + c_i$
 - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- Model recognizes the effect of change in x may occur with a lag
 - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
 - Consider **temporary increase** in x_{it} from level m to $m+1$ at t which lasts only one period
 - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
 - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
 - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $\beta_0 = y_t - y_{t-1}$ immediate change in y due to temporary one-unit increase in x (impact propensity)

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- Model recognizes the effect of change in x may occur with a lag
 - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
 - Consider **temporary increase** in x_{it} from level m to $m+1$ at t which lasts only one period
 - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_t = (m+1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_{t+1} = m\beta_0 + (m+1)\beta_1 + m\beta_2 + c_i$
 - $y_{t+2} = m\beta_0 + m\beta_1 + (m+1)\beta_2 + c_i$
 - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $\beta_1 = y_{t+1} - y_t$ change in y one period after temporary one-unit increase in x

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- Model recognizes the effect of change in x may occur with a lag
 - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
 - Consider **temporary increase** in x_{it} from level m to $m + 1$ at t which lasts only one period
 - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_t = (m + 1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_{t+1} = m\beta_0 + (m + 1)\beta_1 + m\beta_2 + c_i$
 - $y_{t+2} = m\beta_0 + m\beta_1 + (m + 1)\beta_2 + c_i$
 - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $\beta_2 = y_{t+2} - y_{t-1}$ change in y two periods after temporary one-unit increase in x

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- Model recognizes the effect of change in x may occur with a lag
 - effect of new tax credit for children on fertility rate
- Interpretation of coefficients:
 - Consider **temporary increase** in x_{it} from level m to $m + 1$ at t which lasts only one period
 - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_t = (m + 1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_{t+1} = m\beta_0 + (m + 1)\beta_1 + m\beta_2 + c_i$
 - $y_{t+2} = m\beta_0 + m\beta_1 + (m + 1)\beta_2 + c_i$
 - $y_{t+3} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $\beta_3 = y_{t-1}$ change in y is zero three periods after temporary one-unit increase in x

Distributed Lag model

$$y_{it} = x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- Interpretation of coefficients:
 - Consider **permanent increase** in x_{it} from level m to $m + 1$ at t , i.e., ($x_s = m, s < t$ and $x_s = m + 1, s \geq t$)
 - $y_{t-1} = m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_t = (m + 1)\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_{t+1} = (m + 1)\beta_0 + (m + 1)\beta_1 + m\beta_2 + c_i$
 - $y_{t+2} = (m + 1)\beta_0 + (m + 1)\beta_1 + (m + 1)\beta_2 + c_i$
 - $y_{t+3} = (m + 1)\beta_0 + (m + 1)\beta_1 + (m + 1)\beta_2 + c_i$
 - After one period y has increased by $\beta_0 + \beta_1$, after two periods y has increased by $\beta_0 + \beta_1 + \beta_2$ and there are no further increases after two periods
 - Long-run increase in y : $\beta_0 + \beta_1 + \beta_2$ (long-run propensity)

Lagged effects of direct democracy

```
. xtreg nat_rate repdem L1.repdem L2.repdem L3.repdem i.year, fe cl(muniID) i(muniID)

Fixed-effects (within) regression
Group variable: muniID

Number of obs = 3920
Number of groups = 245

R-sq: within = 0.1536
      between = 0.0012
      overall = 0.1235

Obs per group: min = 16
               avg = 16.0
               max = 16

F(19, 244) = 21.63
Prob > F = 0.0000
```

(Std. Err. adjusted for 245 clusters in muniID)

nat_rate	Coef.	Robust				
		Std. Err.	t	P> t	[95% Conf. Interval]	
repdem						
--.	.6364802	.3593924	1.77	0.078	-.0714272	1.344388
L1.	1.201065	.4233731	2.84	0.005	.367133	2.034998
L2.	-.1648692	.4697434	-0.35	0.726	-1.090139	.7604003
L3.	-.5245206	.4109918	-1.28	0.203	-1.334065	.2850239

Long-run effect of direct democracy

```
. lincom repdem + L1.repdem + L2.repdem + L3.repdem  
( 1)  repdem + L.repdem + L2.repdem + L3.repdem = 0
```

nat_rate	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
(1)	1.294485	.4426322	2.92	0.004	.4226175 2.166353

Lags and Leads model

$$y_{it} = x_{i,t+1}\beta_{-1} + x_{it}\beta_0 + x_{i,t-1}\beta_1 + x_{i,t-2}\beta_2 + c_i + \varepsilon_{it}; \quad t = 1, 2, \dots, T$$

- Can use estimate of β_{-1} to test for anticipation effects
 - Consider **temporary increase** in x_{it} from level m to $m+1$ at t
 - $y_{t-2} = \beta_{-1}m + m\beta_0 + m\beta_1 + m\beta_2 + c_i$
 - $y_{t-1} = \beta_{-1}(m+1) + m\beta_0 + m\beta_1 + m\beta_2 + c_i$
- Anticipation effect: $\beta_{-1} = y_{t-1} - y_{t-2}$ change in y in period $t-1$, the period before the temporary one-unit increase in x
- Placebo test: if x causes y , but y does not cause x , then β_{-1} should be close to zero

Leads and Lags

```

. xtreg nat_rate F1.repdem repdem L1.repdem L2.repdem L3.repdem i.year, fe cl(muniID) i(muniID)

Fixed-effects (within) regression
Group variable: muniID

R-sq:  within = 0.1621
      between = 0.0010
      overall = 0.1269

corr(u_i, Xb) = -0.0353

Number of obs = 3675
Number of groups = 245

Obs per group: min = 15
                  avg = 15.0
                  max = 15

F(19, 244) = 20.34
Prob > F = 0.0000

(Std. Err. adjusted for 245 clusters in muniID)

```

nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
repdem						
F1.	.1707685	.3212906	0.53	0.596	-.4620886	.8036255
--.	.6975731	.4397095	1.59	0.114	-.1685376	1.563684
L1.	.8723962	.4619322	1.89	0.060	-.0374873	1.78228
L2.	.014941	.4583628	0.03	0.974	-.8879119	.9177939
L3.	-.2904252	.4108244	-0.71	0.480	-1.09964	.5187895

The Autor Test

- Let D_{it} be a binary indicator equaling 1 if unit i **switched** from control to treatment between t and $t - 1$; 0 otherwise
 - Lags: $D_{i,t-1}$: unit switched between $t - 1$ and $t - 2$
 - Leads: $D_{i,t+1}$: unit switches between $t + 1$ and t
- Include lags and leads into the fixed effects model:

$$y_{it} = D_{i,t+2}\beta_{-2} + D_{i,t+1}\beta_{-1} + D_{it}\beta_0 + D_{i,t-1}\beta_1 + D_{i,t-2}\beta_2 + c_i + \varepsilon_{it}$$

- Interpretation of coefficients:
 - Leads β_{-1}, β_{-2} , etc. test for anticipation effects
 - Switch β_0 tests for immediate effect
 - Lags β_1, β_2 , etc. test for long-run effects
 - highest lag dummy can be coded 1 for all post-switch years

Lags and Leads of Switch to Representative Democracy

```
. list muni_name year repdem switch_t sw_lag1 sw_lag2 sw_lag3 ///
>           sw_lead1 sw_lead2 sw_lead3 in 806/817
```

	muni_name	year	repdem	switch_t	sw_lag1	sw_lag2	sw_lag3	sw_lead1	sw_lead2	sw_lead3
806.	Stäfa	1998	0	0	0	0	0	0	0	0
807.	Stäfa	1999	0	0	0	0	0	0	0	0
808.	Stäfa	2000	0	0	0	0	0	0	0	0
809.	Stäfa	2001	0	0	0	0	0	0	0	0
810.	Stäfa	2002	0	0	0	0	0	0	0	1
811.	Stäfa	2003	0	0	0	0	0	0	1	0
812.	Stäfa	2004	0	0	0	0	0	1	0	0
813.	Stäfa	2005	1	1	0	0	0	0	0	0
814.	Stäfa	2006	1	0	1	0	0	0	0	0
815.	Stäfa	2007	1	0	0	1	0	0	0	0
816.	Stäfa	2008	1	0	0	0	1	0	0	0
817.	Stäfa	2009	1	0	0	0	1	0	0	0

Dynamic Effect of Switching to Representative Democracy

```
. xtreg  nat_rate sw_lag3 sw_lag2 sw_lag1 switch_t ///
>          sw_lead1 sw_lead2 sw_lead3 sw_lead4 sw_lead5 i.year, fe cluster(muniID) i(muniID)

Fixed-effects (within) regression                         Number of obs      =      4655
Group variable: muniID                               Number of groups   =       245

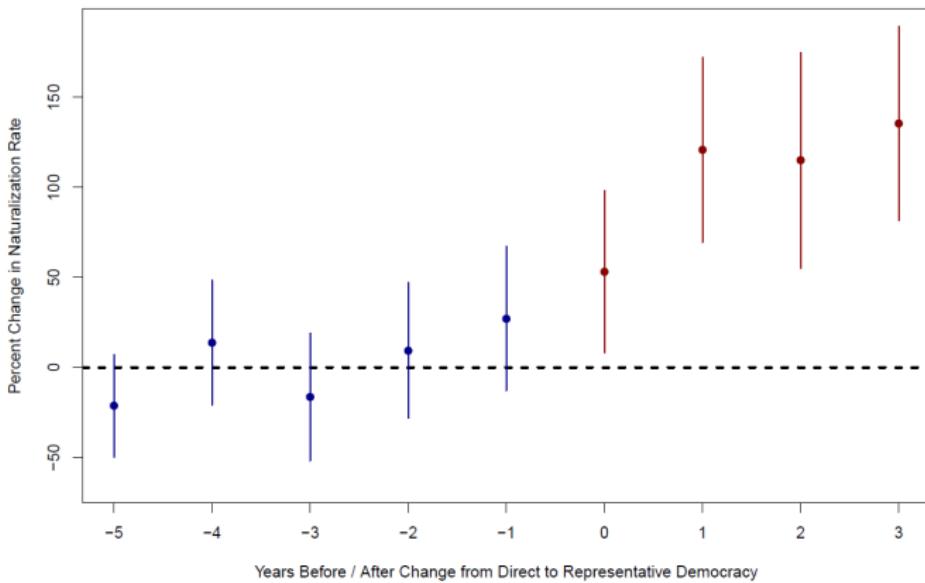
R-sq:  within  = 0.1913                               Obs per group: min =        19
          between = 0.0011                           avg =      19.0
          overall = 0.1601                           max =        19

                                                F(27,244)      =     23.76
corr(u_i, Xb)  = -0.0162                           Prob > F        =  0.0000

                                                (Std. Err. adjusted for 245 clusters in muniID)
```

nat_rate	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
sw_lag3	1.160345	.5080271	2.28	0.023	.1596665	2.161023
sw_lag2	1.743682	.5395212	3.23	0.001	.680969	2.806396
sw_lag1	1.881663	.4880099	3.86	0.000	.9204133	2.842913
switch_t	.7564792	.428627	1.76	0.079	-.0878019	1.60076
sw_lead1	.2138757	.3899881	0.55	0.584	-.5542971	.9820485
sw_lead2	.0843676	.3575292	0.24	0.814	-.61987	.7886051
sw_lead3	.1440446	.3194086	0.45	0.652	-.4851054	.7731945
sw_lead4	.0750194	.2990359	0.25	0.802	-.5140018	.6640405
sw_lead5	-.0942415	.2599789	-0.36	0.717	-.6063307	.4178477

Dynamic Effect of Switching to Representative Democracy



Lagged Dependent Variable

$$y_{it} = \alpha y_{i,t-1} + c_i + \varepsilon_{it}, \quad t = 1, 2, \dots, T$$

- y_{it} could be capital stock of firm i at time t and α the capital depreciation rate
- For simplicity, we assume that ε_{it} are uncorrelated in time (as well as across individuals)
- Note that

$$y_{i,t-1} = \alpha y_{i,t-2} + c_i + \varepsilon_{i,t-1}$$

- So we have $\text{Cov}[y_{i,t-1}, c_i] \neq 0$ and therefore we need to include fixed effects c_i into the regression
- Does this work though?

Lagged dependent variable

With $T = 3$ we have

$$y_{i3} = \alpha y_{i2} + c_i + \varepsilon_{i3}$$

$$y_{i2} = \alpha y_{i1} + c_i + \varepsilon_{i2}$$

and we can take time differences to eliminate c_i (similar to fixed effects)

$$y_{i3} - y_{i2} = \alpha(y_{i2} - y_{i1}) + (c_i - c_i) + (\varepsilon_{i3} - \varepsilon_{i2})$$

$$\Delta y_{i3} = \alpha \Delta y_{i2} + \Delta \varepsilon_{i3}$$

Since ε_{i2} affects both $\Delta y_{i2} = y_{i2} - y_{i1}$ and $\Delta \varepsilon_{i3} = \varepsilon_{i3} - \varepsilon_{i2}$ we get

$\text{Cov}[\Delta y_{i2}, \Delta \varepsilon_{i3}] \neq 0$ and thus still have endogeneity

Models with fixed effects and lagged y do not produce consistent estimators
Might use past levels y_{i1} as an instrument for Δy_{i2} , but

this requires strong assumptions (e.g., no serial correlation in ε_{it})

Cheng, Cheng and Mark Hoestra 2013. “Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine”. Journal of Human Resources, 48(3), pp. 821-854.

Summary

- Cheng and Hoekstra (2013) are interested in whether expansions to “castle doctrine statutes” at the state level increase or decrease gun violence.
- Prior to these expansions, English common law principle required “duty to retreat” before using lethal force against an assailant except when the assailant is an intruder in the home
 - The home is one’s “castle” – hence, “castle doctrine”
 - When intruders threatened the victim in the home, the duty to retreat was waived and lethal force in self-defense was allowed

Castle doctrine law explained

- In 2005, Florida passed a law that expanded self-defense protections beyond the house
 - 2000 to 2010, 21 states explicitly put ?castle doctrine? into statute, and (more importantly) extended it to places outside the home
 - In other words, 21 states removed the duty to retreat in specified circumstances
- Other changes:
 - Presumption of reasonable fear is added
 - Civil liability for those acting under the law is removed

Texas example

- Duty to retreat
 - "For purposes of subsection (a), in determining whether an actor described by subsection (e) reasonably believed that the use of force was necessary, a finder of fact may not consider whether the actor failed to retreat."
 - Also: Language stating a person is justified using deadly force against another "if a reasonable person in the actor's situation would not have retreated" is removed from the statute
- Presumption of reasonableness
 - "Except as provided in subsection (b), a person is justified in using force against another when and to the degree the actor [he] reasonably believes the force is immediately necessary to protect the actor [himself] against the other's use or attempted use of unlawful force. The actor's belief that the force was immediately necessary as described by this subsection is presumed to be reasonable if the actor . . ."
- Civil Liability
 - "A defendant who uses force or deadly force that is justified under Chapter 9 Penal code is immune from civil liability for personal injury or death that results from the defendant's use of force or deadly force, as applicable."

Economic theory

- Workers supply legal or illegal labor
 - Costs: foregone wages, risk of arrest, risk of conviction, penalties conditional on conviction
 - Benefits: valuable goods
- Castle doctrine expansions lowered the (expected) cost of killing someone in self-defense
- If people are rational, then lowering the price of lethal self-defense should increase lethal homicides

Economic theory

- Two types of lethal force in self-defense:
 - True positive: use of lethal force against criminals committing serious crimes
 - False positive: arguments escalate to lethal force that could have been averted with “duty to retreat”; cases of mistaken identity caused by lower expected penalties for being wrong, etc.
- Caveat:
 - Although deterrence is a theoretical possibility, note that the goal of the laws was to protect/enhance victim rights, not deter crime

Castle Doctrine Laws in US, 2000-2010

State	Effective Year	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	2006	Yes	Yes	No	Yes
Alaska	2006	Yes	No	Yes	Yes
Arizona	2006	Yes	Yes	Yes	Yes
Florida	2005	Yes	Yes	Yes	Yes
Georgia	2006	Yes	Yes	No	Yes
Indiana	2006	Yes	Yes	No	Yes
Kansas	2006	Yes	Yes	No	Yes
Kentucky	2006	Yes	Yes	Yes	Yes
Louisiana	2006	Yes	Yes	Yes	Yes
Michigan	2006	Yes	Yes	No	Yes
Mississippi	2006	Yes	Yes	Yes	Yes
Missouri	2007	Yes	No	No	Yes
Montana	2009	Yes	Yes	Yes	No
North Dakota	2007	Yes	No	Yes	Yes
Ohio	2008	Yes	No	Yes	Yes
Oklahoma	2006	Yes	Yes	Yes	Yes
South Carolina	2006	Yes	Yes	Yes	Yes
South Dakota	2006	Yes	Yes	No	No
Tennessee	2007	Yes	Yes	Yes	Yes
Texas	2007	Yes	Yes	Yes	Yes
West Virginia	2008	Yes	Yes	No	No

Economic theory concluded

- Summary:
 - 21 states passed laws removing “duty to retreat” in places outside the home
 - 17 states removed “duty to retreat” in any place one had a legal right to be
 - 13 states include a presumption of reasonable fear
 - 18 states remove civil liability when force was justified under law

Cheng and Hoekstra's identification strategy

- Research design:
 - Estimate the state-level causal effect: what would've happened to these same states had they not passed the law?
 - Compare the changes in outcomes after castle doctrine law adoption to changes in the outcomes in other states in the same region of the country
- Estimation: Panel fixed effects estimation

$$Y_{it} = \beta_1 D_i + \beta_2 T_t + \beta_3 (CDL_{it}) + \alpha_1 X_{it} + c_i + u_t + \varepsilon_{it}$$

- CDL is a dummy variable equalling 1 if the state has a castle doctrine law and zero otherwise
- Note: most of their specifications will include region dummies interacted with year dummies or “region-by-year fixed effects” which means the estimated coefficient must originate from variation within a given region (but across states within that region) in a year

Data

- FBI Uniform Crime Reports (2000-2010)
 - State-level crime rates, or “offenses per 100,000 population”
 - Falsification outcomes: motor vehicle theft and larceny
- Deterrence outcomes:
 - Burglary: the unlawful entry of a structure to commit a felony or a theft
 - Robbery: the taking or attempting to take anything of value from the care, custody or control of a person or persons by force or threat of force or violence and/or putting the victim in fear
 - Aggravated assault: unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury
- Homicide categories
 - ① Total homicides – murder plus non-negligent manslaughter (~14,000 per year)
 - ② Justifiable homicides by private citizens (~250/year)

Standard Errors

- Statistical inference:
 - Cluster the standard errors at the state level
 - Are disturbances random draws from individually identical distribution?
 - It's likely that within a state, unobserved determinants of crime are serially correlated
 - Bertrand, Duflo and Mullainathan (2004) recommend adjusting for serial correlation in unobserved disturbances within states at the level of the treatment
- “Randomization inference”
 - “how likely is it that we estimate effects of this magnitude when using randomly chosen pre-treatment time periods and randomly assigning placebo treatments?”
 - Becoming increasingly commonly done

Region-by-year fixed effects

- Absent passing castle doctrine laws, outcomes in these 21 states would have changed similar to other states in their same region
 - Recall the “region-by-year fixed effects” in the X term
 - By including “region-by-year fixed effects”, they are arguing that unobserved changes in crime are running “parallel” to the treatment states within region over time
 - Need not hold across regions since the across region variation is not being used in this analysis due to the saturation of the model with “region-by-year fixed effects”

Testing the identification strategy

- Assess how including time-varying controls affected estimates
- Examine the effect of the laws on “placebo” outcomes as a falsification (e.g., automobile thefts and larceny)
- Include state-specific linear time trends
 - Alabama, et al. dummy interacted with TREND which equals 1 in 2000, 2 in 2001, . . . , 11 in 2010
 - Forces the identification to come from variation in outcomes around the state-specific linear trend
 - Stronger requirements, in other words
 - Outcomes must be large enough and different enough from a state-specific linear trend otherwise it is collinear with the state-trend
- Include “leads” to test for divergence in the year prior to adoption
- Test for historical tendency for outcomes in adopting states to diverge from control states

Control variables

- Controls (X matrix in earlier equation)
 - Full-time police employment per 100,000 state residents from the LEKOA data (FBI data)
 - Persons incarcerated in state prison per 100,000 residents
 - Shares of white/black men in 15-24 and 25-44 age groups
 - State per capita spending on public assistance
 - State per capita spending on public welfare

Summary Statistics

Table 2: Descriptive Statistics

	Mean (Unweighted)	Mean (Weighted by Population)
Dependent Variables		
Homicides per 100,000 Population	4.8 (2.5)	5.5 (1.9)
Justifiable Homicide by Private Citizens (count)	5.1 (8.2)	11.8 (12.9)
Justifiable Homicide by Police (count)	8.0 (16.9)	23.4 (34.3)
Robberies per 100,000 Population	107.2 (59.6)	143.1 (47.5)
Aggravated Assault per 100,000 Population	267 (131)	296 (114)
Burglary per 100,000 Population	710 (240)	744 (235)
Larceny per 100,000 Population	2,334 (533)	2,328 (532)
Motor Theft per 100,000 Population	331 (178)	381 (174)
Proportion of Robberies in Which a Gun Was Used	0.35 (0.13)	0.37 (0.13)

Step one: Falsification test

- Cheng and Hoekstra (2013) present falsification first to show the reader that they find no association within region over time in the passage of these laws and either larceny rate or motor vehicle theft rate
 - The idea here is to immediately address concerns that what they show you later is due to generic crime trends in those states that pass the laws
 - It's a useful way to assuage doubt people may have, as remember, policy-makers are not just randomly flipping coins when passing laws, but presumably do so because of things they observe on the ground
- Results will be presented separately under six different specifications
 - Each new specification adds more controls
- What should you expect to find on key variables of interest?
 - No statistically significant association between the CDL passage and the placebos; small magnitudes preferably too
 - No association on the one-year lead either
- How do you interpret coefficients?
 - His model is “log outcomes” regressed onto a dummy variable (level), so these are semi-elasticities and approximate percentage changes – but you should transform them by taking the exponential of each coefficient and then differencing it from one to find the actual percentage change
 - Ex: CDL = -0.0137 (column 12, Table 3, “Log (larceny rate)” outcome.)
$$\text{Exp}(-0.0137) = 0.986$$
, and so $1-0.986 = 1.4$. Thus, CDL reduced larceny rates by 1.4 percent, which is not statistically significant.

Results – Falsification Exercise

Table 3: Placebo Tests

OLS - Unweighted						
	7	8	9	10	11	12
Panel A: Larceny						
Castle Doctrine Law	0.00745 (0.0227)	0.00145 (0.0205)	-0.00188 (0.0210)	-0.00445 (0.0226)	-0.00361 (0.0201)	-0.0137 (0.0228)
One Year Before Adoption of Castle Doctrine Law				-0.0103 (0.0114)		
Observation	550	550	550	550	550	550
Panel B: Motor Vehicle Theft						
	Log (Motor Vehicle Theft Rate)					
Castle Doctrine Law	0.0767* (0.0413)	0.0138 (0.0444)	0.00814 (0.0407)	0.00775 (0.0462)	0.00977 (0.0391)	-0.00373 (0.0361)
One Year Before Adoption of Castle Doctrine Law				-0.00155 (0.0287)		
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

Notes: Each column in each panel represents a separate regression. The unit of observation is state-year. Robust standard errors are clustered at the state level. Time-varying controls include policing and incarceration rates, welfare and public assistance spending, median income, poverty rate, unemployment rate, and demographics.

Step two: testing the deterrence hypothesis

- Having found no effect on their placebos, Cheng and Hoekstra (2013) examine the effect of CDL on three deterrence outcomes: burglary, robbery and aggravated assault
 - They will, again, have six specifications per outcome in the “weighted” regression, and then another five for the “unweighted” regression
- What does deterrence look like?
 - Negative signs on the CDL variable is consistent with deterrence – these crimes were “deterring”, in other words
 - Bounds on the magnitudes from the standard errors are used to provide some confidence about the estimates as well

Results – Deterrence

OLS - Weighted by State Population												OLS - Unweighted											
	1	2	3	4	5	6		7	8	9	10	11	12										
Panel A: Burglary							Log (Burglary Rate)							Log (Burglary Rate)									
Castle Doctrine Law	0.0780***	0.0290	0.0223	0.0164	0.0327*	0.0237		0.0572**	0.00961	0.00663	0.00277	0.00683	0.0207										
	(0.0255)	(0.0236)	(0.0223)	(0.0247)	(0.0165)	(0.0207)		(0.0272)	(0.0291)	(0.0268)	(0.0304)	(0.0222)	(0.0259)										
One Year Before Adoption of							-0.0201							-0.0154									
Castle Doctrine Law							(0.0139)							(0.0214)									
Panel B: Robbery							Log (Robbery Rate)							Log (Robbery Rate)									
Castle Doctrine Law	0.0408	0.0344	0.0262	0.0216	0.0376**	0.0515*		0.0448	0.0320	0.00839	0.00552	0.00874	0.0267										
	(0.0254)	(0.0224)	(0.0229)	(0.0246)	(0.0181)	(0.0274)		(0.0331)	(0.0421)	(0.0387)	(0.0437)	(0.0339)	(0.0299)										
One Year Before Adoption of							-0.0156							-0.0115									
Castle Doctrine Law							(0.0167)							(0.0283)									
Panel C: Aggravated Assault							Log (Aggravated Assault Rate)							Log (Aggravated Assault Rate)									
Castle Doctrine Law	0.0434	0.0397	0.0372	0.0362	0.0424	0.0414		0.0555	0.0698	0.0343	0.0305	0.0341	0.0317										
	(0.0387)	(0.0407)	(0.0319)	(0.0349)	(0.0291)	(0.0285)		(0.0604)	(0.0630)	(0.0433)	(0.0478)	(0.0405)	(0.0380)										
One Year Before Adoption of							-0.00343							-0.0150									
Castle Doctrine Law							(0.0161)							(0.0251)									
Observations	550	550	550	550	550	550		550	550	550	550	550	550										
State and Year Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes	Yes	Yes										
Region-by-Year Fixed Effects		Yes	Yes	Yes	Yes	Yes			Yes	Yes	Yes	Yes	Yes										
Time-Varying Controls			Yes	Yes	Yes	Yes				Yes	Yes	Yes	Yes										
Contemporaneous Crime Rates							Yes																
State-Specific Linear Time Trends								Yes															Yes

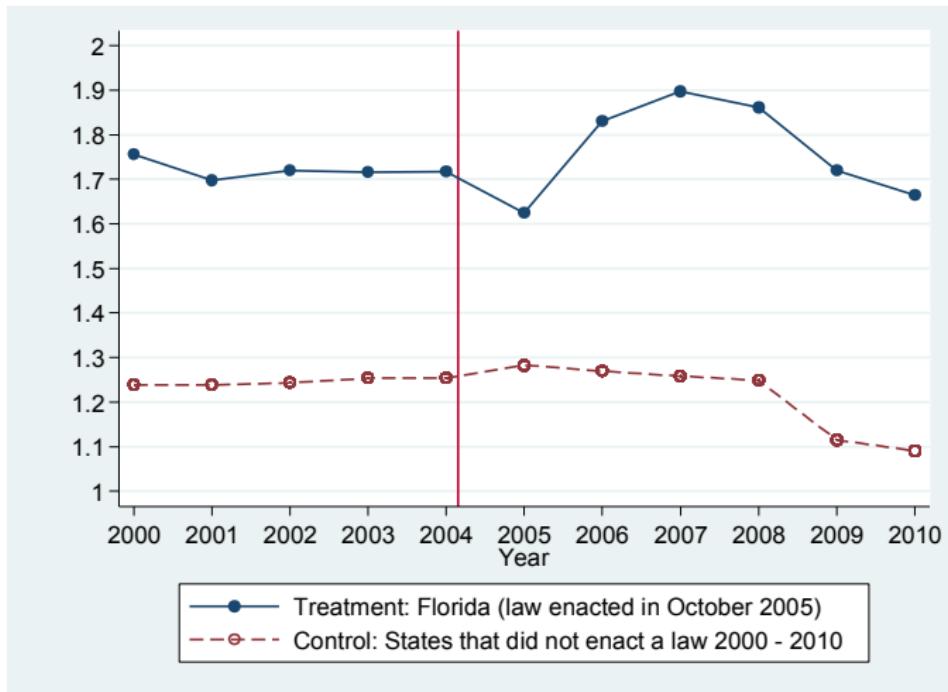
Conclusion

- “In short, these estimates provide strong evidence against the possibility that castle doctrine laws cause economically meaningful deterrence effects” (p. 17)
 - Translation: They can’t find evidence of large deterrence effects
- “Thus, while castle doctrine law may well have benefits to those legally justified in protecting themselves in self-defense, there is no evidence that the law provides positive spillovers by deterring crime more generally” (p. 17)
 - They note in footnote 24 that they cannot measure the benefits to victims whose crimes were deterred, or the benefits from lower legal costs; their focus is limited to whether it deterred the crimes, not whether the net benefits from the laws were positive
 - Obviously, if there is no deterrence, though, then the net benefits are lower from CDL than they would be if they did deter

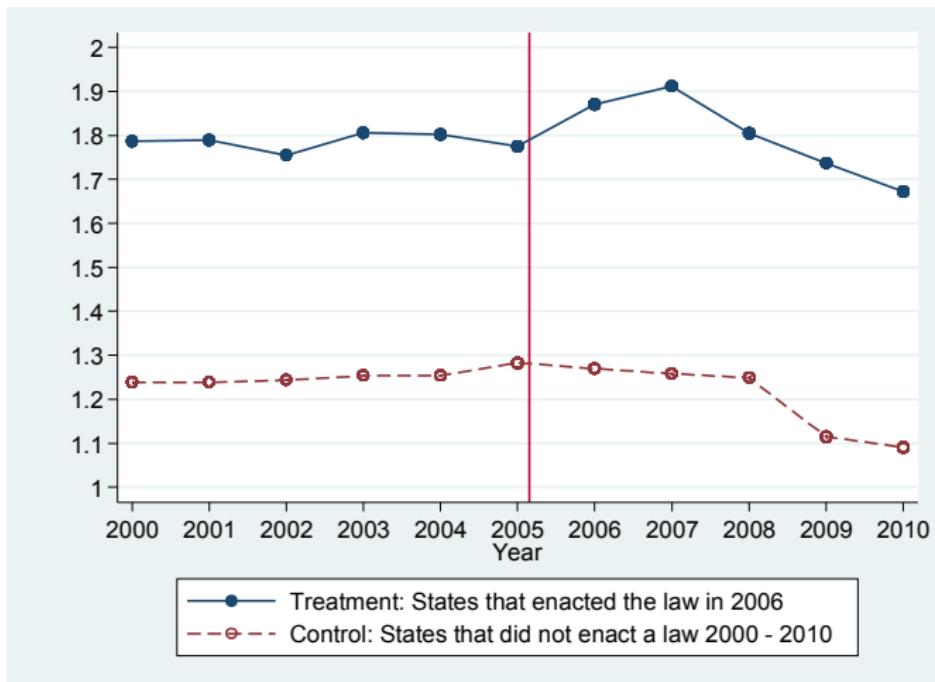
Step 3: Homicides

- The key finding in this study is the very large effect that CDL had on homicides and non-negligent manslaughter
- As the effects are quite large, their strategy is first to present pictures
- The pictures are a bit tricky, though, since they're going to also present pictures for the control and treatment group units
- This is going to be useful for eye-balling the parallel trends pre-treatment.
 - Remember, though – he needs parallel trends within-region – these figures don't show that
 - But you should start with pictures; don't fetishize regression

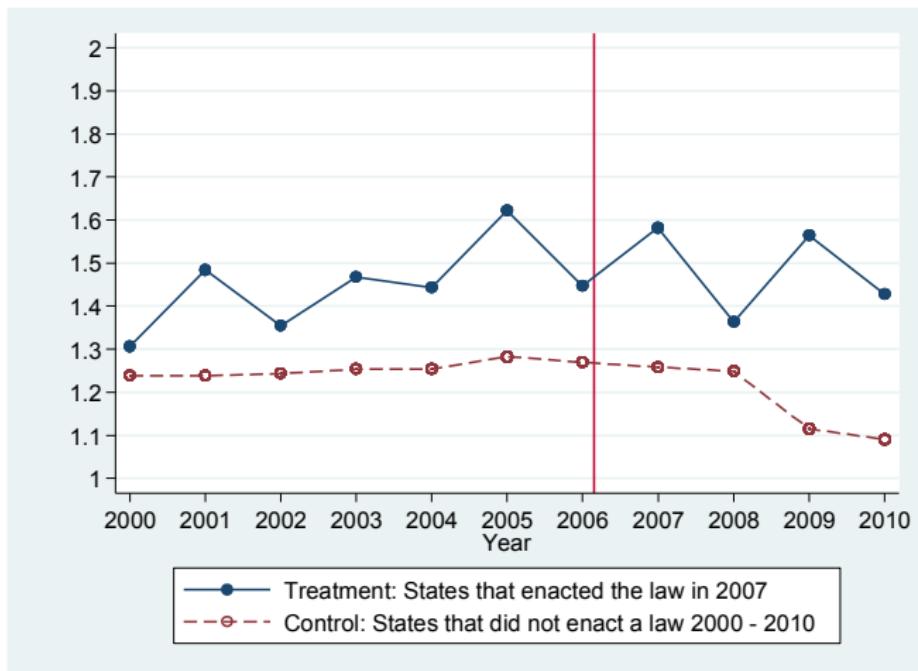
Log Homicide Rates – 2005 Adopter = Florida



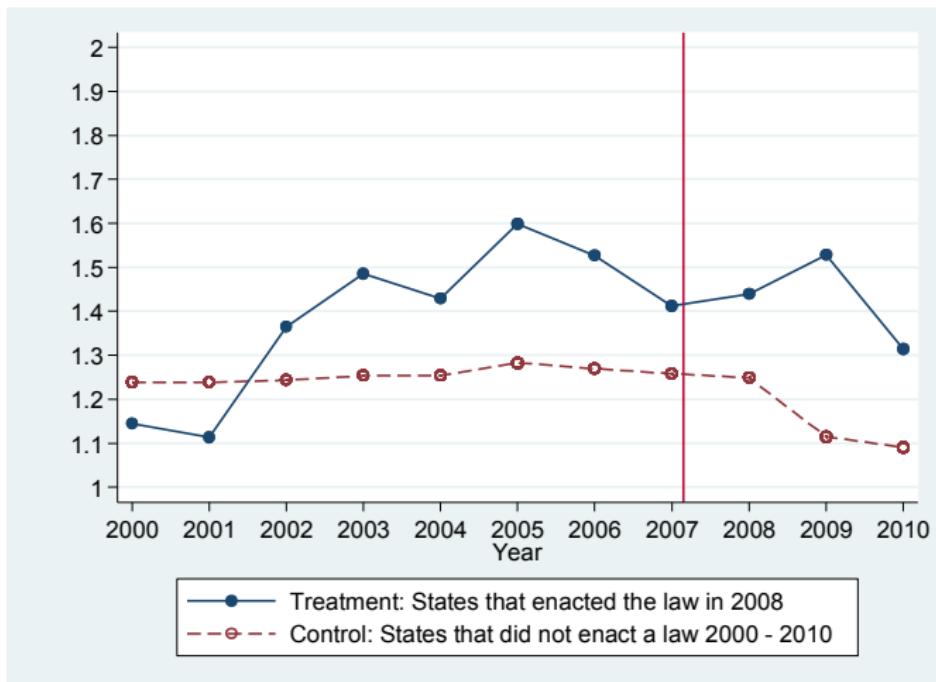
Log Homicide Rates – 2006 Adopter (13 states)



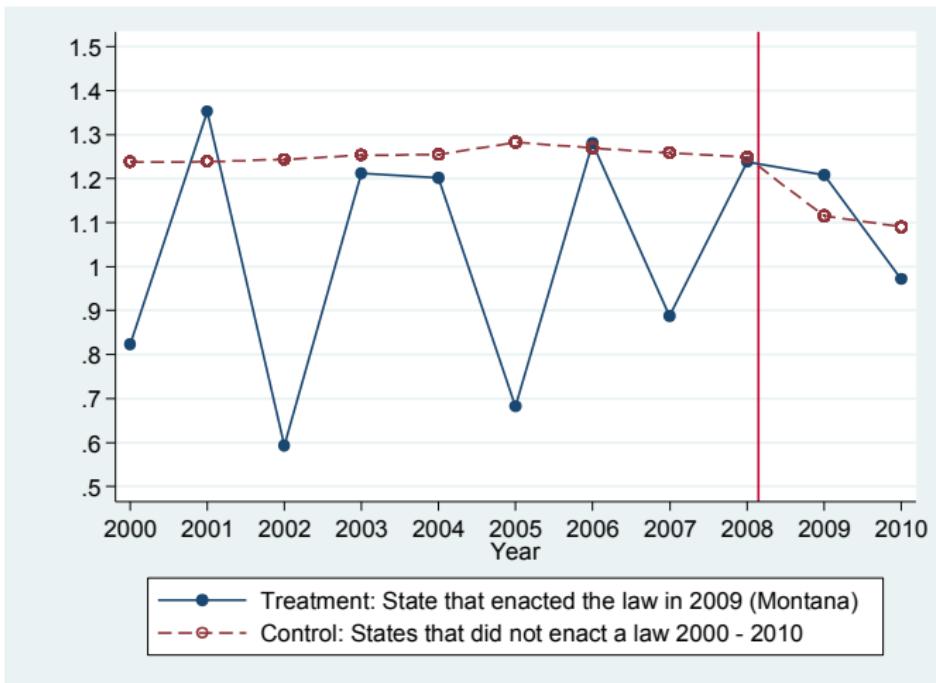
Log Homicide Rates – 2007 Adopter (4 states)



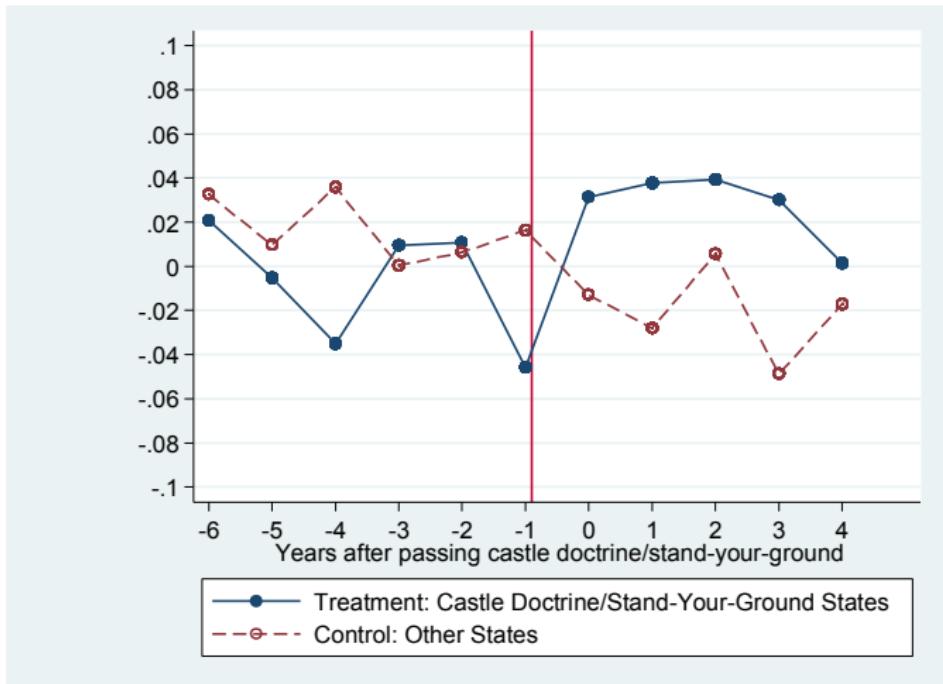
Log Homicide Rates – 2008 Adopter (2 states)



Log Homicide Rates – 2009 Adopter = Montana



Residual log homicide rates



Estimation results

- Before going into the estimation results, here's what you are looking for
 - This second hypothesis wherein reductions in the expected penalties and costs associated with self-defense due to CDL causes lethal violence to increase (non-deterrence escalation of violence) should exhibit a positive association on the DD variable
 - It should be different from zero statistically and economically meaningful
- He will estimate the model using panel fixed effects estimation and “negative binomial count models”
 - Because of the smaller number of annual homicides each year in a state, he moves away from homicide rates in some specifications and looks at “count” outcomes
 - He uses a class of estimators more appropriate for “counts” called “count models”, like the negative binomial estimated with maximum likelihood
 - Results are robust to least squares and count models

Homicide – OLS

	1	2	3	4	5	6
<u>Panel A: Log Homicide Rate (OLS - Weighted)</u>						
Castle Doctrine Law	0.0801** (0.0342)	0.0946*** (0.0279)	0.0937*** (0.0290)	0.0875** (0.0337)	0.0985*** (0.0299)	0.100** (0.0388)
One Year Before Adoption of Castle Doctrine Law				-0.0212 (0.0246)		
Observations	550	550	550	550	550	550
<u>Panel B: Log Homicide Rate (OLS - Unweighted)</u>						
Castle Doctrine Law	0.0877 (0.0638)	0.0811 (0.0769)	0.0600 (0.0684)	0.0461 (0.0764)	0.0580 (0.0662)	0.0672 (0.0450)
One Year Before Adoption of Castle Doctrine Law				-0.0557 (0.0494)		
Observations	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
Contemporaneous Crime Rates					Yes	
State-Specific Linear Time Trends						Yes

Homicide – Negative Binomial; Murder – OLS

	1	2	3	4	5	6
<u>Panel C: Homicide (Negative Binomial - Unweighted)</u>						
Castle Doctrine Law	0.0565* (0.0331)	0.0734** (0.0305)	0.0879*** (0.0313)	0.0783** (0.0355)	0.0937*** (0.0302)	0.108*** (0.0346)
One Year Before Adoption of Castle Doctrine Law				-0.0352 (0.0260)		
Observations	550	550	550	550	550	550
<u>Panel D: Log Murder Rate (OLS - Weighted)</u>						
Castle Doctrine Law	0.0906** (0.0424)	0.0955** (0.0389)	0.0916** (0.0382)	0.0884** (0.0404)	0.0981** (0.0391)	0.0813 (0.0520)
One Year Before Adoption of Castle Doctrine Law				-0.0110 (0.0230)		
Observations	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
Contemporaneous Crime Rates					Yes	
State-Specific Linear Time Trends						Yes

Homicide – identification test

- Question: Did homicide rates of adopting states show a general historical tendency to increase over time relative to other non-adopting states from the same region?
- Method: Move the 11-year panel back one year at a time (covering 1960-2009) and estimate 40 placebo “effects” of passing CDL 1 to 40 years earlier
- Findings

Method	Average estimate	Estimates larger than actual estimate
Weighted OLS	* -0.003	0/40
Unweighted OLS	0.001	1/40
Negative binomial	0.001	0/40

Interpretation

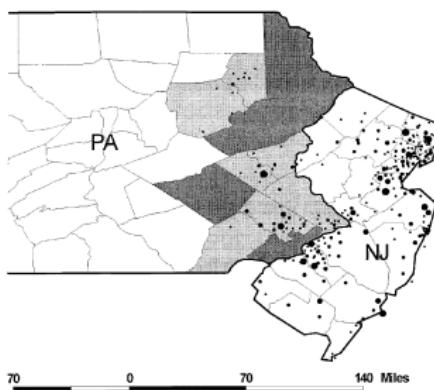
- No evidence that Castle Doctrine/Stand Your Ground Laws deter violent crimes such as burglary, robbery and aggravated assault
- These laws do lead to an 8% net increase in homicide rates, translating to around 600 additional homicides *per year* across the 21 adopting states
 - Unlikely that all of the additional homicides were legally justified
- Economics of crime and behavior: incentives matter in some contexts but not others

Differences-in-differences: Card and Krueger (1994)

- Suppose you are interested in the effect of minimum wages on employment (a classic and controversial question in labor economics).
- In a competitive labor market, increases in the minimum wage would move us up a downward sloping labor demand curve → employment would fall

DD: Card and Krueger (1994)

- Card and Krueger (1994) analyzed the effect of a minimum wage increase in New Jersey using a differences-in-differences (DD) methodology
- In February 1992, New Jersey increased the state minimum wage from \$4.25 to \$5.05. Pennsylvania's minimum wage stayed at \$4.25.



- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase in New Jersey

DD Strategy

- DD is a version of fixed effects estimation. To see this formally:

Y_{ist}^1 : employment at restaurant i , state s , time t with a high w^{min}

Y_{ist}^0 : employment at restaurant i , state s , time t with a low w^{min}

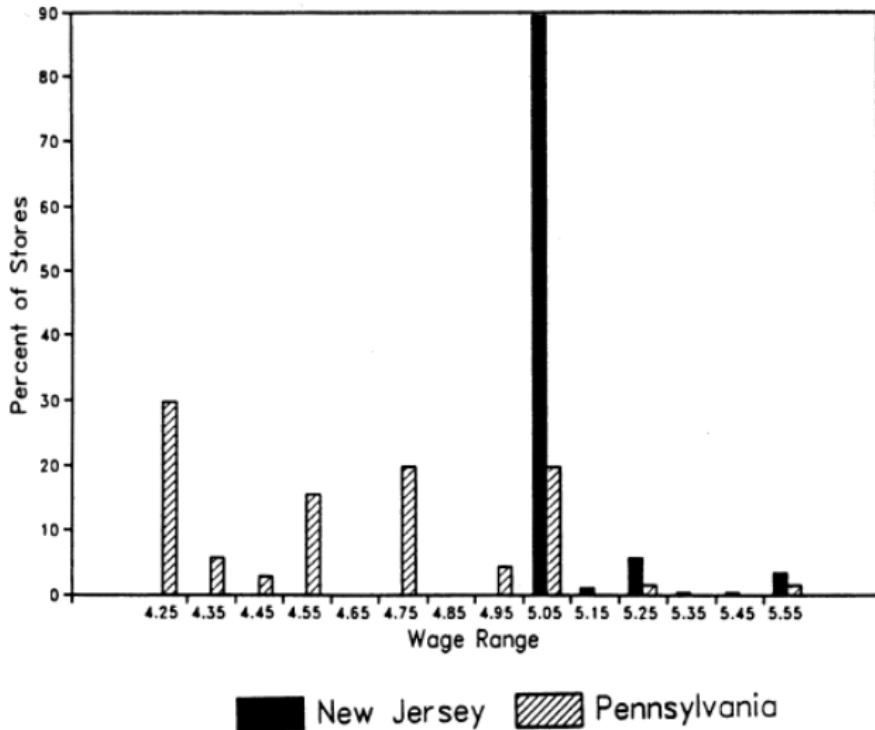
- In practice of course we only see one or the other. We then assume that:

$$E[Y_{its}^0 | s, t] = \gamma_s + \lambda_t$$

- In the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect, γ_s and a year effect, λ_t that is common across states
- Let D_{st} be a dummy for high-minimum wage states and periods
- Assuming $E[Y_{its}^1 - Y_{its}^0 | s, t] = \delta$ is the treatment effect, observed employment can be written:

$$Y_{its} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{its}$$

November 1992



DD Strategy II

- In New Jersey

- Employment in February is

$$E(Y_{ist}|s = NJ, t = Feb) = \gamma_{NJ} + \lambda_{Feb}$$

- Employment in November is:

$$E(Y_{ist}|s = NJ, t = Nov) = \gamma_{NJ} + \lambda_{Nov} + \delta$$

- Difference between November and February

$$E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb) = \lambda_N - \lambda_F + \delta$$

- In Pennsylvania

- Employment in February is

$$E(Y_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb}$$

- Employment in November is:

$$E(Y_{ist}|s = PA, t = Nov) = \gamma_{PA} + \lambda_{Nov}$$

- Difference between November and February

$$E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) = \lambda_N - \lambda_F$$

DD Strategy III

- The DD strategy amounts to comparing the change in employment in NJ to the change in employment in PA.
- The population DD are:

$$\left(E(Y_{ist} | s = NJ, t = Nov) - E(Y_{ist} | s = NJ, t = Feb) \right) - \left(E(Y_{ist} | s = PA, t = Nov) - E(Y_{ist} | s = PA, t = Feb) \right) = \delta$$

- This is estimated using the sample analog of the population means

Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	– 2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	– 0.14 (1.07)
3. Change in mean FTE employment	– 2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Surprisingly, employment *rose* in NJ relative to PA after the minimum wage change

Regression DD

- We can estimate the DD estimator in a regression framework
- Advantages:
 - It's easy to calculate the standard errors
 - We can control for other variables which may reduce the residual variance (lead to smaller standard errors)
 - It's easy to include multiple periods
 - We can study treatments with different treatment intensity.
(e.g., varying increases in the minimum wage for different states)
- The typical regression model we estimate is

$$\text{Outcome}_{it} = \beta_1 + \beta_2 \text{Treat}_i + \beta_3 \text{Post}_t + \beta_4 (\text{Treat} \times \text{Post})_{it} + \varepsilon_{it}$$

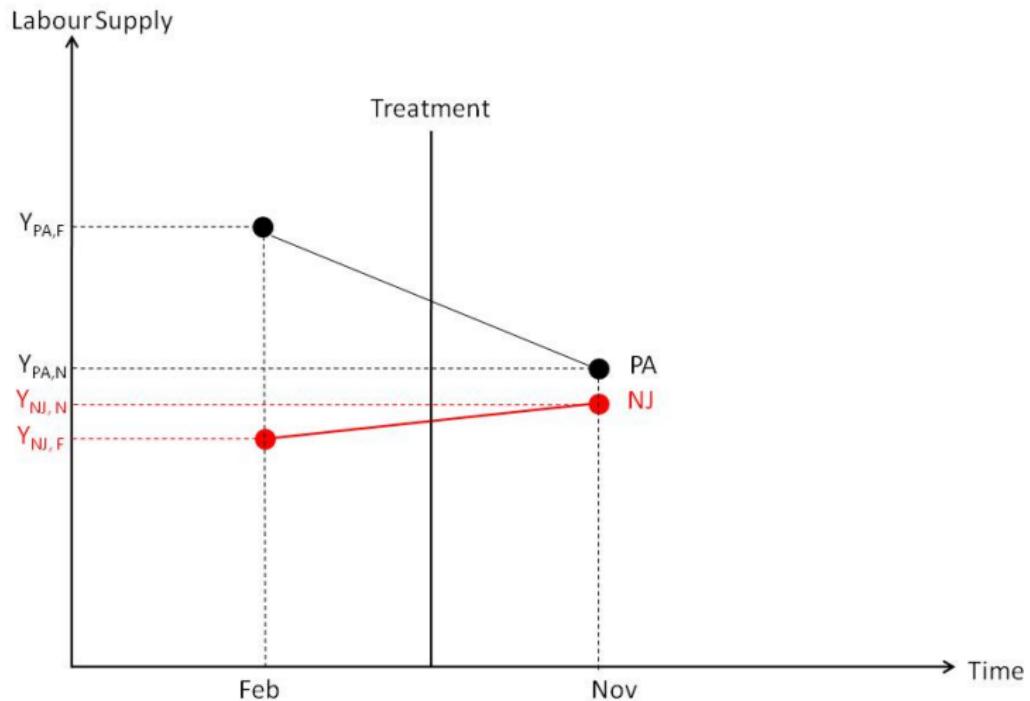
where Treat is a dummy if the observation is in the treatment group and Post is a post treatment dummy

Regression DD - Card and Krueger

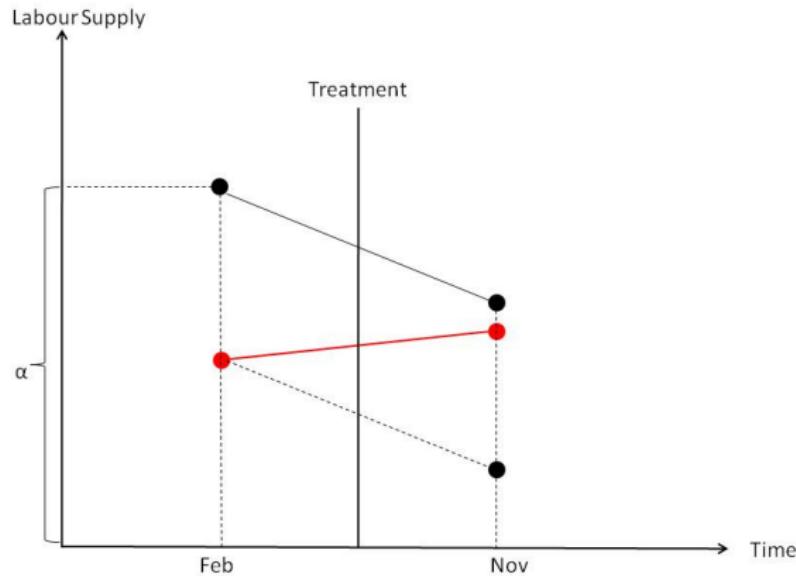
- In the Card and Krueger case, the equivalent regression would be:

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{its}$$

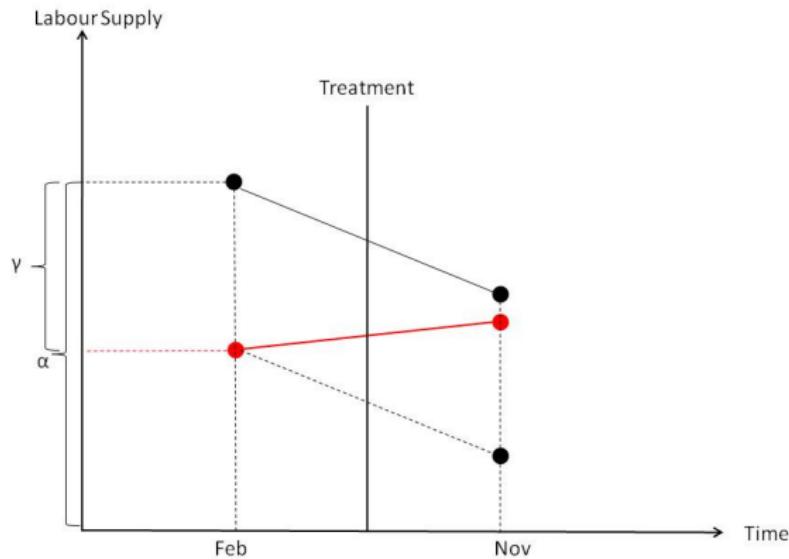
- NJ is a dummy equal to 1 if the observation is from NJ
- d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
 - PA Pre: α
 - PA Post: $\alpha + \lambda$
 - NJ Pre: $\alpha + \gamma$
 - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DD estimate: $(NJ \text{ Post} - NJ \text{ Pre}) - (PA \text{ Post} - PA \text{ Pre}) = \delta$



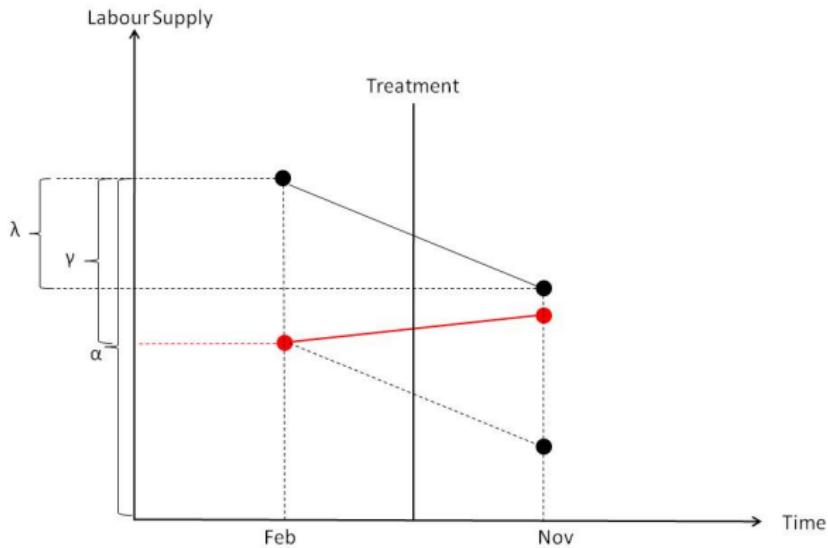
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{ist}$$



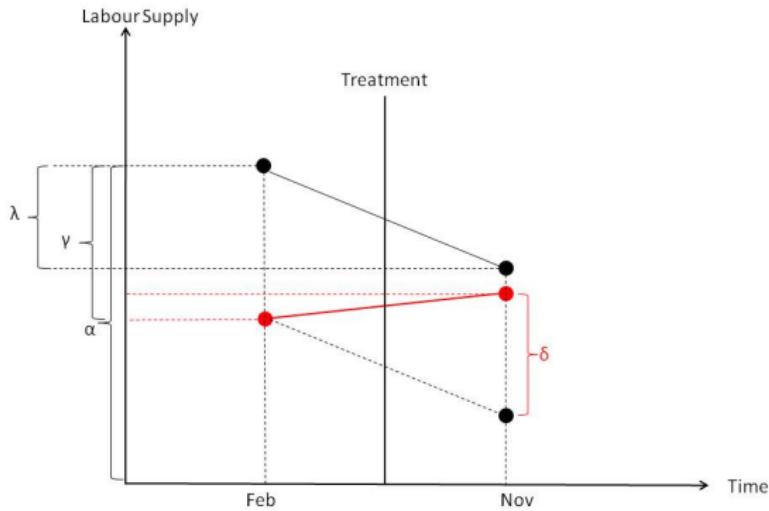
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{ist}$$

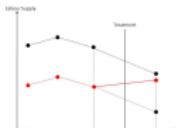


$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta (NJ \times d)_{st} + \varepsilon_{ist}$$



Key assumption of any DD strategy: Parallel trends

- The key assumption for any DD strategy is that the outcome in treatment and control group would follow the same time trend in the absence of the treatment
- This doesn't mean that they have to have the same mean of the outcome
- Parallel trends are difficult to verify because technically one of the parallel trends is an unobserved counterfactual
- But one often will check using pre-treatment data to show that the trends are the same
- Even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)



Regression DD Including Leads and Lags

- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

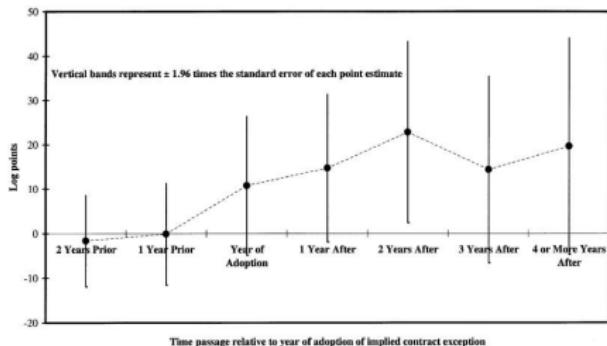
$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-q}^{-1} \gamma_{\tau} D_{s\tau} + \sum_{\tau=0}^m \delta_{\tau} D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

- Treatment occurs in year 0
- Includes q leads or anticipatory effects
- Includes m leads or post treatment effects

Study including leads and lags – Autor (2003)

- Autor (2003) includes both leads and lags in a DD model analyzing the effect of increased employment protection on the firm's use of temporary help workers
- In the US employers can usually hire and fire workers at will
- Some states courts have made some exceptions to this employment at will rule and have thus increased employment protection
- The standard thing to do is normalize the adoption year to 0
- Autor then analyzes the effect of these exemptions on the use of temporary help workers.

Results



- The leads are very close to 0. → no evidence for anticipatory effects (good news for the parallel trends assumption).
- The lags show the effect increases during the first years of the treatment and then remains relatively constant.

Standard errors in DD strategies

- Many papers using DD strategies use data from many years – not just 1 pre and 1 post period
- The variables of interest in many of these setups only vary at a group level (say a state level) and outcome variables are often serially correlated
- In the Card and Krueger study, it is very likely that employment in each state is not only correlated within the state but also serially correlated
- As Bertrand, Duflo and Mullainathan (2004) point out, conventional standard errors often severely underestimate the standard deviation of the estimators – standard errors are biased downward

Standard errors in DD – practical solutions

- Bertrand, Duflo and Mullainathan propose the following solutions:
 - ① Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)
 - ② Clustering standard errors at the group level (in STATA one would simply add `, cluster(state)` to the regression equation if one analyzes state level variation)
 - ③ Aggregating the data into one pre and one post period. Literally works if there is only one treatment data. With staggered treatment dates one should adopt the following procedure:
 - Regress Y_{st} onto state FE, year FE and relevant covariates
 - Obtain residuals from the treatment states only and divide them into 2 groups: pre and post treatment
 - Then regress the two groups of residuals onto a post dummy
- Correct treatment of standard errors sometimes makes the number of groups very small: in the Card and Krueger study the number of groups is only 2.

Threats to validity

- ① Non-parallel trends
- ② Compositional differences
- ③ Long-term effects vs. reliability
- ④ Functional form dependence

Non-parallel trends

- Often policymakers will select the treatment and controls based on pre-existing differences in outcomes – practically guaranteeing the parallel trends assumption will be violated.
- “Ashenfelter dip”
 - Named after Orley Ashenfelter, labor economist at Princeton
 - Participants in job trainings program often experience a “dip” in earnings just prior to entering the program
 - Since wages have a natural tendency to mean reversion, comparing wages of participants and non-participants using DD leads to an upward biased estimate of the program effect.
- Regional targeting. NGOs may target villages that appear most promising, or worse off, which is a form of selection bias and violates parallel trends

Checks for DD Design

- Very common for readers and others to request a variety of “robustness checks” from a DD design
- Think of these as along the same lines as the leads and lags we already discussed
 - Falsification test using data for prior periods (already discussed)
 - Falsification test using data for alternative control group
 - Falsification test using alternative “placebo” outcome that should not be affected by the treatment

Alternative control group – DDD

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES
ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. Treatment Individuals: Married Women, 20–40 Years Old:			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:		−0.062 (0.022)	
B. Control Group: Over 40 and Single Males 20–40:			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:		−0.008: (0.014)	
DDD:		−0.054 (0.026)	

DDD in Regression

$$\begin{aligned}W_{ijt} = & \alpha + \beta_1 X_{ijt} + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\& + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt}\end{aligned}$$

- The DDD estimate is the difference between the DD of interest and a placebo DD (which is supposed to be zero)
- If the placebo DD is non-zero, it might be difficult to convince the reviewer that the DDD removed all the bias
- If the placebo DD is zero, then DD and DDD give the same results but DD is preferable because standard errors are smaller for DD than DDD

Falsification test with placebo outcome

- Cheng and Hoekstra (2013) examine the effect of castle doctrine gun laws on homicides. They investigate the effect of the laws on non-gun related offenses like grand theft auto and find no evidence of an effect (“second order outcomes”)
- Auld and Grootendorst (2004) estimated standard “rational addiction” models (Becker and Murphy 1988) on data with milk, eggs, oranges and apples. They find these plausibly non-addictive goods are addictive, which casts doubt on the research design of all the rational addiction models.
- Several studies found significant network effects on outcomes like obesity, smoking, alcohol use and happiness. Cohen-Cole and Fletcher (2008) use similar models and data and find similar network “effects” for things that *aren't* contagious like acne, height and headaches

Threats to validity – compositional differences

- One of the risks of a repeated cross-section is that the composition of the sample may have changed between the pre and post period
- Hong (2011) uses repeated cross-sectional data from the Consumer Expenditure Survey (CEX) containing music expenditure and internet use for a random sample of households
- Study exploits the emergence of Napster (first file sharing software widely used by Internet users) in June 1999 as a natural experiment
- Study compares internet users and internet non-users before and after emergence of Napster

Figure 1: Internet Diffusion and Average Quarterly Music Expenditure in the CEX

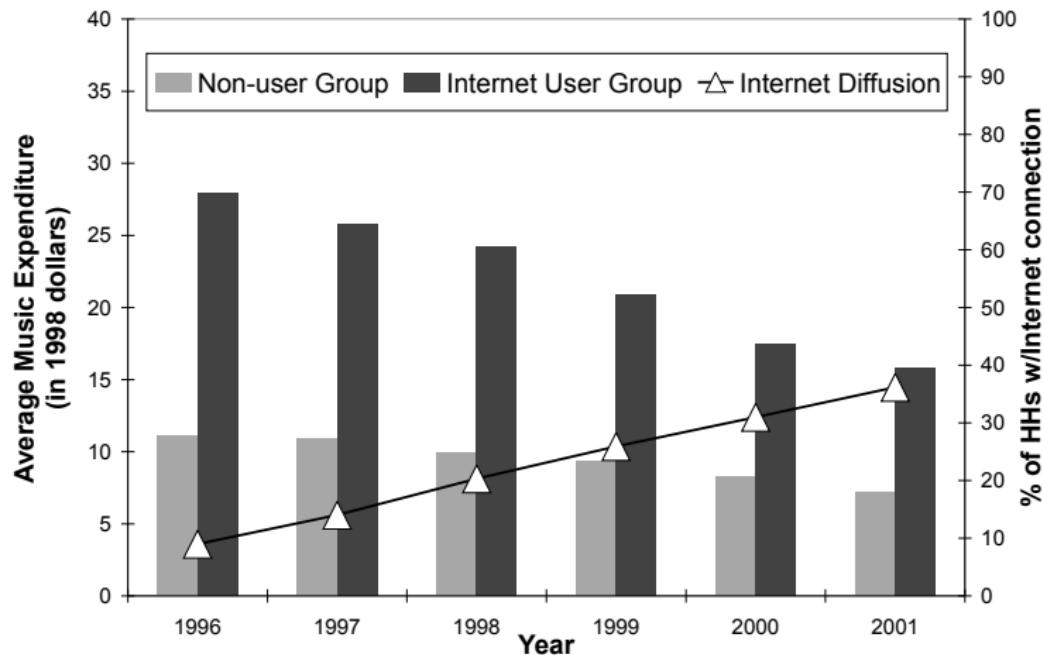


Table 1: Descriptive Statistics for Internet User and Non-user Groups^a

Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	\$25.73	\$10.90	\$24.18	\$9.97	\$20.92	\$9.37
Entertainment	\$195.03	\$96.71	\$193.38	\$84.92	\$182.42	\$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	\$52,887	\$30,459	\$51,995	\$28,169	\$49,970	\$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

Diffusion of the Internet changes samples (e.g., younger music fans are early adopters)

Comparative case studies

- In qualitative comparative case studies, the goal is to reason inductively the causal effects of events or characteristics of a single unit on some outcome, but oftentimes through logic and historical analysis.
 - May not answer the causal questions at all because there is rarely a counterfactual, or if so, it's ad hoc.
 - Classic example of comparative case study approach is Alexis de Toqueville's Democracy in America
- Quantitative comparative case studies are more explicitly causal designs. Usually a natural experiment applied to a single aggregate unit (e.g., city, school, firm, state, country)
- Comparative case studies: compare the evolution of an aggregate outcome for the unit affected by the intervention to the evolution of the same aggregate for some control group (Card 1990; Card and Krueger 1994; Abadie and Gardeazabal 2003)

Motivating example: The Mariel Boatlift

- How do inflows of immigrants affect the wages and employment of natives in local labor markets?
- Card (1990) uses the Mariel boatlift of 1980 as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives



Motivating example: The Mariel Boatlift

- How do inflows of immigrants affect the wages and employment of natives in local labor markets?
- Card (1990) uses the Mariel boatlift of 1980 as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives
- The Mariel Boatlift increased the Miami labor force by 7%
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and four comparison cities (Atlanta, Los Angeles, Houston, Tampa-St. Petersburg)

Motivating example: the Mariel Boatlift

Differences-in-differences estimates of the effect of immigration on unemployment^a

	Group	Year		
		1979 (1)	1981 (2)	1981–1979 (3)
Whites				
(1)	Miami	5.1 (1.1)	3.9 (0.9)	– 1.2 (1.4)
(2)	Comparison cities	4.4 (0.3)	4.3 (0.3)	– 0.1 (0.4)
(3)	Difference Miami-comparison	0.7 (1.1)	– 0.4 (0.95)	– 1.1 (1.5)
Blacks				
(4)	Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
(5)	Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
(6)	Difference Miami-comparison	– 2.0 (1.9)	– 3.0 (2.0)	– 1.0 (2.8)

^a Notes: Adapted from Card (1990, Tables 3 and 6). Standard errors are shown in parentheses.

Comparative case studies

- Advantages:
 - Policy interventions often take place at an aggregate level
 - Aggregate/macro data are often available
- Problems:
 - Selection of control group is often ambiguous
 - Standard errors do not reflect uncertainty about the ability of the control group to reproduce the counterfactual of interest

Synthetic Control method

- A distinctive feature of comparative case studies is that units of analysis are usually aggregate entities, like countries or regions, for which suitable single comparisons often do not exist
- The synthetic control method is based on the observation that, when the units of analysis are a few aggregate entities, a combination of comparison units (a “synthetic control”) often does a better job reproducing the characteristics of a treated unit than any single comparison unit alone.
- Motivated by this consideration, the comparison unit in the synthetic control method is selected as the *weighted average of all potential comparison units* that best resembles the characteristics of the treated unit(s)

Synthetic control method: advantages

- Precludes extrapolation
- Does not require access to post-treatment outcomes in the “design” phase of the study, when synthetic controls are calculated
- Makes explicit the contribution of each comparison unit to the counterfactual of interest
- Allows researchers to use quantitative and qualitative techniques to analyze the similarities and differences between the units representing the case of interest and the synthetic control
- Formalizing the way comparison units are chosen not only represents a way of systemizing comparative case studies, it also has direct implications for inference

Synthetic control method: estimation

- Suppose that we observe $J + 1$ units in periods $1, 2, \dots, T$
- Unit “one” is exposed to the intervention of interest (that is, “treated”) during periods $T_0 + 1, \dots, T$
- The remaining J are an untreated reservoir of potential controls (a “donor pool”)
- Let Y_{it}^N be the outcome that would be observed for unit i at time t in the absence of the intervention
- Let Y_{it}^I be the outcome that would be observed for unit i at time t if unit i is exposed to the intervention in periods $T_0 + 1$ to T .
- We aim to estimate the effect of the intervention on the treated unit $(\alpha_{1T_0+1}, \dots, \alpha_{1T})$ where

$$\alpha_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N$$

for $t > T_0$ and Y_{1t} is the outcome for unit one at time t

Synthetic control method: implementation

- Let $W = (w_2, \dots, w_{J+1})'$ with $w_j \geq 0$ for $j = 2, \dots, J+1$ and $w_2 + \dots + w_{J+1} = 1$. Each value of W represents a potential synthetic control
- Let X_1 be a $(k \times 1)$ vector of pre-intervention characteristics for the treated unit. Similarly, let X_0 be a $(k \times J)$ matrix which contains the same variables for the unaffected units.
- The vector $W^* = (w_2^*, \dots, w_{J+1}^*)'$ is chosen to minimize $\|X_1 - X_0 W\|$, subject to our weight constraints
- Let Y_{jt} be the value of the outcome for unit j at time t . For a post-intervention period t (with $t \geq T_0$) the synthetic control estimator is:

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}$$

Synthetic control method: estimation

- That is to reproduce the counterfactual Y_{it}^N we find the combination of untreated units that best resembles the treated unit before the intervention in terms of the values of k relevant covariates (predictors of the outcome of interest)
- Example: 1988 California's tobacco control program (Proposition 99):
 - Treated unit: California
 - Outcome of interest: tobacco consumption in California after 1988
 - Potential controls: other US states
 - Covariates: predictors of state-level tobacco consumption measured before 1988

Synthetic control method: estimation

- Abadie, et al. consider $\|X_1 - X_0 W\| = \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)}$ where V is some $(k \times k)$ symmetric and positive semidefinite matrix
- Let X_{jm} be the value of the m -th covariates for unit j
- Typically, V is diagonal with main diagonal v_1, \dots, v_k . Then, the synthetic control weights w_2^*, \dots, w_{J+1}^* minimize:

$$\sum_{m=1}^k v_m \left(X_{1m} - \sum_{j=2}^{J+1} w_j X_{jm} \right)^2$$

where v_m is a weight that reflects the relative importance that we assign to the m -th variable when we measure the discrepancy between the treated unit and the synthetic controls

Synthetic control estimation: estimation

- The choice of V is important
 - W^* depends on the choice of V
 - The synthetic control $W^*(V)$ is meant to reproduce the behavior of the outcome variable for the treated unit in the absence of the treatment
 - Therefore, the weights v_1, \dots, v_k should reflect the predictive value of the covariates

Synthetic control estimation: estimation

- Choice of v_1, \dots, v_k can be based on
 - Subjective assessment of the predictive power of each of the covariates, or calibration inspecting how different values for v_1, \dots, v_k affect the discrepancies between the treated unit and the synthetic control
 - Use regression to assess the predictive power of the covariates
 - Minimize mean square prediction error (MSPE):

$$\sum_{t=1}^{T_0} \left(Y_{1t} - \sum_{j=2}^J w_j^*(V) Y_{jt} \right)^2$$

- Cross-validation
 - Divide the pre-treatment period into an initial **training** period and a subsequent **validation** period
 - For any given V , calculate $W^*(V)$ in the training period.
 - Minimize the MSPE of $W^*(V)$ in the validation period

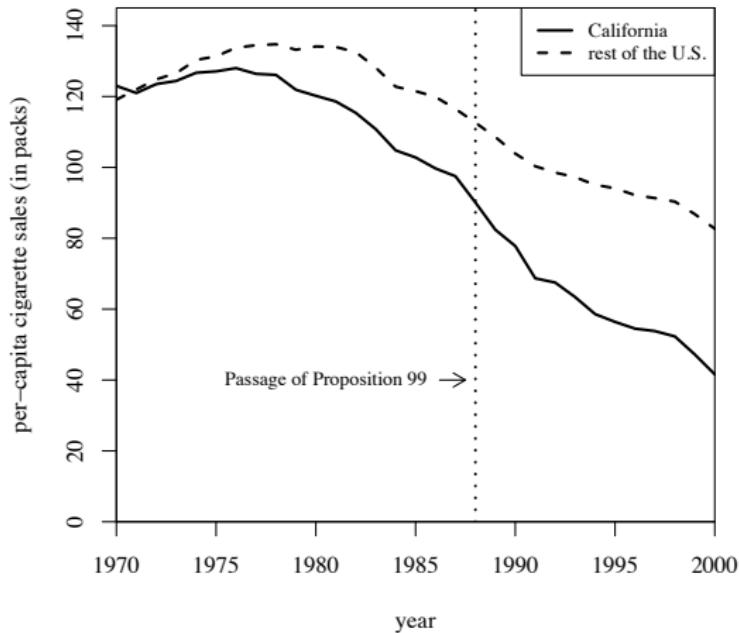
What about unobserved factors?

- Comparative case studies are complicated by unmeasured factors affecting the outcome variables as well as heterogeneity in the effect of observed and unobserved factors
- However, as we will see, if the number of pre-intervention periods in the data is large, matching on pre-intervention outcomes allows us to control for heterogeneous responses to multiple unobserved factors
- Intuition: only units that are alike in observed and unobserved determinants of the outcome variable as well as in the effect of those determinants on the outcome variable should produce similar trajectories of the outcome variable over extended periods of time

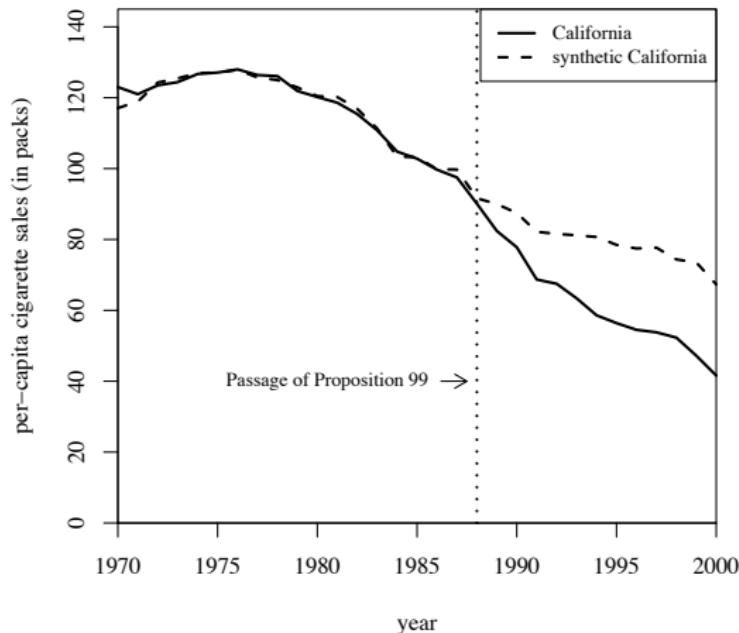
The Application: California's Proposition 99

- In 1988, California first passed comprehensive tobacco control legislation:
 - increased cigarette tax by 25 cents/pack
 - earmarked tax revenues to health and anti-smoking budgets
 - funded anti-smoking media campaigns
 - spurred clean-air ordinances throughout the state
 - produced more than \$100 million per year in anti-tobacco projects
- Other states that subsequently passed control programs are excluded from donor pool of controls (AK, AZ, FL, HI, MA, MD, MI, NJ, OR, WA, DC)

Cigarette Consumption: CA and the Rest of the US



Cigarette Consumption: CA and synthetic CA

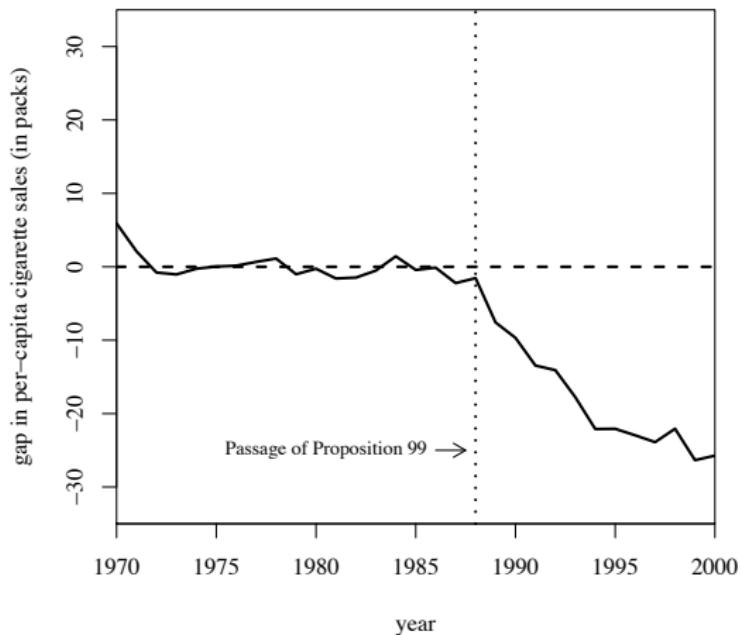


Predictor Means: Actual vs. Synthetic California

Variables	California		Average of 38 control states
	Real	Synthetic	
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15-24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

Note: All variables except lagged cigarette sales are averaged for the 1980-1988 period (beer consumption is averaged 1984-1988).

Smoking Gap between CA and synthetic CA



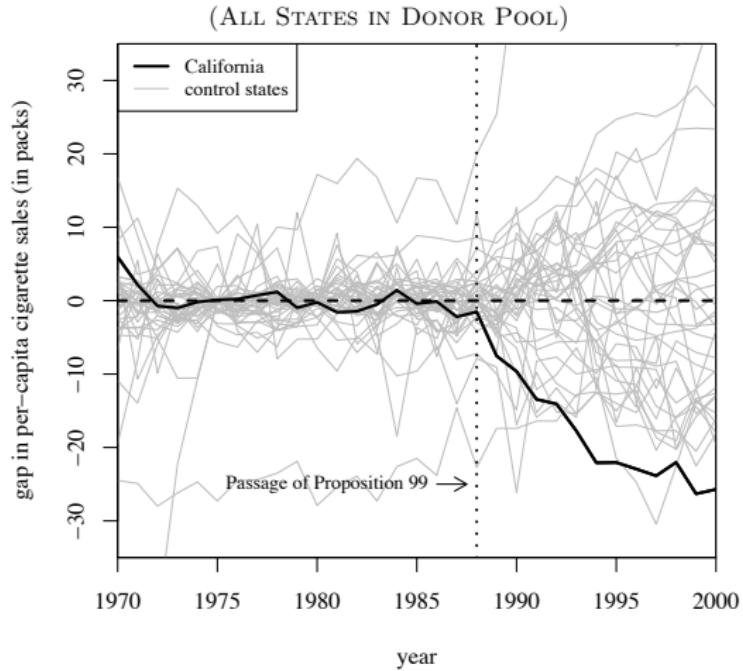
Inference

- To evaluate the significance of our results, we conduct a series of placebo experiments where we reassign the tobacco control program to states other than California
- We proceed as follows:
 - Iteratively apply the synthetic method to each country/state in the donor pool and obtain a distribution of placebo effects
 - Compare the gap (RMSPE) for California to the distribution of the placebo gaps. For example the post-Prop. 99 RMSPE is:

$$RMSPE = \left(\frac{1}{T - T_0} \sum_{t=T_0+1}^T \left(Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \right)^2 \right)^{\frac{1}{2}}$$

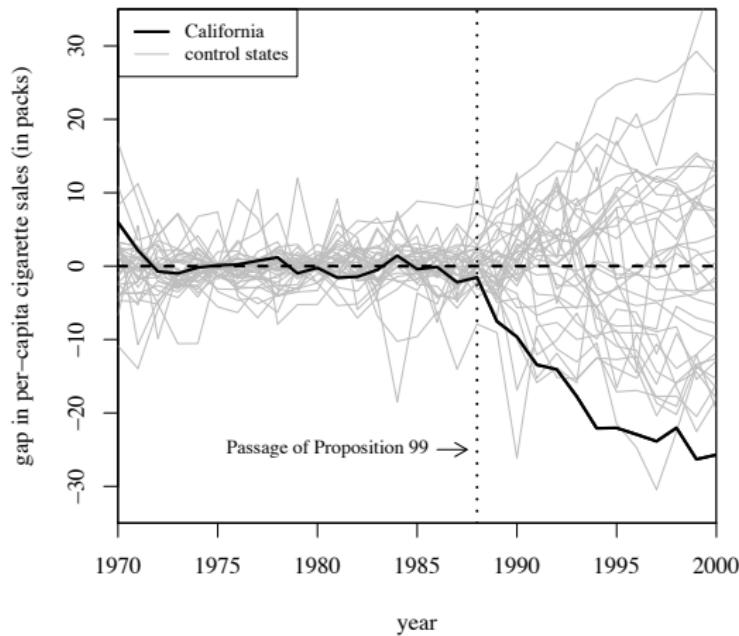
- Question is whether the effect estimated by the synthetic control for California is large relative to the effect estimated for a state chosen at random

Smoking Gap for CA and 38 control states



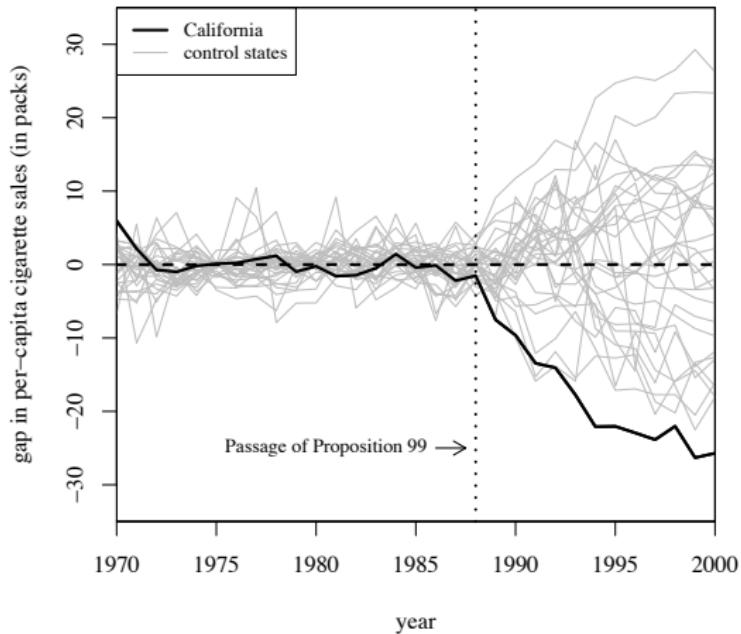
Smoking Gap for CA and 34 control states

(PRE-PROP. 99 MSPE \leq 20 TIMES PRE-PROP. 99 MSPE FOR CA)



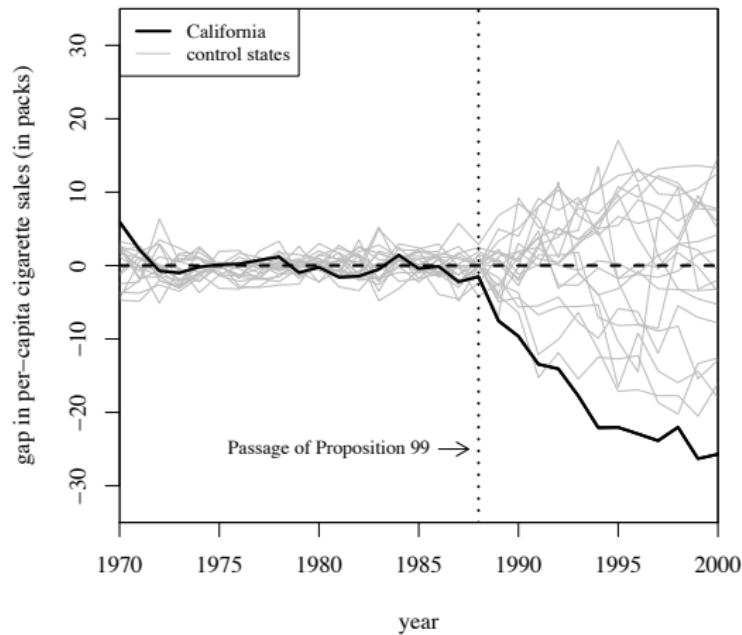
Smoking Gap for CA and 29 control states

(PRE-PROP. 99 MSPE \leq 5 TIMES PRE-PROP. 99 MSPE FOR CA)

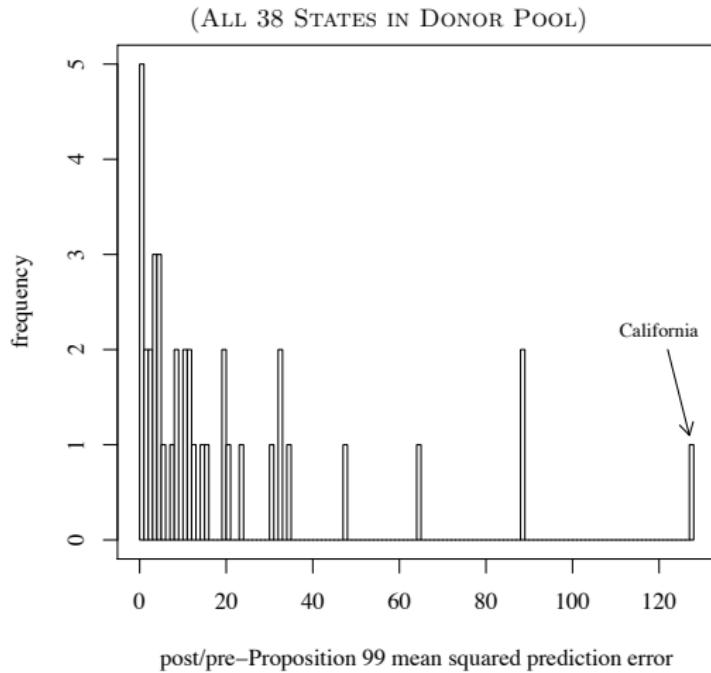


Smoking Gap for CA and 19 control states

(PRE-PROP. 99 MSPE \leq 2 TIMES PRE-PROP. 99 MSPE FOR CA)



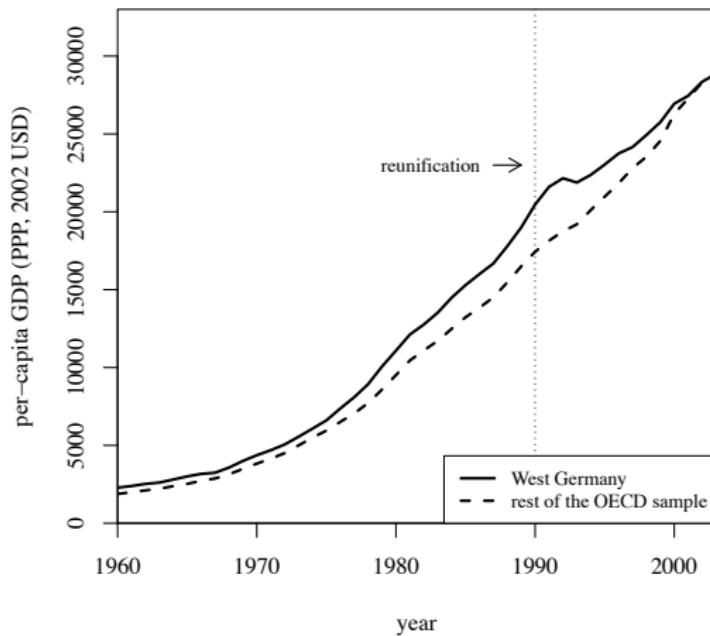
Ratio Post-Prop. 99 RMSPE to Pre-Prop. 99 RMSPE



The Application: The 1990 German Reunification

- Cross-country regressions are often criticized because they put side-by-side countries of very different characteristics
- “What do Thailand, the Dominican Republic, Zimbabwe, Greece and Bolivia have in common that merits their being put in the same regression analysis? Answer: For most purposes, nothing at all.” (Harberger 1987)
- The synthetic control method provides a data-driven procedure to select a comparison unit
- Application: the economic impact of the 1990 German reunification in West Germany
- Donor pool is restricted to 16 OECD countries

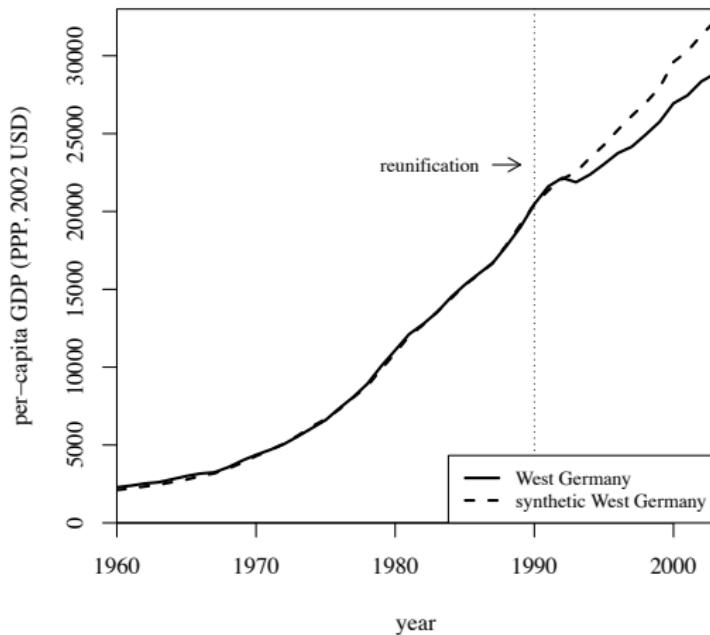
West Germany and the OECD sample



Covariate averages before 1990

	West Germany	Synthetic West Germany	OECD Sample
GDP per-capita	15808.9	15800.9	8021.1
Trade openness	56.8	56.9	31.9
Inflation rate	2.6	3.5	7.4
Industry share	34.5	34.4	34.2
Schooling	55.5	55.2	44.1
Investment rate	27.0	27.0	25.9

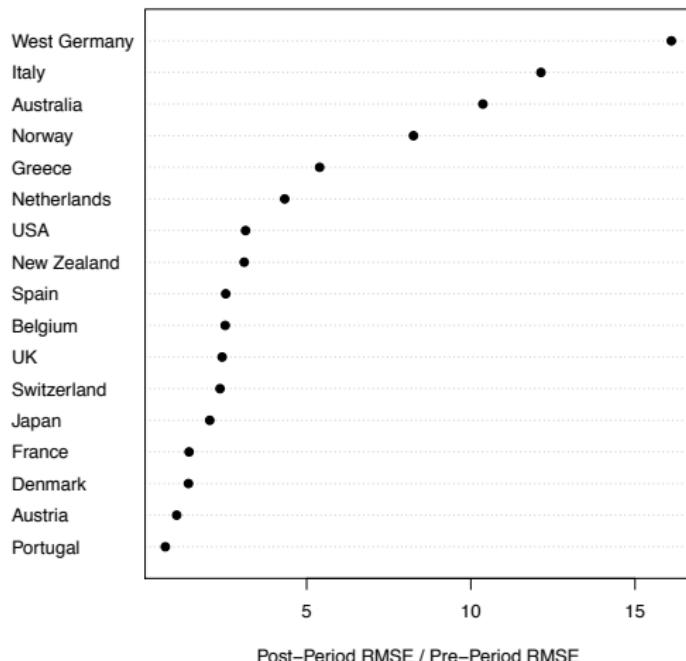
West Germany and synthetic West Germany



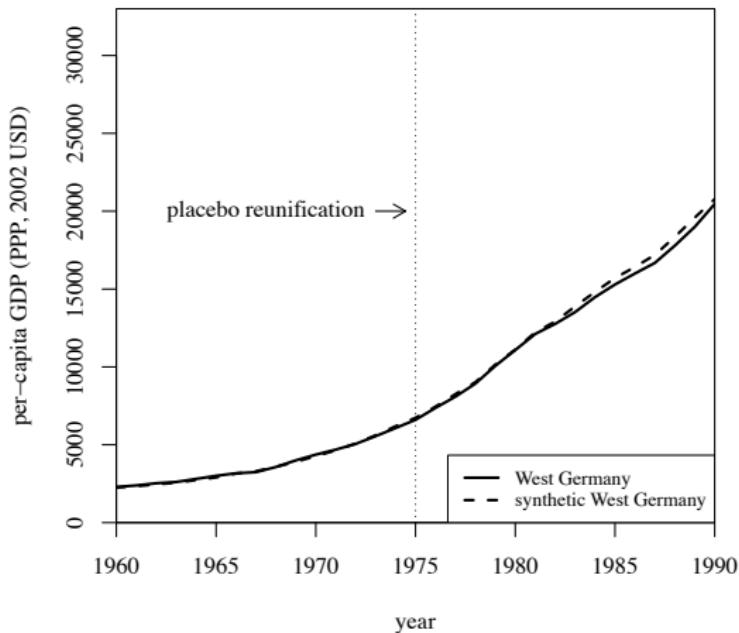
Country Weights in the Synthetic West Germany

Country	Weight	Country	Weight
Australia	0	Netherlands	0.10
Austria	0.42	New Zealand	0
Belgium	0	Norway	0
Denmark	0	Portugal	0
France	0	Spain	0
Greece	0	Switzerland	0.11
Italy	0	United Kingdom	0
Japan	0.16	United States	0.22

Post-Reunification to Pre-Reunification RMSPE Ratio



Placebo Reunification: 1975



Comparison to Regression

- Constructing a synthetic comparison as a linear combination of the untreated units with coefficients that sum to one may appear unusual
- However in fact a regression-based approach does exactly this, albeit in an implicit way
- In contrast to the synthetic control method, the regression approach does not restrict the coefficients of the linear combination that define the comparison unit to be in between zero and one, therefore allowing *extrapolation* outside the support of the data

Comparison to regression

- Let T_1 be the number of post-intervention periods and:

X_1 : $(k \times 1)$ -matrix of covariates for treated unit

X_0 : $(k \times J)$ -matrix of covariates for control units

Y_0 : $(T_1 \times J)$ -matrix of post-intervention outcomes for control units

- Let

$$\hat{B} = (X_0 X_0')^{-1} X_0 Y_0'$$

be the $(k \times T_1)$ matrix of regression coefficients of Y_0 on X_0

- That is, each column of \hat{B} contains the regression coefficients of Y_0 on X_0 for a post-intervention period.
- A regression-based counterfactual of the outcome for the treated unit in absence of the treatment is given by the $(T_1 \times 1)$ vector $\hat{B}' X_1$

Comparison to regression

- Notice that

$$\hat{B}' X_1 = Y_0 W^{\text{reg}}$$

where

$$W^{\text{reg}} = X_0' (X_0 X_0')^{-1} X_1$$

- As a result, the regression-based estimate of the counterfactual of interest is a linear combination of post-treatment outcomes for the untreated units, with weights W^{reg}
- Let ι be a $(J \times 1)$ vector of ones. The sum of the regression weights is $\iota' W^{\text{reg}}$. It can be proven that

$$\iota' W^{\text{reg}} = 1$$

Synthetic vs. Regression Weights

Country	Synthetic Control Weight	Regression Weight	Country	Synthetic Control Weight	Regression Weight
Australia	0	0.12	Netherlands	0.10	0.14
Austria	0.42	0.26	New Zealand	0	0.12
Belgium	0	0	Norway	0	0.04
Denmark	0	0.08	Portugal	0	-0.08
France	0	0.04	Spain	0	-0.01
Greece	0	-0.09	Switzerland	0.11	0.05
Italy	0	-0.05	UK	0	0.06
Japan	0.16	0.19	USA	0.22	0.13

Final thoughts

- A good research design is one you are excited to tell people about – that's basically what characterizes *all* research designs, whether propensity score matching or regression discontinuity designs, in some respects
- Most important thing is to be honest with yourself and the reader
- Always check for covariate balance in everything that you do
- Causality is easy and hard. Don't get confused which is the hard part and which is the easy part. Don't get enamored by statistical modeling that obscures the identification problem from plain sight. Always understand what assumptions you *must* make, be clear which parameters you are and are not identifying, and don't be afraid of your answers.

Articles and Books on the Syllabus I