# Lecture notes for applied microeconometrics

## Contents

1 Syllabus Review, Quiz, Causal analysis, Stata simulation .......................... 5
   1.1 Stata simulation and bootstrapping ........................................ 5
   1.2 Variance, Covariance and Matrix Algebra review .......................... 6
   1.3 Matrix algebra representation of variances ............................... 6
   1.4 OLS review ........................................................................ 6
   1.5 Derivation of OLS in (basically) 1 step with (basically) 1 assumption .... 8
   1.6 Evaluating consistency of an estimator ..................................... 8
   1.7 Deriving standard errors of an estimator ................................... 9
      1.7.1 Numerical Example ....................................................... 10
   1.8 Matching STATA standard errors in MATA .................................. 10
      1.8.1 Calculating OLS standard error ..................................... 11
      1.8.2 Calculating robust standard errors .................................. 11
      1.8.3 Clustered standard errors ............................................ 11

2 Identifying variation ............................................................................. 12
   2.1 Conditional independence assumption ....................................... 12
   2.2 Evaluating Empirical Work ....................................................... 14
      2.2.1 OLS without controls .................................................... 15
      2.2.2 OLS with controls ......................................................... 15
      2.2.3 OLS with impressive controls ....................................... 15
      2.2.4 Random Assignment .................................................... 16
      2.2.5 Instrumental Variables .................................................. 16
   2.3 Fixed Effect .......................................................................... 17
   2.4 Means comparisons .................................................................. 18
   2.5 Concluding Remarks .............................................................. 18

3 OLS Bias ............................................................................................ 20
   3.1 A few random points ............................................................... 20
      3.1.1 FWL review ................................................................. 20
   3.2 Inconsistency in OLS ............................................................... 20
      3.2.1 Omitted Variable Bias (OVB) ........................................ 21
      3.2.2 “Solutions” to omitted variable problem ........................... 23
      3.2.3 Good controls, useless controls and bad controls .......... 23
   3.3 Simultaneity .......................................................................... 24
   3.4 Measurement Error ................................................................. 24
      3.4.1 Classical measurement error ......................................... 24
      3.4.2 Solving Measurement error using an instrument ............... 26
      3.4.3 Non-random measurement error ..................................... 27

4 Linearity Assumption in OLS, Propensity Score Matching and Weighting .... 29
   4.1 Linearity assumption in OLS .................................................... 29
   4.2 Exact Matching ...................................................................... 31
   4.3 Comparison of matching to regression ..................................... 33
   4.4 Propensity Score Matching ..................................................... 35
4.5 Propensity Score Matching Estimation .............................................. 36
  4.5.1 Kernel Matching ............................................................... 37
  4.5.2 Investigating Common Support ............................................... 37
  4.5.3 Regression or Matching? ...................................................... 38
4.6 Weighting Estimator ................................................................. 38
  4.6.1 Math behind weights .......................................................... 42

5 Best practices to “test” CIA ............................................................... 45
  5.1 Selection on Observables .......................................................... 45
  5.2 Best practices to explore plausibility of CIA ................................. 45
    5.2.1 Means comparison of treatment and control (balance test) .......... 45
    5.2.2 Test of balance for continuous treatments ................................ 46
    5.2.3 A more relaxed test of balance ........................................... 46
    5.2.4 Gradually add covariates to main regression ............................ 48
    5.2.5 Failing the CIA specification tests ....................................... 49
    5.2.6 Falsification tests ............................................................ 49
  5.3 The problem with OLS, matching and weighting ............................ 49
    5.3.1 Lalonde (1986) .................................................................. 50
  5.4 OLS is Conditional Means Comparison .......................................... 50

6 Instrumental Variables ....................................................................... 50
  6.1 External vs Internal Validity ......................................................... 50
  6.2 Instrumental Variables(IV) ......................................................... 51
    6.2.1 IV implementation .............................................................. 52
    6.2.2 2 stage least squares (2SLS) vs Indirect Least Squares (ILS) ...... 52
    6.2.3 IV Terminology .................................................................. 53
    6.2.4 IV assumptions ................................................................. 54
    6.2.5 Example: Random Assignment .............................................. 54
    6.2.6 Example: Random Assignment with imperfect compliance ........ 55
    6.2.7 Intent to Treat (ITT) ............................................................ 56
    6.2.8 Recovering the LATE from the ITT. ........................................ 57
    6.2.9 Types of people in IV analyses .............................................. 58
    6.2.10 Example: (Angrist 1990) Veterans Draft Lottery ..................... 59
    6.2.11 Split-sample IV ................................................................. 62
    6.2.12 Other covariates; continuous endogenous regressors .............. 62
    6.2.13 IV Implementation in Practice ............................................. 63

7 Instrumental Variables: Issues, Evaluation, and Testing ...................... 64
  7.1 The weak instrument problem ..................................................... 64
    7.1.1 1 endogenous regressor and 1 instrument ............................... 64
    7.1.2 Multiple endogenous regressors ......................................... 65
  7.2 How biased are bad instruments? ................................................. 66
  7.3 The LATE interpretation ............................................................. 67
  7.4 Testing your instruments ........................................................... 67
    7.4.1 Hausman Tests and Overidentification Tests ........................... 67
    7.4.2 Test whether Z predicts things that it shouldn’t predict ............ 67
    7.4.3 An invalid test: $Y = \gamma Z + \beta D + \varepsilon$ .......................... 67
  7.5 How I evaluate instruments ....................................................... 68
  7.6 How to find instruments ............................................................ 68
7.7 Best practices when writing IV based papers
   7.7.1 First stage best practices
   7.7.2 Run the reduced form
   7.7.3 Interpret the IV estimate results
7.8 Examples!
   7.8.1 You can only have 1 first stage!

8 Regression Discontinuity Design (RDD)
   8.1 Sharp RD
   8.2 Implementing RD
   8.3 Parametric Estimation
      8.3.1 Linear estimation
      8.3.2 Global Polynomial
   8.4 Parametric with excluding some observations.
   8.5 Non-parametric estimation
      8.5.1 Local Linear Regression
      8.5.2 Local Linear Regression with Kernel smoothing
      8.5.3 A needlessly complicated phrase: local linear regression with a rectangular kernel and bandwidth h
      8.5.4 Bandwidth selection
      8.5.5 Judging discontinuities in figures
   8.6 Fuzzy RD
   8.7 RD issues
   8.8 RD empirical testing
      8.8.1 Histogram test
      8.8.2 Covariate smoothness test
   8.9 RD is not a before and after comparison
   8.10 RD estimates a LATE
   8.11 Regression Kink Designs

9 Panel data
   9.1 Review: Dummy variables and FWL
   9.2 Between and Within Variation
   9.3 Different levels of fixed effects
   9.4 Multiple fixed effects
   9.5 Is there any variation left? (perfect colinearity)
      9.5.1 Firm by Year fe vs Firm and Year fe
      9.5.2 Too many fixed effects
      9.5.3 Observable controls
   9.6 Thinking about identifying variation: Example 2
      9.6.1 Isolating the clean variation
      9.6.2 Addressing remaining biases
   9.7 What is the identifying variation? Is it clean?
      9.7.1 Complex colinearity problems
   9.8 Classical measurement error in fixed effects
   9.9 Fixed effects vs Aggregation vs Random effects vs Pooled OLS
   9.10 Clustering Standard Errors
      9.10.1 Level of clustering
10 Differences-in-differences

10.1 Basic Idea in Algebraic Form .................................................. 108
  10.1.1 Before/After ............................................................... 108
10.2 Graphs ................................................................. 109
10.3 Testing Common Trend Assumption .......................................... 111
  10.3.1 Graphical test ............................................................ 111
10.4 Implementation through fixed effects ....................................... 112
10.5 Why DID fails ................................................................. 113
10.6 Multiple dimension DID (state-year fe) ...................................... 114
10.7 DID extensions: multi-valued treatments; covariates ...................... 115
  10.7.1 Including covariates ...................................................... 115
  10.7.2 Continuous treatments .................................................. 115
  10.7.3 Common Trends Test with many treatment and control groups .... 115
10.8 DIDID ................................................................. 116
  10.8.1 DIDID Implementation .................................................. 118
  10.8.2 DIDID example: Many tennessee residents without children lost public
        health care ................................................................. 119
10.9 DID is IV ................................................................. 120

11 Missing data

11.1 Types of missing data ...................................................... 122
11.2 Addressing missing data .................................................... 125
  11.2.1 Sample selection bias .................................................. 125
11.3 Addressing missing data .................................................... 126
  11.3.1 Quantile regression at the median .................................... 126
  11.3.2 Lee bounds ............................................................... 126
1 Syllabus Review, Quiz, Causal analysis, Stata simulation

Syllabus Review

Pretest

This is my third time teaching the course. Syllabus will probably change since my time estimates are probably bad.

1.1 Stata simulation and bootstrapping

In order to better understand methods, it’s useful to run computer simulations where you know the true data generating process.

Suppose I write in stata the following code:

```stata
#d;
clear
set obs 1000
gen x=rnormal(3,1)
gen e=rnormal(0,1)
gen y=2*x+e
```

What is the causal impact of \(x\) on \(y\) in this case? If I run "reg y x" will I get the correct answer?

Run it a few times and show that I get roughly the right answer. Note that I don’t get exactly the same thing every time even thought the data generating process is identical each time. The variation in the estimates of beta is what the standard error is trying to capture.

I can look at the distribution of the betas with the following code:

```stata
cap program drop ols
program define ols, rclass
clear
set obs 1000
gen x=rnormal(3,1)
gen e=rnormal(0,1)
gen y=2*x+e
reg y x
return scalar b=_b\[x\]
end
```

**run once to make sure it works
ols

**simulate
simulate beta_estimates=r(b),reps(500):ols

Class questions:

- What does the 500 represent and what does the 1000 represent?

- What would happen to the standard error estimate if I increased the 1000 to 2000? What if I increased the 500 to 5000?
OLS review

1.2 Variance, Covariance and Matrix Algebra review

All regression analyses relate to variances and covariances.

\[ \text{Variance}(X) = E[(X_i - \bar{X})(X_i - \bar{X})] = E(X_i^2) - \bar{X}^2 \]

- We subtract off \( \bar{X}^2 \) just for scaling purposes. The important part is \( E(X_i^2) \)
  * Often, \( \bar{X} = 0 \) in which case \( \text{Var}(X) = E(X_i^2) \).
  * Similarly if \( \bar{X} = 0 \) or \( \bar{Y} = 0 \) then \( \text{cov}(X,Y) = E(XY) \).

1.3 Matrix algebra representation of variances

Note that \( E(X_i^2) \) is simply the average \( X_i^2 \) which is \( \frac{\sum X_i X_i}{N} \)

Written out this is simply

\[
\frac{x_1 x_1 + x_2 x_2 + \ldots + x_n x_n}{N} \tag{1}
\]

In matrix form we can rewrite the top part of equation (1) in matrix form as:

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n
\end{bmatrix}
\]

If we define \( X \) as:

\[
X = \begin{bmatrix}
  x_1 \\
  x_2 \\
  \vdots \\
  x_n
\end{bmatrix}
\]

Then the equation for variance can be easily written in matrix form.

- If \( \bar{X} = 0 \) then the variance is, \( \frac{X'X}{N} \).
- Similarly, if \( X \) or \( Y \) has mean zero, the covariance is \( \frac{X'Y}{N} \).

1.4 OLS review

In the two variable case, we have

\[ Y = \beta_0 + \beta_1 X + \varepsilon \]

What is the consistent estimate of \( \beta \)?

Either:

\[ \hat{\beta} = \frac{\text{cov}(X,Y)}{\text{Var}(X,X)} \quad \text{or} \quad \hat{\beta} = (X'X)^{-1}X'Y \quad \text{(students answer)} \]

Note that \( (X'X)^{-1} \) corresponds to \( \text{var}(X,X) \) and \( X'Y \) corresponds to \( \text{cov}(X,Y) \)

Question: What about the \( N \) and the fact that \( X \) might not be zero?
Answer: The N’s cancel out from numerator and denominator and we include a constant in the regression which acts to de-mean everything.

Exercise: Write out a case with 4 observations, 1 covariate (x) and a constant and convince yourself that the matrix form matches the non-matrix form.

What assumptions do we need for \( \hat{\beta} \) to be a consistent estimate of \( \beta \)?

- \( \text{cov}(X, \varepsilon) = 0 \)

What about all the other assumptions? e.g.

1. No serial correlation
2. No heteroskedasticity
3. Normal error term
4. No colinearity
5. \( E[\varepsilon] = 0 \)

The main assumption is the engine and the wheels. Other assumptions are like windshield wipers.

Given that we don’t need the other assumptions for consistency, we might as well not make those assumptions. That said, it’s worth understanding which of those assumptions are plausible.

1. No serial correlation – Sometimes plausible for random sample, implausible for panel or time series. Implausible for grouped observations e.g. kids in classroom.
2. No heteroskedasticity – Almost definitely wrong for 2 distinct reasons.
   - (a) Small companies are more volatile than big companies.
   - (b) Even if variance of Y is constant over Xs, the linear functional form will create heteroskedastic error structure. E.g. If true functional form is quadratic, the variance of the errors will be smaller around where the true functional form intersects the linear approximation. OLS is linear approximation of a non-linear CEF.
3. Normal error term – This might be true, but be careful to avoid thinking of error term as a small normal thing. Error term distribution mimics Y.
4. No colinearity – This assumption means no PERFECT colinearity. If it’s just that x’s are highly correlated this doesn’t matter.
5. \( E[\varepsilon] = 0 \) – This assumption is free.
Error term normality (assumption 3)

Q. — What do we know about the residual?

A. — The residual distribution will look fairly similar to the distribution of Y, except it will be centered around zero.

**Important Point:** The residual has a lot of economic meaning. If the $R^2$ is small (almost always true), the residual is basically the same as the outcome. $\varepsilon$ is the part of Y that is unexplained. **Never think about the residual as a mean zero random normal small variance thing.**

How do I know that the other assumptions don’t matter?

1.5 Derivation of OLS in (basically) 1 step with (basically) 1 assumption

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Take covariance of either side

$$cov(Y, X) = cov(\beta_0 + \beta_1 X + \varepsilon, X)$$
$$= cov(\beta_0, X) + \beta_1 cov(X, X) + cov(\varepsilon, X)$$
$$= cov(\beta_0, X) + \beta_1 cov(X, X) + cov(\varepsilon, X)$$

(Note that $cov(\varepsilon, X) = 0$ and covariance with a constant is always zero.)

$$cov(Y, X) = \beta_1 var(X)$$

$$\frac{cov(Y, X)}{var(X)} = \hat{\beta}_1$$

In matrix algebra form, the same derivation is:

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

$$X'Y = (X'X)\beta + X'\varepsilon$$

(Multiply each side by $X$)

$$(X'X)^{-1}X'Y = \hat{\beta}$$

1.6 Evaluating consistency of an estimator

Steps to evaluate consistency:

1. Write out $\beta$ and plug in for the true $Y$ in terms of $X$ and $\varepsilon$.

2. Invoke assumption of $cov(x, \varepsilon) = 0$ (or $X'\varepsilon = 0$) and take expectations

**Consistency Proof**
1. 

\[ \hat{\beta} = (X'X)^{-1}X'Y \]
\[ \hat{\beta} = (X'X)^{-1}X'(X\beta + \epsilon) \] (Plug in for Y)
\[ \hat{\beta} = (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'\epsilon \]
\[ \hat{\beta} = \beta + (X'X)^{-1}X'\epsilon \]

2. In expectation \(X'\epsilon = 0\) since \(cov(X, \epsilon) = 0\). As a result, we have consistency \(E[\hat{\beta}] = \beta\)

1.7 Deriving standard errors of an estimator

Steps to derive standard error:

1. Write out \(\beta\) and plug in for the true Y in terms of X and \(\epsilon\). (same as first step of consistency proof)
2. Take the variance of \(\hat{\beta} - \beta\)

Variance Derivation

\[ \hat{\beta} = \beta + (X'X)^{-1}X'\epsilon \]
\[ var(\hat{\beta} - \beta) = var((X'X)^{-1}X'\epsilon) \]

How do we calculate \(var((X'X)^{-1}X'\epsilon)\)?

- We know that \((X'X)^{-1}X'\epsilon\) has mean zero since \(E[X'\epsilon] = 0\).
- Recall that \(var(z) = E[z^2] - 0^2\)
- In matrix form, we have \(ZZ'\) if we want the full variance covariance matrix.

This means that

\[ var((X'X)^{-1}X'\epsilon) = (X'X)^{-1}X'\epsilon \epsilon'X(X'X)^{-1} \]

Let's review what this means.

- \(\hat{\beta}\) is a vector that is \(k \times 1\)
- \(V = (X'X)^{-1}X'\epsilon \epsilon'X(X'X)^{-1}\) is a \(k \times k\) matrix
  * The diagonal elements of V are the variances of \(\beta_1, \beta_2...\beta_k\).
  * The off-diagonal elements of V are the covariances between \(\beta_1, \beta_2...\beta_k\).
1.7.1 Numerical Example

To calculate $V$ we need actual numbers. $X$ is directly observed, but $\varepsilon$ is not.

$$
X = \begin{bmatrix}
1 & 36 & 103 & 1 \\
1 & 28 & 108 & 0 \\
1 & 34 & 95 & 0 \\
1 & 19 & 82 & 1 \\
1 & 41 & 101 & 1 \\
\end{bmatrix}
\begin{array}{c}
\text{cons} \\
\text{age} \\
\text{iq} \\
\text{female} \\
\end{array}
$$

We estimate $\varepsilon$ using $e$, the vector of residuals.

We calculate $e$ using $e = Y - X\hat{\beta}$ where $\hat{\beta}$ is defined in terms of $X$ and $Y$

$$
e = \begin{bmatrix}
8 \\
-5 \\
0 \\
-4 \\
1 \\
\end{bmatrix} \quad ee' = \begin{bmatrix} 8 \\
-5 \\
0 \\
-4 \\
1 \\
\end{bmatrix} = \begin{bmatrix} 64 & -40 & 0 & -32 & 8 \\
-40 & 25 & 0 & 20 & -5 \\
0 & 0 & 0 & 0 & 0 \\
-32 & 20 & 0 & 16 & -4 \\
8 & -5 & 0 & -4 & 1 \\
\end{bmatrix}
$$

We calculate $\hat{\beta}$ and $\text{var}(\hat{\beta})$ simply by plugging and chugging.

1.8 Matching STATA standard errors in MATA

Sysuse auto and show that I can calculate $\hat{\beta}$ using mata. Show that $\text{var}(\hat{\beta})$ is not matched.

In MATA, I failed to match the STATA standard errors. What went wrong?

Note that the variance-covariance matrix is based on $V$, but $V$ cannot be consistently estimated because there are too many unknowns in $ee'$. It is not possible to consistently estimate the standard errors without additional assumptions beyond $\text{cov}(X, \varepsilon) = 0$.

We always start with:

$$
\text{var}(X'X)^{-1}X'\varepsilon = (X'X)^{-1}X'ee'X(X'X)^{-1}
$$

OLS, huber-white/robust, cluster each make different assumptions about $ee'$.

**OLS assumes:**

- homoskedasticity
- no autocorrelation

To match the OLS standard errors, you must impose these assumptions on $ee'$ and also use the $N/(N-K)$ DOF correction.

**Huber-white (robust) standard errors assumes:**

- no autocorrelation
clustered standard errors assumes:

- no autocorrelation across clusters, but allows for autocorrelation within clusters.
- clustered standard errors are only consistently estimated as the number of clusters tends towards infinity ($k \to \infty$).

1.8.1 Calculating OLS standard error

We leave the $(X'X)^{-1}X'$ alone and make assumptions regarding $\varepsilon\varepsilon'$. We assume:

$$
\varepsilon\varepsilon' = \begin{bmatrix}
64 & -40 & 0 & -32 & 8 \\
-40 & 25 & 0 & 20 & -5 \\
0 & 0 & 0 & 0 & 0 \\
-32 & 20 & 0 & 16 & -4 \\
8 & -5 & 0 & -4 & 1 \\
\end{bmatrix} \rightarrow \begin{bmatrix}
\frac{64 + 25 + 0 + 16 + 1}{5} & 0 & 0 & 0 & 0 \\
0 & \frac{64 + 25 + 0 + 16 + 1}{5} & 0 & 0 & 0 \\
0 & 0 & \frac{64 + 25 + 0 + 16 + 1}{5} & 0 & 0 \\
0 & 0 & 0 & \frac{64 + 25 + 0 + 16 + 1}{5} & 0 \\
0 & 0 & 0 & 0 & \frac{64 + 25 + 0 + 16 + 1}{5} \\
\end{bmatrix}
$$

The standard errors are the square roots of the diagonal elements of:

$$
\text{var}(\hat{\beta} - \beta) = (X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}
$$

1.8.2 Calculating robust standard errors

Robust standard errors removes homoskedasticity assumption.

1.8.3 Clustered standard errors

Clustered standard errors assumes $\varepsilon\varepsilon'$ is block diagonal. If the first cluster had 2 observations and the second cluster had 3 observations, it would take the form:

$$
\begin{bmatrix}
\sigma_1^2 & \sigma_{12} & 0 & 0 & 0 \\
\sigma_{21} & \sigma_2^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\
0 & 0 & \sigma_{43} & \sigma_4^2 & \sigma_{45} \\
0 & 0 & \sigma_{53} & \sigma_{54} & \sigma_5^2 \\
\end{bmatrix}
$$
2 Identifying variation

*Today is an overview of many different topics. I will be sloppy!*

Usually, we want to study the impact of a variable on an outcome.

I call the variable X and the outcome Y.

**Undergrad thought process: consider everything that may cause Y**

Example:

**Q.** — Does smoking increase mortality?

**A.** — Undergrad method: Start with Mortality = $β_{\text{Smoking}} + ε$. Make a list of all the other determinants of Y and control for those.

This is the wrong way to address the question of does X cause Y. Even though Y is the outcome, we don’t care particularly about it’s causes.

**You should no longer write papers about “the causes of Y”**.

2.1 **Conditional independence assumption**

$β$ will be consistent so long as:

$$E[Y_i | X_i = 1; \text{covariates}] = E[Y_i | X_i = 0; \text{covariates}]$$

In words: We don’t need to control for everything, we just need enough controls so that potential outcomes don’t depend on $X_i$.

**Why should you focus on X, rather than Y?**

**In order to study the impact of X on Y, there must be variation in X!**

Example: What is the impact of the number of kids you have on income?

**Q.** — What happens in our regression if there is not variation in X? What if there is no variation in Y?

Assuming there is variation in X, (e.g. some people have 2 kids and other’s have 3) we can study the correlation of Y,X.

$$Y = Xβ + ε$$

Can we interpret $β$ as the causal impact?

No! If any factors cause X and Y or if Y causes X, then $β$ will be wrong.

If someone is trying to prove that X causes Y, most of the thought needs to be about what causes X. When considering what to control for in a regression, the main goal is to control for the factors that cause X and also cause Y.

What causes the variation in X?
• If the variation in X is partly caused by factors that also cause Y, this is an omitted variable problem.

• If the variation in X is partly caused by Y, this is simultaneity.

Usually, there will be multiple reasons why X varies:

Ex. Why is their variation in the number of kids in different families?

1. Some people have a stronger preference for kids.
2. Some people don’t like working, so they have more kids.
3. Some people belong to a religion that encourages having many kids.
4. Some people have twins and end up with an extra kid as a result.
5. Some people live in areas that are not as kid friendly.
6. There may be cultural shifts over time such that families from the 1970s have more/fewer kids compared to families from the 1990s.
7. China introduced a 1 child policy in some regions, but certain regions were exempt.
8. ...many other factors.

Above I have listed 6 of the reasons why there is variation in X. This correlates to 6 different types of variation.

**Informal definition**: “bad variation” or “dirty variation”: variation in X that is either caused by Y or related to factors which cause Y directly.

**Informal definition**: “exogenous variation” or “clean variation”: variation in X that is not caused by anything related to Y.

I could study the question of how is income impacted by number of kids using many different identification strategies.

Definition: **Identification Strategy**: The technique the researcher uses to try to identify a causal estimate.

Each identification strategy uses a different portion of the variation in X. The goal is to use “clean variation”.

When I say “uses a portion of variation” this means “correlates this portion of variation of X with Y”

Identification strategies possible for this question include:

• OLS with no controls → Uses ALL of the variation
• OLS controlling for various factors → Uses ALL of the variation except for variation caused by the variables controlled for.

• Instrumental Variables → Uses ONLY variation caused by the instrument

• Regression Discontinuity → Uses ONLY variation caused by the discontinuity rule

• Fixed effects → Uses ALL variation except for variation that occurs across units.

• Aggregation → Uses ALL variation except for variation that occurs within units.

• Difference-in-Difference → Uses ONLY variation created by a policy.

• others...

Definition: **Identifying Variation**: The variation used by the identification strategy.

**Identification strategies** There are two different types of identification strategies.

1. Control for as much of the dirty variation as you can. Hope that the variation left over is clean.
   * Researchers don’t really understand the source of identifying variation in this type of study.
   * Researchers hope that the rest of the variation in X is “random” or unrelated to Y.

2. Only use a particular source of variation
   * Researchers know the source of identifying variation in this type of study.
   * Researcher needs to convince reader that the identifying variation only impacts X and not Y.

### 2.2 Evaluating Empirical Work

**Evaluating empirical work**

• Figure out what is the identifying variation.

• Do you think the identifying variation is *completely* clean.

• If any part of the identifying variation is dirty, then the estimator is biased. You should then think about the direction of the bias.

**Example.**

1. Impact of the number of kids had on whether you work. Possible identification strategies:

   1. OLS without controls
   2. OLS with basic controls
   3. OLS with impressive controls
4. Random-Assignment

5. IV

6. Person Fixed Effects (dummy variables for each woman)

7. Several others.

### 2.2.1 OLS without controls

1. If we run a simple OLS regression of Y on X, this uses ALL of the variation and will be biased. (why?)

### 2.2.2 OLS with controls

2. If we control for religion, education level, region dummies, time dummies the remaining identifying variation is based on the variation not crossed out:

   1. Some people have a stronger preference for kids.
   2. Some people don’t like working, so they have more kids.
   3. Some people belong to a religion that encourages having many kids.
   4. Some people have twins and end up with an extra kid as a result.
   5. Some people live in areas that are not as kid friendly.
   6. There may be cultural shifts over time such that families from the 1970s have more/fewer kids compared to families from the 1990s.
   7. China introduced a 1-child policy in some regions, but certain regions were exempt.
   8. ....many other factors

   *If ANY of the variation used is dirty, then the estimate will be biased.*

   IMPORTANT: Bias is not dichotomous. It’s possible that an estimate could be biased, but perhaps it’s not likely to be very biased. This can be good enough sometimes.

### 2.2.3 OLS with impressive controls

3. If we control for survey measures of kid preference, opinions on work, twinning rates, religion, education level, region dummies, time dummies the remaining identifying variation is based only on things we haven’t thought of:

   1. Some people have a stronger preference for kids.
   2. Some people don’t like working, so they have more kids.
   3. Some people belong to a religion that encourages having many kids.
   4. Some people have twins and end up with an extra kid as a result.
5. Some people live in areas that are not as kid friendly.

6. There may be cultural shifts over time such that families from the 1970s have more/fewer kids compared to families from the 1990s.

7. China introduced a 1-child policy in some regions, but certain regions were exempt.

8. ....other factors?

Will this regression be biased? If it’s still biased, is it better than regular OLS? DISCUSS WITH CLASS

Some key points

- If nobody can think of other causes of X that also cause Y, then OLS can be good enough. This is rare though!

- Suppose we could control for every single factor. Then there would be no bias, but there would also be no variation left in X!

2.2.4 Random Assignment

4. Suppose we could do random assignment for 100 families. If we could randomly assign the number of kids, then we would control for all factors except the random assignment. The identifying variation would be only the random assignment which by construction doesn’t impact Y.

2.2.5 Instrumental Variables

5. Usually, we can’t do random assignment. If we can think of another cause of X that doesn’t cause Y directly, this is called an instrument. If we can think of an Instrument, then we can isolate the variation in X caused by the instrument and use that identifying variation to estimate the impact of X on Y. Usually, instruments are very very difficult to find.

I’ve chosen the example regarding kids because there is a pretty good instrument. What is it?

Instrument: Twins!

Discuss as class: Is the variation caused by Twins clean? Why or why not? How do we use only one portion of variation? How do we use only the variation caused by Z?

1. Regress X on Z.

2. Predict X

3. The predicted Xs, will vary much less than actual X. All the variation in predicted Xs is variation caused by Z.
Isolating Clean Variation It is possible to combine multiple identification strategies. Example.

When we use the variation caused by twins we have:

Identifying variation

1. Some people have twins and end up with an extra kid as a result.
   a. Extra kids caused by natural random twins.
   b. Extra kids caused by fertility drugs, usually taken by older women.

   It might be that part of the variation caused by twins is dirty, since older women might be more likely to work for other reasons.
   
   This is no problem! Just combine the IV method with the controlling method. We can use just the variation caused by Z, holding fixed variation caused by other factors. If we control for age and education of the woman, this makes the identifying variation more clean. Ideally, we could control for whether the woman took fertility drugs and only use variation caused by naturally occurring twins.

2.3 Fixed Effect

6. How did hours worked change after having an extra child. In a 2-period context, this is the same as the before-after estimator, or the first difference estimator. Under the assumption that people don’t change over time, the identifying variation is based on the things not crossed out below:

   1. Some people have a stronger preference for kids.
   2. Some people don’t like working, so they have more kids.
   3. Some people belong to a religion that encourages having many kids.
   4. Some people have twins and end up with an extra kid as a result.
   5. Some people live in areas that are not as kid friendly.
   6. There may be cultural shifts over time such that families from the 1970s have more/fewer kids compared to families from the 1990s.
   7. China introduced a 1 child policy in some regions, but certain regions were exempt.
   8. ....many other factors****THESE ARE NOW PARTIALLY CONTROLLED FOR

The fixed effect estimator is very powerful because it controls for observed and unobserved factors. When we include individual fixed effect, the identifying variation is only the WITHIN person variation. Fixed effects is also called the "WITHIN" estimator.
Example of dirty within variation: Suppose age makes people work more. This would be correlated with the identifying "Within" variation and would directly cause Y.

The BETWEEN variation is based on comparing across people. The WITHIN variation is based on comparing within each individual.

2.4 Means comparisons

Most identification strategies stem from some comparison of means. In order to evaluate an identification strategy, you need to understand the means comparison being performed.

1. OLS without controls – literally compares means of treated and controls
2. OLS with basic controls – compares means conditional on covariates
3. IV – loosely compares means of instrumented and not-instrumented
4. Person Fixed Effects – compares means of each person to their average
5. DID – difference in changes of means across two locations.
6. RD – compares means on either side of cutoff.

2.5 Concluding Remarks

Regression relates variation in X to variation in Y.

If X goes up and Y goes up for the same observation, this means that X increases Y.

Whenever X changes, this is called variation in X. Bivariate regression relates all the variation in X to variation in Y.

Standard errors are large in 1 of two cases:

1. Not a lot of variation in X.
   - If you are looking at impact of height on weight, if almost everyone in the sample is the same height, you will base all your estimates on those with different heights and so you don’t have much confidence in the estimates
2. A lot of unexplained variation in Y
   - Suppose on average tall people weigh more. If there are many exceptions to this rule, the standard errors will rise.

Controlling for covariates tells OLS which part of the X variation not to use. The identifying variation is the part left over. The OLS estimates are based on relating this left over variation to Y.

Q. —— What happens if the controls are correlated with X but not Y?
A. — *Standard errors will rise, estimates are unaffected*

Q. — What happens if the controls are correlated with Y but not X?

A. — *Standard errors fall, estimates are unaffected*

If there isn’t much identifying variation, the standard errors will be big. If the identifying variation isn’t clean, we get bias.

IV isolates one part of the X variation and only uses that. Since there won’t be a lot of X variation used, IV will tend to have large standard errors.
3 OLS Bias

3.1 A few random points

Means comparison is the simplest type of empirical method. It is simply the comparison of average outcomes for two groups.

We can do means comparison either by calculating the means, or by running OLS.

Suppose I want to calculate the gap in weight between males and females.

I can calculate $W_m - W_f$ or alternatively I can run a regression $W_i = \alpha + \beta_{Female} + \epsilon$

To see that these are equivalent, consider an example for the 1st, 3rd and 5th observation are women and there are 5 total observations. In this case:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

Write out all the matrix multiplication of $(X'X)^{-1}X'Y$ until you are convinced that $\hat{\beta}$ will be equal to the difference in average outcomes. $\alpha$ will give the average height for males.

Once we include controls, the coefficient on a dummy variable is a means comparison, conditional on covariates.

3.1.1 FWL review

Recall that by FWL, we can do multivariate regression in 2 steps. Suppose we want to run $Wage = \alpha + \beta_{Female} + \beta_2 Education + \epsilon$.

1. $Education = \beta_2 Female + \mu$
2. $Wage = \beta_2 \mu + \nu$

The reason that this works is that $\mu$ is basically still education, it’s just demeaned by Female.

Q. — Why didn’t we have to residualize the outcome variable?

A. — Consistency is all about the identifying variation in $X$. There will always be lots of “other things” that cause $Y$ that we ignore.

Thinking about FWL as just residualizing $X$ conforms to the idea of conditional independence assumption (CIA). If we assume CIA, this means that once we control for all the covariates, the potential outcomes are uncorrelated with $X$. This means that we are thinking of conditional $X$ as coming from a randomized experiment.

3.2 Inconsistency in OLS

Inconsistency comes from $X \not\perp \epsilon$. This is also known as $X$ being endogenous. There are three possible sources for endogeneity.
Sources of endogeneity

1. Omitted Variable
   (a) If a variable that is related to X and Y is not controlled for, then the error term will be correlated with X.

2. Simultaneity
   (a) If Y causes X, then ε causes X since ε is basically just Y.

3. Measurement Error in an Independent Variable
   (a) If X is measured with error then this can be thought of a special type of omitted variable bias. The omitted variable is the inverse of the measurement error.

Other problems

1. Mis-specification of the model
   (a) If you model a quadratic relationship with a linear term, the regression results may be misleading.
   (b) When the functional form is mis-specified, you still get a consistent estimate of the linear relationship. Linear relationships are often of interest since it tells you “on average, does increasing X make Y increase?”

3.2.1 Omitted Variable Bias (OVB)

This is the most common cause of inconsistency. In most cases, there are no good ways to address the omitted variable problem, but it’s important to understand how an omitted variable biases your estimates. Most identification strategies aim to solve the OVB problem.

The direction of the bias

An omitted variable, Q is problematic if and only if it satisfies two properties.

1. The variable Q is correlated with X
2. The variable Q directly impacts Y

If we understand how Q relates to X and Y, then it is possible to understand the bias created by leaving out Q.

- If X and Q are positively correlated the impact of Q on Y gets attributed to X.
- if X and Q are negatively correlated, then the opposite of the impact of Q on Y gets attributed to X.
Example

Suppose we are interested in the impact of education on wages. We could run the regression:

\[ W = \hat{\beta}E + \epsilon \]  

(2)

In this regression, there is an omitted variable of motivation. For simplicity, I’ll assume that this motivation is the only omitted variable.

- Motivation causes you to get more education.
- Motivation directly causes increased wages.

Q. — If we estimate equation (2) the estimate of \( \hat{\beta} \) will be biased because of OVB, but what direction will the bias be?

A. — Since motivation and education are positively correlated, the impact of motivation will be attributed to education. This will make education seem more valuable than it actually is and \( \hat{\beta} \) will be too large.

We can confirm this intuition with a derivation of consistency.

Recall: Steps to evaluate consistency:
1. Write out \( \beta \) and plug in for the true \( Y \) in terms of \( X \) and \( \epsilon \).
2. Invoke assumption of \( \text{cov}(x, \epsilon) = 0 \) (or \( X' \epsilon = 0 \)) and take expectations

What is the expected value of \( \hat{\beta} \) when estimating equation (2)?

Note that we will run the regression given by equation (2), but the true model is \( W = \beta E + \gamma M + \mu \). Note that, \( E \) is correlated with \( \epsilon \), but \( E \) is uncorrelated with \( \mu \). \( \gamma \) is the impact of motivation on wages.

1. Begin with our estimate of \( \hat{\beta} \).

\[
\hat{\beta} = (E'E)^{-1}E'W \\
\hat{\beta} = (E'E)^{-1}E'(\beta E + \gamma M + \mu) \quad \text{(Plug in for } W \text{ using truth)} \\
\hat{\beta} = (E'E)^{-1}E'\beta E + (E'E)^{-1}E'\gamma M + (E'E)^{-1}E\mu \\
\hat{\beta} = \beta + (E'E)^{-1}E'M\gamma + (E'E)^{-1}E\mu
\]

2. In expectation \( (E'E)^{-1}E\mu = 0 \) since \( E \not\perp \mu \). As a result, we have

\[
\hat{\beta} = \beta + (E'E)^{-1}E'M\gamma
\]

In this case, \( \hat{\beta} \) is not a consistent estimate of \( \beta \). The size and direction of the bias is given by: \( (E'E)^{-1}E'M\gamma \).
We can rewrite this bias term in a simpler form, by noticing that \((E'E)^{-1}E'M\) is of the form 
\((XX')^{-1}X'Y\). \((E'E)^{-1}E'M\) is thus the regression coefficient from a regression of motivation on education. For notational simplicity, define this coefficient as \(\delta\) where \(\delta\) comes from the regression \(M = \delta E + \nu\).

Returning to the bias formula, we have:

\[
\hat{\beta} = \beta + \delta \gamma
\]  

(3)

This formula is incredibly useful and coincides perfectly with the intuitive discussion regarding omitted variable bias. In this case, \(\delta\) is positive since people with more motivation have more education. \(\gamma\) is also positive since motivation increases income. This means that the bias is positive so our estimated \(\hat{\beta}\) is too large.

Referring back to the intuitive discussion at the beginning of the OVB section, make sure you fully understand how the OVB mathematical formula given by equation (3) matches the language re-copied below:

- If \(X\) and \(Q\) are positively correlated the impact of \(Q\) on \(Y\) gets attributed to \(X\).
- If \(X\) and \(Q\) are negatively correlated, then the opposite of the impact of \(Q\) on \(Y\) gets attributed to \(X\).

If you are commenting on a paper, don’t leave a comment as “there is an omitted variable”. You should always say something like, “there is an omitted variable and this will bias your estimate downward”. In writing your own papers, it is frequently useful to provide a discussion of likely direction of bias. Naturally, it is best if the coefficient isn’t biased at all, but if you can argue that your estimated \(\hat{\beta}\) is a lower/upper bound, this can be useful.

### 3.2.2 “Solutions” to omitted variable problem

Theoretically, one solution to OVB is to control for a proxy of the omitted variable. For example, if we are concerned that motivation biases our estimates, we can control for a survey question regarding motivation. Similarly, if we are concerned that ability is an omitted variable, we could control for SAT scores. In some cases controlling for a proxy of the omitted factor will get us a better estimate of the impact of \(X\) on \(Y\). Occasionally, the inclusion of a new control is a sufficient contribution to be publishable. Most of the time though, it is necessary to deal with OVB in a more direct way by using an identification strategy other than simple OLS. The rest of the course will discuss potential empirical strategies to use when omitted variables are a big problem.

### 3.2.3 Good controls, useless controls and bad controls

1. Good controls are variables that you include in the regression in order to kill dirty identifying variation.
2. Useless controls are variables that you include in the regression that aren’t correlated with \(X\). Often it’s worthwhile to include these controls to reduce your standard errors, but fundamentally, these controls don’t do that much.
3. Bad controls are variables that are caused by \(X\). Controlling for these variables is called “over controlling”.

23
Example
Suppose we want to understand the extent of the gender gap in wages.

If we regress wages on gender, we cannot interpret the coefficient causally because there are many omitted factors.

I am considering controlling for: Work experience, education, occupation, the unemployment rate, zip code,

- Controlling for experience and education may be examples of good controls. These controls are correlated with gender and also cause income. These could be considered bad controls if we think gender causes education or experience.

- Controlling for the unemployment rate in your area is probably a useless control. It definitely impacts wages, but it isn’t related to gender unless certain areas have more/less women.

- Controlling for occupation is a bad control since occupation could be caused by gender. Controlling for occupation is over controlling.

3.3 Simultaneity
When a variable is omitted, it is theoretically possible to add controls to address the issue. If we have simultaneity (Y causes X), we can’t take this approach because you can’t control for Y since it is the dependent variable. Essentially the only way to credibly address simultaneity is by finding an instrument. Recall that an instrument is an exogenous variable that causes X, but has no direct effect on Y.

3.4 Measurement Error
Relative to OVB, measurement error is a relatively minor concern. That said, it is often more of a problem than researchers acknowledge and it is useful to understand it’s consequences.

3.4.1 Classical measurement error
Classical measurement error is random, with mean zero. The defining characteristic here is the randomness i.e. it is uncorrelated with X and Y.

Example: It is hard to think of a realistic assumption where measurement error is perfectly random. The most basic example is simply errors made in the coding of variables or data entry errors, but these are unlikely to generate truly random error.

Intuitively, classical measurement error will make us more likely to find a zero effect because the regression will sometimes observe changes in X (caused by errors) that don’t correspond to a change in Y. This is called attenuation bias.

Suppose we have measured the true variable \( X_T \) with error and we observe only X. We have:

\[
X = X_T + \nu
\]

Because we assume the measurement error is random, we assume that \( \nu \perp X_T \) and \( \nu \perp Y \).
Q. —- Is ν ⊥ X?

A. —- No! if ν is large and positive, X will be larger. In fact, a 1 unit change in ν causes a 1 unit change in X!

Assuming the true model is \( Y = X\beta + \varepsilon \) and the model estimated is \( Y = X\hat{\beta} + \mu \).

Plugging in for \( X_T \), the model that we end up estimating is:

\[
\begin{align*}
    y &= X_T\beta + \varepsilon \\
    y &= \beta(X - \nu) + \epsilon \\
    &= \beta X + \left(\varepsilon - \beta \nu\right) \\
    &= \beta X + \text{error term}
\end{align*}
\]

When we run the regression of \( Y = X\hat{\beta} + \mu \), the error term is \( \varepsilon - \beta \nu \).

This means that X is correlated with the error term and we will have bias. In a sense, the omitted variable here is the negative measurement error ν.

Q. —- What will the direction of the bias be?

A. —- To evaluate the direction of OVB, we need to know the correlation between X and the omitted variable and the correlation between the omitted variable and Y. The correlation between X and −ν is always negative since X and ν are positively correlated. What is the impact of ν on Y? It is tempting to think the impact is zero since it is just measurement error. However, according to equation (4), when ν increases, Y changes by −β. The reason this is happening is that once we control for the level of X, increasing ν is the same as decreasing \( X_T \).

- \( \text{cov}(X, -\nu) < 0 \)
- \( \text{cov}(Y, -\nu) = \text{sign of } \beta \)

If β > 0 then β is biased down, if β < 0 then β is biased up. Either way, β is biased towards zero.

Notice that measurement error is like an omitted variable that always will bias \( \hat{\beta} \) in the opposite direction of β.

To determine the magnitude of bias caused by measurement error we evaluate the consistency using the same steps as usual. I will do this in the bivariate case.

Note that the \( \hat{\beta} \) estimate is in terms of X, not \( X_T \).
\[
\hat{\beta} = \frac{\text{cov}(X, Y)}{\text{var}(X)}
\]
\[
= \frac{\text{cov}(X, \beta X_T + \varepsilon)}{\text{var}(X)}
\]
(Plug in for Y using truth (including the true X))
\[
= \frac{\beta \text{cov}(X, X_T) + \text{cov}(X, \varepsilon)}{\text{var}(X)}
\]
(by assumption, \(\text{cov}(X, \varepsilon) = 0\))
\[
= \frac{\beta \text{cov}(X, X_T)}{\text{var}(X)}
\]
(We can plug in for measured X in terms of true \(X_T\) and the error term.)
\[
= \frac{\beta \text{cov}(X_T + \nu, X_T)}{\text{var}(X_T + \nu)}
\]
(This simplifies because \(\text{cov}(\nu, X_T) = 0\))
\[
= \frac{\beta \text{var}(X_T)}{\text{var}(X_T) + \text{var}(\nu)}
\]

The final result is that \(\hat{\beta}\) will be a scaled down version of \(\beta\) since we know that \(\frac{\sigma_{X_T}}{\sigma_{X_T} + \sigma_{\nu}} < 1\).
The extent of the attenuation bias depends on the relative size of the variance of \(X\) and \(\nu\).
Notice that if \(\sigma_{\nu}\) is very large, this will introduce a lot of attenuation which is an intuitive result. The attenuation multiplier here is very similar to a standard signal to noise ratio.

**Measurement error in Y vs measurement error in X**

Q. --- What will happen if we have measurement error in Y?

A. --- *We will not get bias*

Measurement error in Y is fundamentally different than measurement error in X.

- When \(X\) is measured with error, the regression sometimes sees \(X\) move and sees no corresponding change in \(Y\) so it concludes that \(X\) doesn’t cause \(Y\) as strongly. The estimated coefficient will thus be attenuated.
- With \(X\) measured with error, part of the identifying variation is due to measurement error.
- When \(Y\) is measured with error, the regression sometimes sees \(Y\) move for no reason it can understand. This error in \(Y\) simply ends up in the residual. Even if \(Y\) is measured with error, it remains the case that when \(X\) causes true \(Y\) to increase by 100, measured \(Y\) will increase by 100 on average.

Measurement error in Y just increases the total variability of the residual and makes it harder to detect whether \(X\) causes \(Y\). The estimated coefficient will still be consistent, but the standard errors will become larger.

### 3.4.2 Solving Measurement error using an instrument

The basic problem with measurement error is that part of the identifying variation is due to measurement error that doesn’t actually change True \(X\). The solution to this problem, is simple: **Don’t use that part of identifying variation.**
Q. —- How can we remove the measurement error from the identifying variation?

A. —- If we had a measure of the measurement error, we could control for the measurement error in OLS and this would eliminate that portion of the identifying variation. In practice, this is unrealistic because we don’t know the measurement error. The solution is therefore to use and instrument to focus on ONLY the variation caused by the instrument as identifying variation. This solves the measurement error problem because we are no longer using the noisy variation for identification.

In many cases people just note that their estimates may be attenuated and they don’t address the measurement error.

3.4.3 Non-random measurement error

Q. —- What is an example of non-random measurement error?

A. —- Most types of measurement error are probably non-random. Example: Rich people under-report income and poor people over report income. The direction of the measurement error in income is correlated with your level of income.

Q. —- What is an example of non-normal measurement error?

A. —- Any measurement error in a dummy e.g. smoking status.

Nonrandom measurement error can be a much more serious problem because the direction of bias is sometimes ambiguous.

To evaluate the direction of bias, we do the same steps as earlier to get to:

\[ \hat{\beta} = \beta \frac{\text{cov}(X_T + v, X_T)}{\text{var}(X_T + v)} = \beta \frac{\text{Var}(X_T) + \text{Cov}(v, X_T)}{\text{var}(X_T) + \text{var}(v) + 2\text{Cov}(v, X_T)} \]

In this case, the multiplier on \( \beta \) could be larger or smaller than 1.

Q. —- Under what conditions will \( \beta \) be biased away from zero? Under what conditions will \( \beta \) be the wrong sign?

A. —- If \( \text{Var}(X_T) + \text{Cov}(v, X_T) > \text{var}(X_T) + \text{var}(v) + 2\text{Cov}(v, X_T) \) we get anti-attenuation.

\[ \text{Var}(X_T) + \text{Cov}(v, X_T) > \text{var}(X_T) + \text{var}(v) + 2\text{Cov}(v, X_T) \]

\[ -\text{Cov}(v, X_T) > \text{var}(v) \]

So we get positive bias if \( \text{Cov}(v, X_T) \) is negative and of larger magnitude than \( \text{var}(v) \). Example: Income is measured with error, but the error is negative for rich people and positive for poor people. This means that \( \text{Cov}(v, X_T) < 0 \).
Q. — What is the intuition of why we get anti-attenuation bias in this case?

A. — With negative covariance, whenever we see $X$ move, the true value of $X_T$ has actually moved even more. The large change in $X_T$ causes $Y$ to change by a large amount. We will conclude that relatively small changes in $X$ cause large changes in $Y$.

A. — Example: Suppose we are interested in the impact of income on smoking. Suppose that the true impact is that a $\$1000$ increase in income, increases smoking by 10%. If we have mean reverting measurement error, we will observe an increase of measured income of $\$500$ that corresponds to a true increase of $\$1000$. The true increase of $\$1000$ will cause smoking to drop by 10%. We will thus think that increasing income by $\$500$ causes a 10% drop in smoking, implying that a $\$1000$ increase in income would cause a 20% drop in smoking.
4 Linearity Assumption in OLS, Propensity Score Matching and Weighting

Notational change!
In this lecture I will follow Mostly Harmless econometrics and refer to my treatment variable of interest as “D” and the covariates controls as “X”.

Recall, that the treatment on the treatment effect (what OLS estimates) is given by:

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

The first term is directly observed, but we need to estimate the counterfactual $E[Y_{0i}|D_i = 1]$

If we are willing to assume that D is randomly assigned, then we can use $E[Y_{0i}|D_i = 0]$ as the counterfactual.
In OLS, we are rarely in a position where D is randomly assigned so instead we assume that conditional on covariates, D is randomly assigned.

Conditional Independence Assumption (CIA)
We assume $E[Y_{0i}|X_i, D_i = 1] = E[Y_{0i}|X_i, D_i = 0]$

This is the main assumption made in OLS. It is equivalent to $X \perp \epsilon$ and is sometimes discussed as assuming only “Selection on observables”.

Conceptually it says that D is randomly assigned holding X fixed.

4.1 Linearity assumption in OLS

However, OLS also makes an assumption of linearity of the covariates. OLS does not literally hold X’s fixed, it holds the linearly related part fixed.

Q. —- How big of a deal is the linearity assumption in OLS?
What happens if the true model is $Y = X + X^2 + D + \epsilon$ and we estimate $Y = X + D + \epsilon$
We could investigate things the “right” way by writing out the consistency derivation and exploring the bias term.

Instead, I will do things the “wrong” way and do some simulations.

Simulation of non-linearity

```plaintext
clear
set obs 1000
gen x=rnormal(0,1)
gen d=2*x+rnormal(1,1)
```

*This creates a correlation between x and d.*

29
```plaintext
gen e=rnormal(0,.1)
gen y=10*x-x*x+3*d+e

**we get bias if we fail to control for x.
reg y d

**If we control for X and it’s square, we get correct estimates
gen x2=x*x
reg y d x x2
**notice that we could have done this estimation using factor notation
reg y d c.x##c.x

**What if instead we just run a linear model?
*The linearity between Y and X is unimportant!
reg y d x

**This is true, even though X and D2 are correlated
corr x d2

**The reason is that X and D2 are uncorrelated conditional on X.

**We only get bias if we create non-linearity between D and X in addition to X and Y.

Let’s look at that:
clear
set obs 1000
gen x=rnormal(0,1)
gen d=x+2*x^2 + rnormal(0,1)
gen e=rnormal(0,.1)
gen y=10*x-x*x+3*d+e

**linear regression get’s wrong answer
reg y d x

**add correct functional form to get right answer:
reg y d c.x##c.x

In general, we won’t know the correct functional form between D and X and Y and X.
A lot of relationships are close enough to linear so that linear regression is good enough.

If we think that a particular control is very non-linear, we can always use dummy variables.
reg y d i.x
```
Q. — Why does non-linear relationship between X and Y not create bias in D?

A. — If we mis-specify the relationship between X and Y, the non-linear mis-specification goes into the error term. So long as D and X have a linear relationship, there will be no bias.

Essentially, mis-specification can be thought of as an omitted variable problem where the omitted variable is the non-linear relationship between X and Y. So long as D is uncorrelated with that omitted variable, it doesn’t create bias.

If you think through your functional specification seriously, “linearity assumption” is usually not that big of a deal in my opinion.

My rationale:

1. Violation of linearity assumption only occurs when X has a non-linear relationship with D and Y
2. We can always dummy out a continuous covariate. This avoids any functional form assumptions.
3. In the extreme case, if every variable is modeled as a series of dummy variables and fully interacted with all other covariates, there can be no functional form mistakes. In practice, this is rarely done because it becomes unwieldy.

CAVEAT Often functional form issues severely bias estimates because the researcher has not thought through the functional form carefully enough. I am not saying that linearity is an innocuous assumption. I am saying that it can be dealt with directly, whereas X \perp \epsilon can not be easily fixed.

One way to address the OLS linearity assumptions is by flexibly modeling each variable and exploring robustness to different functional forms for the X’s.

An alternative approach is using a matching estimator.

4.2 Exact Matching

The idea behind matching is to more explicitly estimate the counterfactual \( E[Y_0|X_i, D_i = 1] \).

Essentially, we match each individual in the treated group to an untreated individual with the same covariates. We then invoke the CIA assumption and simply take the difference between treatment and control in each match.

Matching Steps:

1. Use the full sample of treatment and controls and divide into cells according to X’s.

   Definition: A cell is a unique combination of covariates.

   • If X’s are gender and race, the cells would be: white females, black females, white males, etc.
   • If one of the X’s was education, then an example cell would be: white females with 14 years of education.
• Even with a small number of covariates, we need a lot of sample size to make sure that every cell has enough people.

2. Within each cell, the X’s are the same, but some people will be treated (D=1) and others will be untreated (D=0). We assume CIA, so the untreated people are assumed to be a good counterfactual for the treated people.

3. For each treated person, we consider the counterfactual to be average outcome of the untreated people in their cell.

4. The treatment effect for each person is the difference between their actual outcome and the estimated counterfactual.

5. We average all the treatment effects for the treated people to get an estimate of the treatment on the treated (TOT).

• Note that we could calculate the average treatment effect for a subgroup if we wanted to. E.g. we could calculate the average treatment effect for just the black individuals to get the “treatment on the black treated”.

Example

<table>
<thead>
<tr>
<th>id</th>
<th>X</th>
<th>D</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>0</td>
<td>55</td>
</tr>
</tbody>
</table>

The treated individuals are person 2,3,5,7,8. The chart below calculates the treatment effect for each person.

<table>
<thead>
<tr>
<th>id</th>
<th>Y</th>
<th>Counterfactual</th>
<th>Treatment effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>15</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>27.5</td>
<td>2.5</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The matching estimator is thus (5+10+2.5)/3 = 5.8

A few points:

• The matching estimator used the untreated observations as the counterfactuals only. This is why it is estimating the TOT as opposed to the ATE.
The observations with $X \geq 3$ were never used at all. This is because there is no variation in treatment status.

Matching estimators only estimate the impact of the treatment where there is “common support”.

- **Definition:** Common support is the regions of $X$ where there are both treatment and control observations.
- This means that matching isn’t really even estimating the TOT. It is estimating the TOT for people in the common support region. In this case, it is the TOT for people with $X=1$ or $X=2$. Since matching is estimating a TOT only in the common support region, it is important to show that the common support region is large and not unusual.

I like that matching only estimates treatment effects in the area of common support since we don’t really want to use control individuals who have very different observables than any of the treated individuals.

The matching estimator is explicitly doing what undergrads *think* OLS is doing.

### 4.3 Comparison of matching to regression

**Q.** Does OLS get us something similar to matching?

**A.** *If we run a regression in Stata: reg y X D we get a coefficient on D of $-1.21$*

**Q.** Why doesn’t OLS get us similar answer to matching if it is supposed to be conceptually similar?

**A.** *In this case, the relationship between Y and X is not quite linear and the relationship between X and D is highly non-linear*

**Common support in linear regression**

In linear regression, the common support idea breaks down. We can compare a treated person with $X=2$ to a control person with $X=5$ because we assume linearity in $X$.

Conceptually, linear regression is doing the following steps

1. Calculate the linear relationship between $X$ and $Y$.
2. Use this linear projection to infer the outcome for the $X=5$ control if they had had $X=2$. Call this projection $Y_{0i}^{Projection}$.
3. Compare the $X=2$ treated person outcome to $Y_{0i}^{Projection}$.

Essentially, the linearity in OLS allows us to make comparisons across covariate cells and so we don’t require common support. The downside is the linearity assumption may be poor (as
it was in this case). The benefit of linear OLS is that it uses more data, so the standard errors will be smaller.

**Q.** —— What if we control for education perfectly flexibly using dummy variables e.g. in Stata: reg y i.x d

**A.** —— If we include dummy controls for education, OLS can no longer compare across covariate cells, so OLS is only using the area of common support. We get an estimated coefficient of 5. This is much closer to 5.8 and very different from -1.21

Even though OLS and matching didn’t get the exact same thing, you can think of OLS and matching basically being the same so long as OLS is estimated in a flexible fashion. People sometimes discuss matching estimators as if they are dramatically different from OLS, but this isn’t the case. A matching estimator does more or less the same thing as OLS, but it uses a more flexible functional form.

**Common support in flexible regression**

If we include dummy variables, we once again are only using the area of common support. We can see this by re-running the regression only on id < 3.

If we run the regression of reg y i.x d if id < 3 we get the exact same estimate of 5 that we got before.

**IMPORTANT:** With flexible controls, OLS only uses observations where there is common support. (The number of observations shown in the regression will still show the total sample size, but the effective sample size can be much smaller.) This is literally true. Dropping observations with no variation in treatment will yield identical estimates to including these observations.

**Q.** —— Why doesn’t OLS get exactly the same point estimate as matching?

**A.** —— OLS and matching both combine the treatment effects for the treated individuals, but the weights are different.

In matching, we take a straight average across all the treated individual. This is equivalent to weighting each cell based on fraction of treated individuals in the cell.

In OLS, the weighting is based on the variance of the treatment within each cell. If you are a cell with a high or low proportion of treated individuals, the estimated treatment effect in this cell will receive relatively little weight because the variance (p*(1-p)) is relatively low. OLS puts the highest weight on cells where there are an equal number of treated and control individuals. In general, OLS puts weight where there is a lot of covariate variance and a small amount of residual variance.

For example, in the above example, we have 2/3s of the treated observations have X=1 and the average treatment effect when X=1 is 7.5. 1/3 of the treated observations have X=2 and the average treatment effect when X=2 is 2.5.
7.5 \times \left(\frac{2}{3}\right) + 2.5 \times \left(\frac{1}{3}\right) = 5.8

This is equivalent to the simple average of the treated individuals.

\frac{(5+10+2.5)}{3} = 5.8

In OLS, the estimated coefficient is a weighted average of the same treatment effects, but the weights are different. The exact weights are given in mostly harmless econometrics section on matching. These weights are essentially adding more weight on cells with more treatment variance and also adds more weights based on cell size.

In this case, the variance in cell X=1 and cell X=2 are identical \((p \times (1-p)) = \left(\frac{1}{2} \times \frac{3}{2}\right)\) and both cells have 3 observations, so the weights on the two treatment effects are the same.

OLS: \(7.5 \times \frac{1}{2} + 2.5 \times \frac{1}{2} = 5\)

In most cases, flexible OLS and matching will give you very similar estimates since they are weighted averages of the same estimated treatment effects.

### 4.4 Propensity Score Matching

In practice, matching is almost never done as exact matching because with multiple covariates, there are simply too many cells. This is known as the curse of dimensionality Propensity score matching is conceptually identical to exact matching, but is feasible to implement, even with many continuous covariates. The propensity score differs for each person and is the probability each person is treated. The \(P(D_i = 1)\) for each individual is estimated based on their covariates. We then match people based on the propensity scores. The idea is to match people who have very similar propensities to take up treatment, but one is treated and one is not treated.

**Propensity Score:** The estimated probability that you take up treatment.

The Propensity Score Matching procedure is:

1. Estimate a logit regression of \(D_i = \text{covariates}\). Unlike, most of microeconometrics, this case, we are interested in the best prediction of \(D_i\), so we will potentially include flexible functions of many of the relevant covariates. We could also do this step with OLS or probit.

2. We use the logit regression to get a predicted probability of \(\hat{D}_i = 1|X_i\) for each individual in the sample. This predicted probability is called the propensity score.

3. We match people based on the propensity score. The matching procedure can be done in many different ways, but it will match individuals based on having similar propensity scores rather than identical propensity scores.

4. The remainder of the procedure is identical to the exact matching case previously discussed.

Types of Propensity Score Matching:
1. Nearest neighbor matching
2. Nearest K neighbor matching
3. Kernel matching
4. Stratification
5. Many others

1. Nearest neighbor matching – matches each treated person to nearest propensity score in untreated group.

2. Nearest K neighbor matching – matches each treated person to nearest k people in untreated group.

3. Kernel matching – Like nearest k matching, but with weights (more details below).

4. Stratification – Split sample into blocks by propensity score (e.g. 5 quintiles or 100 percentiles).

5. Many others

### 4.5 Propensity Score Matching Estimation

Mathematically, we can all the different matching estimators can be written in the form:

$$\text{TOT: } \frac{1}{N_T} \sum_{i \in \{D=1\}} \left( Y_{1,i} - \sum_{j \in \{D=0\}} \phi(i, j) Y_{0,j} \right)$$  \hspace{1cm} (7)

The $\phi(i, j)$ specifies who is included in the match for person $i$ (and to what extent).

- If we want to do nearest neighbor matching, we would set $\phi(i, j) = 1$ if person $j$ is person $i$’s nearest neighbor and $\phi(i, j) = 0$ otherwise.

- If we want to do stratification, we set $\phi(i, j) = \frac{1}{N_{C,j}}$ if person $j$ is in the same strata as person $i$’s and $\phi(i, j) = 0$ otherwise. $N_c$ is the number of control individuals in person $i$’s strata. This is just taking an average of the control outcomes in person $i$’s strata.

- The Kernel matching method is more nuanced:

$$\phi(i, j) = \frac{K(p(x)_j - p(x)_i)}{\sum_{j=1}^{N_{C,j}} K(p(x)_j - p(x)_i)},$$  \hspace{1cm} (8)

where $K$ is a kernel function.

The idea behind a kernel is described below.
4.5.1 Kernel Matching

Kernel matching is a form of matching in which the untreated individuals are used to calculate the counterfactual for person $i$, but the controls more similar to person $i$ are given more weight.

**Kernel functions:** A symmetric density function with its maximum when the argument is zero and decreases on either side. Some examples of Kernel Functions:

- The triangular kernel: $1 - |x|$ for $|x| \leq 1$. (This looks like an upside down triangle)
- Epanechnikov kernel: $\frac{3}{4}(1 - x^2)$ for $|x| \leq 1$ (This looks like an upside down parabola)
- The gaussian kernel.
- The rectangular kernel. (equal weights on all observations in some range, zero weight outside the range).

All of these kernel functions provide different ways to assign weights to all of the observations in your data. The triangle kernel as written above gives zero weight outside of $|x| \leq 1$. It gives observations at $x=0$ much of the weight, and observations near $x=1$ or $x=-1$ almost no weight. The gaussian kernel gives positive weight to all observations in your data, but puts the most weight around $x=0$.

Returning back to equation (8), the weights are calculated so that the largest weight is when $p(x)_j$ is close to $p(x)_i$. The denominator in equation (8) is just to make sure the weights $(\phi(i, j))$ sum to 1.

For some kernels, for example the gaussian kernel, each treated individual is matched to all of the control individuals, but the counterfactual for each treated individual will be a different weighted average of the control’s outcomes. For treated individuals with high propensity scores, the treated individuals with high propensity scores will be given the most weight in calculating the counterfactual.

**Aside: Kernel density weighting**

Kernel density weighting is a general technique to make discrete measures continuous or calculating a running average. This is also known as Kernel smoothing.

4.5.2 Investigating Common Support

Since matching methods only use the area of common support, it is good practice to show this area empirically. If the area of common support is very small, this means that the estimated treatment effects will be based on relatively few observations.

Showing the area of common support is fairly straight-forward when using propensity score matching methods.

- Create a smooth histogram of the propensity score for the treatment group. (you can use a kernel smoother if necessary)
• Overlay a smooth histogram of the propensity score for the control group.
• The overlapping region shows the area where treatment and control have the same propensity scores e.g. the area of common support.

When implementing matching, it is expected that you think about and present the region of common support. Even when you’re using OLS, it is good to think about the region of common support.

If you have a small region of common support it doesn’t bias estimates, but it does make things more tenuous since the estimates are based on relatively few observations. If you have a small region of common support in OLS, you are either basing your estimates on a small number of observations, or you are relying heavily on the linearity assumption (as was the case in the numerical example).

4.5.3 Regression or Matching?
Essentially, the two methods are very similar and are of the same type. They both rely on a selection on observables framework in which the CIA assumption is made. If the treated and untreated groups differ in unobservable ways that are related to the outcome, both regression and matching will fail. In practice, there are some differences:

Advantages to Regression
• Regression is a bit easier to implement in general.
• Continuous treatment doesn’t lend itself to matching.
• Standard errors have a closed form solution.
  – While bootstrapping allows for the calculation of the matching standard errors, this can be computationally intensive and time consuming. Furthermore, Abadie (2006, econometrica) shows that bootstrapping actually fails for nearest neighbor matching.
• Standard errors will be smaller since we can extrapolate to use observations outside of the area of common support.

Advantages to Matching
• Matching is less parametric (though the propensity score still needs to be estimated parametrically).
• Matching makes it easy to explicitly see the area of common support.

4.6 Weighting Estimator
The math here is a little complicated, but the idea and implementation are simple.

I will first go over the idea and implementation, then I will go over the math behind it.
A general note on weights

Weights are a much more general concept than the weighting estimator. Many datasets come with sampling weights in which certain observations represent more than their proportional share. For example, some surveys oversample minorities (to increase their sample size) but in order to get analyses representative of the population, it is necessary to downweight the minorities that were oversampled. In general, Stata accepts weights for most commands. When using weights, Stata will put more emphasis on observations with higher weight. If person N has twice the weight as person M, it is essentially like we have 2 observations identical to person N and 1 observation for person M. Stata can calculate weighted means in this way, but it can also calculate weighted standard deviations, perform weighted regressions, weighted histograms, etc.

The most typical use of weights is to make a sample representative of a population. For example:

- If a survey surveyed 60% women, we could reweight the sample data so that it becomes representative of the population.

But we can use weights in lots of ways!

- If smokers are mostly men, we could reweight the smokers so that they become representative of the population.
  - The weighted smokers would still all be smokers, but less weight would be placed on the men and more weight would be placed on the women.
  - Note that the weighted sample of smokers would not be representative of smokers, but would be representative of the population in terms of gender.

- We can reweight any sample to become representative of any group if we specify the correct weights.

Weighting Estimator

The weighting estimator uses the propensity score idea, but rather than matching treated individuals to control individuals, we reweight the data so that the treated and control groups become counterfactuals.

If we want to estimate the ATE, we need to weight the treated and untreated individuals so they are representative of the average individual. We then simply compare treated and controls.

If we want to estimate the TOT, we leave the treated individuals unchanged and weight the untreated individuals so they are representative of a treated individual. We then simply compare treated and controls.
What weights do we use?

The weights are based on the propensity score and are given below:

<table>
<thead>
<tr>
<th>Goal:</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated individuals → average individuals</td>
<td>( \frac{1}{p(x)} )</td>
</tr>
<tr>
<td>Untreated individuals → average individuals</td>
<td>( \frac{1}{1-p(x)} )</td>
</tr>
<tr>
<td>Untreated individuals → treated individual</td>
<td>( \frac{p(x)}{1-p(x)} )</td>
</tr>
<tr>
<td>Treated individuals → untreated individual</td>
<td>( \frac{1-p(x)}{p(x)} )</td>
</tr>
</tbody>
</table>

where \( p(x) \) is the propensity score

**Intuition:** Consider the weight which makes untreated individuals look like treated individuals. If we want the untreated individuals to look like they have \( D=1 \), we put the most weight on the untreated individuals who have characteristics that would predict them to have \( D=1 \). In other words, we put a lot of weight on the untreated individuals with high propensity scores. Untreated individuals with low propensity scores have characteristics very different from \( D=1 \), so we put very little weight on these individuals.

Note that we can think of the weight \( \frac{p(x)}{1-p(x)} \) as split into two parts. The \( \frac{1}{1-p(x)} \) portion makes the untreated individuals look like average individuals. The multiplying by \( p(x) \) portion makes the average individual look like treated individuals.

The implementation of the weighting estimator is very easy. Suppose we are interested in estimating the TOT.

1. We want to compare the treated individuals to a counterfactual.
2. The counterfactual we use is the untreated individuals using weights of \( \frac{p(x)}{1-p(x)} \).
3. To calculate the TOT, simply compare the means between the treated and weighted untreated individuals.

**Cautionary note:** Individuals with extreme propensity scores will get a huge amount or tiny amount of weight. A tiny amount of weight on one observation doesn’t bother me terribly because it won’t have a huge impact on results to drop a single observation. Observations that get a huge amount of weight can be problematic, particularly if the tiny propensity score was so tiny due to sampling error.

Unlike in OLS or matching, here we have an entire distribution of treated individuals and an entire distribution of counterfactuals.

So long as we are weighting the untreated individuals, we can compare the entire distributions to one another.

Go to STATA example.
clear
sysuse nlsw88
gen lwage=log(wage)

*We are interested in the impact of union on wages.
reg lwage union
reg lwage union ttl_exp tenure smsa south collgrad i.race

**note that union works are not the same as non-union workers.
summ collgrad if union==1
summ collgrad if union==0

summ south if union==1
summ south if union==0
**we will use inverse propensity score weights to make the
**non-union worker look like the union workers.

**Step 1. Get the propensity scores
logit union ttl_exp tenure smsa south collgrad i.race
predict ps
keep if ps!

**Step 2. Use the propensity scores to create the appropriate weights
**these weights will make non-union workers look like union workers.
gen c_to_t_weight=ps/(1-ps) if union==0
replace c_to_t_weight=1 if union==1

**It works!
summ collgrad if union==0
summ collgrad if union==1
summ  collgrad [aw=c_to_t_weight] if union==0

summ south if union==0
summ south if union==1
summ  south [aw=c_to_t_weight] if union==0

**Now we assume that the weighted non-union workers are a good control group and we look
**supposedly, unions raise average wages and also reduce the standard deviation of wag
**we can compare both of these using the summ statement
summ lwage if union==0
summ lwage if union==1
summ  lwage [aw=c_to_t_weight] if union==0

reg lwage union [aw=c_to_t_weight]
**the controls do nothing here since it’s already weighted!
reg lwage union ttl_exp tenure smsa south collgrad [aw=c_to_t_weight]
kdensity wage
twoway (kdensity wage if union==1) (kdensity wage if union==0), legend(lab(1 "union") lab(2 "non-union"))
twoway (kdensity wage if union==1) (kdensity wage [aw=c_to_t_weight] if union==0), legend(lab(1 "union") lab(2 "non-union weighted to look like union"))

Mostly harmless econometrics provide an explanation and proof of why we use these weights on page 82 (see the footnote combined with the previous page).

I will explain where the weights come from using a more general framework from Dinardo Fortin Lemieux (Econometrica 1996)

4.6.1 Math behind weights

Suppose we are interested in the impact of unions on wages. In this case, D=1 refers to a unionized worker and D=0 is a non-union worker. For simplicity, I will think of a case with a single covariate.

Let’s first write down what we can directly observe in data:

- \( g(w) \) – overall wage distribution
- \( h(x) \) – overall distribution of covariate
- \( g(w|D = 0) \) – distribution of non-union wages
- \( g(w|D = 1) \) – distribution of union wages

If we wanted to calculate \( g(w|D = 0) \) in Stata, we would type "histogram wage if Union==0".

By the definition of conditional distributions, we have

\[
g(w) = \int f(w|x)h(x)dx
\]  

where \( f(w|x) \) is the distribution of wages conditional on \( x \).

This is analogous to the law of iterated expectations. Essentially, if we integrate over all possible values of \( X (h(x)) \), we get the overall distribution back.

Using this form, we can rewrite the two conditional distributions as:

\[
g(w|D = 0) = \int f^{[D=0]}(w|X)h(x|D = 0)dx
\]

\[
g(w|D = 1) = \int f^{[D=1]}(w|X)h(x|D = 1)dx
\]

where \( f^{[D=0]}(w|X) \) is the distribution of wages for non-union workers conditional on \( x \).
Note that the total expression in equations 10 and 11 are directly observable in the data.

Our goal here is to get a counterfactual distribution for the unionized workers. We can directly observe \( \int f^{(D=1)}(w|X)h(x|D = 1)dx \) and we want to estimate the counterfactual \( \int f^{(D=0)}(w|X)h(x|D = 1)dx \).

Note the difference between the observed wage distribution for unionized workers and the unobserved counterfactual wage distribution. The key point is that in both expressions, we are integrating over the same covariate distribution. The difference is that for the counterfactual, we are looking at the wage distribution if the workers with the characteristics of unionized workers were paid according to the non-unionized wage distribution.

Mathematically, we just need to get an estimate of the counterfactual using only information that we can observe in the data.

To do this, recall Bayes Law: \( P(A|B) = \frac{P(B|A)P(A)}{P(B)} \). Notice that we can rewrite this as \( P(B) = \frac{P(B|A)P(A)}{P(A|B)} \).

In terms of distributions, we have:

\[
h(x) = \frac{h(x|D = 0)\text{Prob}(D = 0)}{\text{prob}(D = 0|x)}
\]  
\[\text{(12)}\]

Similarly, we can write:

\[
h(x) = \frac{h(x|D = 1)\text{Prob}(D = 1)}{\text{prob}(D = 1|x)}
\]  
\[\text{(13)}\]

Combining equations 12 and 13 we get:

\[
\frac{h(x|D = 1)\text{Prob}(D = 1)}{\text{prob}(D = 1|x)} = \frac{h(x|D = 0)\text{Prob}(D = 0)}{\text{prob}(D = 0|x)}
\]  
\[\text{(14)}\]

Rearranging terms we get:

\[
h(x|D = 1) = h(x|D = 0) \frac{\text{Prob}(D = 0) \text{Prob}(D = 1|x)}{\text{prob}(D = 0|x) \text{Prob}(D = 1)}
\]  
\[\text{(15)}\]

We will use this result to plug in for \( h(x|D = 1) \) in terms of \( h(x|D = 0) \) to estimate the counterfactual.

Recall: We can directly observe \( \int f^{(D=1)}(w|X)h(x|D = 1)dx \) and we want to estimate the counterfactual \( \int f^{(D=0)}(w|X)h(x|D = 1)dx \).

Using the result from equation 15, we can rewrite what we want in terms of what we can observe. The counterfactual becomes
Counterfactual = \int f^{D=0}(w|X)h(x|D = 1)dx

= \int f^{D=0}(w|X)h(x|D = 0) \frac{Prob(D = 0)}{prob(D = 0|x)} \frac{Prob(D = 1|x)}{Prob(D = 1)} dx

= \int f^{D=0}(w|X)h(x|D = 0) \theta dx

The implication of the previous line is important. It says that we can rewrite the counterfactual in terms of the observed distribution for the untreated group. All we need to do is weight by parameter \( \theta \) where the weights are given by \( \frac{Prob(D = 0)}{prob(D = 0|x)} \frac{Prob(D = 1|x)}{Prob(D = 1)} \). Once we get the counterfactual distribution, we can compare the observed distribution of union wages to the counterfactual distribution of union wages in any way we like.

In general, STATA (and other software) will always normalize the weights so that they sum to 1. This means that any constant multiplicative term in the weight is irrelevant if it doesn’t vary by individual. (e.g. With 2 observations, weights of 3 and 1 are identical to weights of 6 and 2).

Ignoring the terms in \( \theta \) that don’t vary by person, the weights become \( \theta = \frac{Prob(D = 1|x)}{prob(D = 0|x)} \).

Note that the propensity score is an estimate of \( Prob(D = 1|x) \) and \( (1 - p\text{score}) \) is an estimate of \( prob(D = 0|x) \).

As such in order to weight the control group to be representative of the treated group, we weight by \( \frac{p\text{score}}{1 - p\text{score}} \).
5 Best practices to “test” CIA

5.1 Selection on Observables

OLS, matching and weighting are all in the category of “selection on observables”. The word “selection” refers to the idea that the treated observations (D=1) chose to take up treatment and are therefore a select group of individuals, not comparable to the control group (D=0).

In OLS, matching and weighting, we assume that the treated observations are only select in terms of their observable factors. Once we control for these observable factors, the remaining variation in D is random (this is the CIA assumption).

Q. — How do we evaluate the CIA assumption?

A. — Usually, we just think about potential simultaneity or omitted variables. There are also empirical “tests” that can inform our thinking. The reason the word “tests” is in quotes is because it is wrong to think of these as definitive tests of the CIA assumption. Fundamentally, it is untestable. The tests we run will help us gauge how plausible CIA is, but ultimately, we must still rely on our best judgement.

5.2 Best practices to explore plausibility of CIA

In a paper, usually these explorations would come after the main results in a section called “Specification Tests”. It is sometimes also possible to include some of these tests towards the beginning of a paper in the initial descriptive statistics.

5.2.1 Means comparison of treatment and control (balance test)

To get a basic understanding of the selection processes, it is a good idea to begin by considering the average characteristics of D=1 vs D=0.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Treated Individuals</th>
<th>Control Individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Mean: 30, SD: 18</td>
<td>Mean: 24, SD: 14</td>
</tr>
<tr>
<td>Percent black</td>
<td>Mean: 0.34, SD: 0.2</td>
<td>Mean: 0.21, SD: 0.18</td>
</tr>
<tr>
<td>covariate3</td>
<td>Mean: 2.4, SD: 0.4</td>
<td>Mean: 4.5, SD: 0.3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

This is known as examining whether there is covariate balance. In general, any paper based on a randomized control trial will show covariate balance as the first table. Even if you don’t have a randomized control trial, it’s a good idea to consider whether the covariates are balanced. In most cases, you will not have balanced covariates, but if you do, it makes the CIA assumption very plausible.
In practice, relatively few papers actually show the balance test because it often doesn’t help
the authors’ case. If you actually have balance, you can add 2 extra columns that show the
difference in the means and a t-test on the difference.

Q. — What is the logic behind this test?

A. — If there is balance, this means that there is no selection on observables (i.e. treatment is
unrelated to observables). If there is no selection on observables, this makes it plausible that
there is no selection on unobservables.

Important Point: The CIA allows for selection on observables. If you fail the balance test
this is not a direct violation of the CIA. That said, if you fail the balance test, then you need to
argue the following: “Even though the treatment is related to observable characteristics like
age, race, and education, I assume that it is not related to unobservable factors like
motivation.” This argument is often implausible.

5.2.2 Test of balance for continuous treatments
If you are interested in the impact of a continuous treatment, it is not easy to split the sample
into “treatment” and “control” to assess the balance in covariates. There are two alternatives
that get at the same idea.

• Regress D on all the covariates. This tells you whether the covariates are systematically
related to the treatment, which is analogous to the test of balance.

• Discretize D into above and below the median and then show the dichotomous balance
test.

Predicted outcome test
In either the continuous or discrete case, a nice summary test is as follows:

1. Use all the exogenous covariates to predict the outcome and created predicted values \( \hat{Y} \).
2. Plot the \( \hat{Y} \) against D (or run the regression).
3. Ideally, the impact of D on \( \hat{Y} \) is found to be zero.
4. I like this test because it is a direct analogue of the CIA assumption.

5.2.3 A more relaxed test of balance
Outside of a random experiment, you will rarely pass the balance test. You don’t really have
to pass the balance test though, since the whole point of OLS is that we can control for
observable differences between treatment and control. The problem is that once you control
for all the covariates, you have no more ways to test the plausibility of the CIA. To get around
this issue, you can implement a test of conditional balance.

Conditional Balance Test: Once a subset of observables are controlled for, are the rest of the
covariates balanced?
If we can show that covariates i, j and k are unrelated to D once we control for covariates a, b, c, d, and e, this makes it more plausible that all the unobservables are unrelated to D once we control for covariates a, b, c, d, and e.

One way to implement this test is:

- Regress covariate i, j or k on D and covariates a, b, c, d and e. The coefficient on D tells us whether D is related to i, j or k conditional on a-e.
### 5.2.4 Gradually add covariates to main regression

Almost every regression based paper includes a table with the following structure.

<table>
<thead>
<tr>
<th>Table 2: Impact of Layoff on Higher Education Enrollment</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) enroll     (2) enroll     (3) enroll     (4) enroll     (5) enroll</td>
</tr>
<tr>
<td>Parent laid off at ages 15-17   -0.111*** (-0.0401)  -0.100*** (-0.0393)  -0.0774*** (-0.0377)  -0.0870*** (-0.0373)  -0.0848*** (-0.0377)</td>
</tr>
<tr>
<td>Parent laid off either at ages 15-17 or 21-23  -0.0176 (-0.0330)  -0.0213 (-0.0324)  -0.0208 (-0.0314)  -0.00456 (-0.0315)  -0.00409 (-0.0317)</td>
</tr>
<tr>
<td>Parent laid off in both periods   -0.114* (-0.0649)  -0.102 (-0.0657)  -0.103* (-0.0622)  -0.107* (-0.0613)  -0.105* (-0.0603)</td>
</tr>
<tr>
<td>Female                             0.0956*** (-0.0154)  0.105*** (-0.0141)  0.106*** (-0.0140)  0.106*** (-0.0139)  0.106*** (-0.0139)</td>
</tr>
<tr>
<td>SOE                                -0.114*** (-0.0241)  -0.0735*** (-0.0220)  -0.0673*** (-0.0216)  -0.0652*** (-0.0216)</td>
</tr>
<tr>
<td>Black                              -0.0622*** (-0.0245)  -0.0236 (-0.0226)  -0.0236 (-0.0226)  -0.0101 (-0.0226)  -0.0123 (-0.0226)</td>
</tr>
<tr>
<td>Father high school                0.0864*** (-0.0197)  0.0456** (-0.0203)  0.0370* (-0.0203)  0.0370* (-0.0203)</td>
</tr>
<tr>
<td>Father some college               0.206*** (-0.0234)  0.150*** (-0.0242)  0.144*** (-0.0243)  0.144*** (-0.0243)</td>
</tr>
<tr>
<td>Father college                    0.309*** (-0.0278)  0.242*** (-0.0289)  0.236*** (-0.0289)  0.236*** (-0.0289)</td>
</tr>
<tr>
<td>Father some grad                   0.337*** (-0.0281)  0.260*** (-0.0293)  0.251*** (-0.0297)  0.251*** (-0.0297)</td>
</tr>
<tr>
<td>Mother high school                0.0838*** (-0.0208)  0.0561*** (-0.0212)  0.0569*** (-0.0213)  0.0569*** (-0.0213)</td>
</tr>
<tr>
<td>Mother some college                0.222*** (-0.0242)  0.181*** (-0.0249)  0.179*** (-0.0250)  0.179*** (-0.0250)</td>
</tr>
<tr>
<td>Mother college                     0.308*** (-0.0303)  0.245*** (-0.0311)  0.241*** (-0.0316)  0.241*** (-0.0316)</td>
</tr>
<tr>
<td>Mother some grad                   0.345*** (-0.0318)  0.278*** (-0.0329)  0.273*** (-0.0333)  0.273*** (-0.0333)</td>
</tr>
<tr>
<td>Home owner                         0.0301* (0.0163)  0.0329** (0.0163)  0.0329** (0.0163)  0.0329** (0.0163)</td>
</tr>
<tr>
<td>Income Category                     y            y            y            y            y</td>
</tr>
<tr>
<td>Year of Birth Dummies              y            y            y            y            y</td>
</tr>
<tr>
<td>N                                  4030         4030         4030         4030         4030</td>
</tr>
<tr>
<td>R-sq                               0.009        0.042        0.204        0.220        0.227</td>
</tr>
</tbody>
</table>

In the above table, the treatment variable of interest (D) is “Parent laid off at ages 15-17”. The first column shows a barebones regression with few covariates included. Each subsequent column adds in additional controls and the preferred specification is column (5) on the far right.

This table should be thought of as a type of conditional balance test.
If adding a covariate doesn’t change the coefficient on D, this means that the covariate is either unrelated to D, or unrelated to Y. If none of the observables are related to D and Y, this makes it plausible that none of the unobservables are related to D and Y.

People often discuss this type of table as demonstrating the robustness of the result, but this is not really the point. Robustness means that the results are not sensitive to the particular specification (e.g. shows you didn’t cherry pick a particular set of covariates and functional form to get a particular result). **Showing a table like the one above is not simply showing that the results are not sensitive, the table is providing important evidence regarding the plausibility of the CIA.**

A less precise way to think about this type of table is to argue that since the coefficient on D didn’t change from columns (1) - (5), it is likely that it would remain stable if we added on more columns in which we controlled for other unobserved factors.

**Important point:** If adding covariates changes the coefficient on D, it makes it less plausible that CIA holds, but it still is possible that CIA holds.

### 5.2.5 Failing the CIA specification tests

It is possible to fail all of these tests and still argue that the CIA holds. Each of the above tests has the basic idea of testing for independence between D and observables to assess the likelihood of independence between D and unobservables. Even if it turns out that every observable characteristic is related to D and Y, theoretically, so long as we control for all these characteristics, it is possible that the remaining unobservables are conditionally independent. This type of argument will be an uphill battle though.

### 5.2.6 Falsification tests

Falsification tests aim to show an absence of an effect where we would not expect an effect. These types of tests usually involve switching Y with some other outcome that could not be impacted by D.

A common falsification test examines how D impacts past outcomes that were determined prior to D. For example, if we aim to show small class sizes increase student test scores, it would be useful to show that being in a small class in 2nd grade does not “impact” 1st grade test scores.

Example: A paper published in a good journal and featured in a TED talk argues that obesity is contagious (he shows that people with obese friends are more likely to be obese). A simple falsification test by Jason Fletcher finds that height and other immutable factors are also found to be “contagious”.

### 5.3 The problem with OLS, matching and weighting

The main issue in OLS, matching or weighting is that we must assume that D is randomly assigned, conditional on covariates (CIA assumption). Except for special circumstances, it is rarely possible to convince the reader that the CIA holds perfectly.
5.3.1 Lalonde (1986)

In a very influential paper, Lalonde compares experimental estimates from an employment program to various econometric specifications that all rely on a regression type framework. In his paper, he has a randomly assigned control and treatment group, but he also creates a number of other control groups that are constructed based on observable characteristics.

Lalonde finds that the regression estimates are very far from the experimental estimates. This created a huge stir, as it suggested that the methods most economists use are inadequate. Several follow-up papers suggested that matching estimators obtain estimates closer to the experimental ideal, but the basic notion that economists need to pay close attention to the source of identifying variation remains.

5.4 OLS is Conditional Means Comparison

If you don’t think the CIA assumption is plausible in your context, OLS is still a very useful tool. Just as you might compare means and summary statistics to describe the data, you can use OLS to compare conditional means to describe the data in a more nuanced way. In general, you should think of OLS as a tool that provides conditional means comparison.

6 Instrumental Variables

6.1 External vs Internal Validity

Up until now, we have mostly been thinking about how to obtain consistent estimates of the impact of D on Y. If we allow for heterogeneous treatment effects, it is possible that we are getting a good estimate of the treatment effect for a particular group, but this estimate is not generalizeable.

Definitions

Internal Validity: Are you correctly estimating the causal impact within the context of your sample.

External Validity: Can you generalize your results to other populations or other contexts?

Everything that we have discussed with respect to consistence is about internal validity.

Estimates might be internally valid, but if the initial dataset is not representative of the entire population, the estimates may not be externally valid.

Even if the sample is representative of the population, the context of the study might be unusual such that it doesn’t generalize to other settings.

External validity is less important than internal validity. If a study is internally invalid, we have learned nothing. If a study is internally valid, but externally invalid, we have at least learned something about our particular context. In general, you should be much more concerned about internal validity then about whether your dataset is representative of the population.
6.2 Instrumental Variables (IV)

Instrumental variables are used to address the following problems encountered in OLS, matching or weighting.

1. omitted variable bias
2. measurement error
3. simultaneity or reverse causality.

Each of these issues makes $D \not\perp \varepsilon$ so OLS will be biased. These issues also bias matching and weighting in much the same way. If there are omitted variables or simultaneity, then the CIA assumption is violated so the matching and weighting estimators will be biased. If there is measurement error in D, then we get bias when using matching or weighting because we are mis-classifying treatment and control.

**Identifying variation**

In OLS, matching and weighting, we cross out variation that we can control for. The remaining variation in D is assumed to be random (CIA).

Instrumental variables are a fundamentally different type of identification strategy. We use only the variation caused by the instrument.\(^1\)

Instead of assuming that the “left-over” variation is clean, we must assume that the variation caused by the instrument is clean.

**Q.** What does it mean for an instrument to cause only clean variation?

**A.** *The instrument is dirty if the instrument Z directly impacts Y*

The idea in instrumental variables is illustrated by the following diagram. Suppose we are in a situation where some of the causes of variation in D, also cause Y.

![Diagram](attachment:image.png)

U is the omitted variable in this context.

Suppose there is one source of variation that we are sure is clean. We can use only this variation to examine the impact of D on Y. The idea is shown in the figure below. Here we are assuming that Z is clean and causes variation in D. When Z changes, it should not impact Y

---

\(^1\)For the rest of the topics in this course, the identifying variation will be much more specific and it is very important that you be able to identify and describe the identifying variation. Once you know what the identifying variation is, you can consider whether it is clean, and also whether much identifying variation exists.
directly, but it should impact D directly. If Y changes when Z changes, this means that D impacts Y.

\[ \text{Z} \rightarrow \text{D} \rightarrow \text{Y} \]

In this context, we assign any relationship between Z and Y to D. If it turns out that Z directly impacts Y through a different channel, we will be inappropriately interpreting the relationship between Z and Y as indicative of D causing Y.

### 6.2.1 IV implementation

We begin with the model we want to estimate:

\[ Y = \rho D + \epsilon \]

We are concerned that \( D \not\perp \epsilon \).

If we are willing to assume that Z has no direct impact on Y, then we have \( Z \perp \epsilon \).

Assuming \( \text{cov}(Z, \epsilon) = 0 \) we can easily get an estimate of \( \gamma \).

\[
\begin{align*}
Y &= \rho D + \epsilon \\
\text{Cov}(Y, Z) &= \rho \text{Cov}(D, Z) + \text{Cov}(Z, \epsilon) \\
\text{Cov}(Y, Z) &= \rho \text{Cov}(D, Z) \\
\frac{\text{Cov}(Y, Z)}{\text{Var}(Z)} &= \rho \\
(\text{Take covariance of either side with } Z) \\
\frac{\text{Cov}(D, Z)}{\text{Var}(Z)} &= \rho \\
(\text{In expectation, we get } \text{Cov}(Z, \epsilon) = 0)
\end{align*}
\]

It is easier to think of things in terms of regression coefficients so I divide numerator and denominator by \( \text{Var}(Z) \).

\[
\hat{\rho}_{IV} = \frac{\text{Cov}(Y, Z)}{\text{Var}(Z)} \frac{\text{Cov}(D, Z)}{\text{Var}(Z)}
\]  

(16)

The numerator is a regression of Y on Z, the denominator is a regression of D on Z.

### 6.2.2 2 stage least squares (2SLS) vs Indirect Least Squares (ILS)

2SLS and ILS are two separate ways of estimating the IV estimator.

**ILS:**

In ILS, we estimate \( \gamma \) exactly as described above in 2 steps.

1. \( D = a + \gamma Z + \zeta \)
2. \( Y = \alpha + \delta Z + \varepsilon \)

The IV estimator is \( \rho_{IV} = \frac{\delta}{\gamma} \)

**2SLS:**

In 2SLS, we again have 2 steps.

1. \( D = a + \gamma Z + \zeta \)
2. \( Y = \alpha + \rho \hat{D} + \varepsilon \) Where \( \hat{D} \) are the predicted values of \( D \) based on (1).

Note that 2SLS is equivalent to ILS. (see page 121 of Mostly Harmless for details).

\[
\begin{align*}
Y &= \alpha + \rho \hat{D} + \varepsilon \quad (17) \\
Y &= \alpha + \rho(a + \gamma Z) + \varepsilon \quad \text{(Plug in for} \hat{D}) \\
Y &= \alpha + \rho a + \rho \gamma Z + \varepsilon \quad (18)
\end{align*}
\]

Equation 17 estimates the 2nd step from 2SLS and gets an estimate of \( \rho \). Equation 18 shows that if we regress \( Y \) on \( Z \), the resulting coefficient is \( \rho \gamma \). This means that we can get the same estimate of \( \rho \) either by using 2SLS, or by regressing \( Y \) on \( Z \) and dividing the coefficient on \( Z \) by \( \gamma \).

### 6.2.3 IV Terminology

The constant \( \alpha \) is not necessarily the same across these regressions; I just ran out of greek letters.

- **Causal relationship of interest**
  \( Y = \alpha + \rho D + \eta \)

- **First-Stage Regression**
  \( D = a + \gamma Z + \zeta \)

- **Second-Stage Regression**
  \( Y = \alpha + \rho \hat{D} + \nu \)

- **Reduced Form**
  \( Y = \alpha + \delta Z + \varepsilon \)

2SLS runs the first stage, and then the second stage yields an estimate of \( \rho \).

Indirect least squares, divides the reduced form coefficient (\( \delta \)) by the first stage coefficient (\( \gamma \)) to get an estimate of \( \rho \).
6.2.4 IV assumptions

The above discussion already is suggestive of the conditions necessary for IV, but below I state them explicitly.

4 IV assumptions

1. $\text{Cov}(Z, D) \neq 0$ (Relevance condition)
2. $\text{Cov}(Z, \eta) = 0$ (Exclusion Restriction)
3. $\gamma_i \geq 0 \ \forall i$ or $\gamma_i \leq 0 \ \forall i$ (Monotonicity assumption)$^2$
4. Z is conditionally independent of potential outcomes of D and Y.

If the relevance condition fails, then there is no first stage. Mathematically, this means that $\gamma = 0$, so the ILS estimator becomes undefined. The relevance condition isn’t an assumption. If Z doesn’t impact D, then you can’t do IV.

The Exclusion Restriction is analogous to the CIA in OLS. The exclusion restriction cannot be tested and must be evaluated based on your own judgement.

The Monotonicity assumption requires that Z always causes D in the same direction. This assumption is not necessary if we assume a constant effects framework, but it is important once we allow for heterogeneous treatment effects.

The conditional independence assumption is different from the exclusion restriction. Even if Z is randomly assigned, if it directly impacts Y then it violates assumption 2, but it satisfies assumption 4.

Most of the time, we are concerned about the exclusion restriction.

6.2.5 Example: Random Assignment

The ideal instrument is $T_i$, a randomly assigned treatment variable. Suppose we are interested in D, and T is a variable that randomly assigns observations to D=1 or D=0.

Note that T is a GREAT instrument. It has a very very strong impact on D, and it has no direct impact on Y.

The first stage is:

$$D = a + \gamma T + \zeta \quad \text{(First-Stage)}$$

In this case, $\gamma = 1$ since T dictates D.

The reduced form is:

$$Y = \alpha + \delta T + \epsilon \quad \text{(Reduced Form)}$$

$^2$This assumption is only required for the LATE interpretation discussed in section 6.2.9
The IV estimator is $\hat{\rho}_{IV} = \frac{\text{Reduced Form}}{\text{First Stage}} = \frac{\delta}{\gamma}$.

In this case, $\gamma = 1$ and the reduced form is the same as simply regressing $Y$ on $D$. As such, the IV estimator is the same as simple OLS in this case.

6.2.6 Example: Random Assignment with imperfect compliance

Suppose $T_i$ is randomly assigned, but not all individuals comply with their assigned treatment status. For example, in exploring the impact of a weight-loss drug, suppose that only 50% of the people assigned to take the drug actually take the drug.

Imperfect compliance creates a problem for the researcher such that simple means comparison between $D=1$ and $D=0$ will no longer be consistent.

Intuition:

<table>
<thead>
<tr>
<th>Group C</th>
<th>Randomly $T=0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group T</td>
<td>Randomly $T=1$</td>
</tr>
</tbody>
</table>

Since $T$ is randomly assigned, we know that the average characteristics of those in group C are the same as the characteristics of group T.

If everyone in group C has $D=0$ and everyone in group T has $D=1$, we simply compare mean outcomes to get the ATE.

Suppose, we have imperfect compliance such that when $T=1$, sometimes $D=0$. (we still assume that $D=0$ whenever $T=0$).

Now we have:

| Group C: $D=0$ | Randomly $T=0$ |
Within the T=1 box, D is not randomly assigned.

This means that group A and group B are not necessarily similar.

Q. —— Are group A and group C similar?

A. —— No! Group C is representative of the population, group A is the set of individuals who refuse treatment.

There are several bad comparisons

- Group A vs Group B is a bad comparison
- D=1 vs D=0 is a bad comparison (since group A+C is not comparable to group B)

There is only one clean comparison:

- The only comparable groups are T to C (A+B vs C)

Since the only comparable groups are T and C, we will compare average outcomes between T and C to help learn about the impact of D.

Note that this is a strange comparison since 50% of group T has D=0.

6.2.7 Intent to Treat (ITT)

The comparison of T to C is called the intent to treat (ITT). The idea is that we compare the randomly assigned treatment group (T=1) to the randomly assigned control group (T=0) regardless of their actual treatment status (D). In this regard, we are getting the impact of being randomly assigned to the treatment group. If there is perfect compliance, then ITT = ATE.

In many cases, it can be argued that the ITT is an interesting parameter.

Example: Suppose the government randomly gives students vouchers to attend better schools. Not all students will use the voucher, so compliance is imperfect. The ATE is the average test score change caused by attending a better school. The ITT is the average test score change caused by receiving a voucher. From the governments perspective, the ITT might be more relevant since it tells us the likely total impact of sending out vouchers (including the effects of non-compliance). That said, even in this context, the ATE is still very useful to know.
**Estimating the average treatment effect with imperfect compliance**

**Important Point:** If there is imperfect compliance it is not possible to get an estimate of the ATE. The reason is that we can only estimate the return to the treatment for group B.

Group B is not representative of the population. Group B is the set of individuals who take up treatment because they were told to.

The average treatment effect for group B is called the Local Average Treatment Effect or LATE. This terminology refers to the fact that the estimate is only relevant for a subset of the population. I am more specific regarding exactly how to think about group B and the LATE in section 6.2.9.

### 6.2.8 Recovering the LATE from the ITT.

Suppose the weight loss drug is randomly assigned and there is 50% compliance. Group T can be split into (A+B).

Comparing group (A+B) vs C, we find that C lost 15 pounds and (A+B) lost 25 pounds.

Thus, \( \text{ITT} = 10 \)

Since (A+B) and C are initially identical on average, the difference in outcomes must be coming from group B.

This means that group B is decreasing its own weight to such a large extent that the overall decrease in group T is 10. Since Group B makes up only 50% of group T, Group B must have lost 20 pounds.

Note that the estimate of 20 pounds is not the ATE since it only tells us the average benefit for Group B. Group B is not representative of the population since only the aggregate A+B is representative of the population.

If group B was only 25% of group T, it would have to lose 40 pounds in order to make group T lose 10 pounds.

In general, we have \( \text{LATE} = \frac{\text{ITT}}{P(D=1|T=1)} \) where the denominator is the fraction of people with \( D=1 \) in the \( T=1 \) group.

**This is the same as the IV estimator**

Even though T does not perfectly predict D, T is still a great instrument. It is strongly predictive of D and it has no direct impact on Y.

The first stage is:

\[
D = a + \gamma T + \zeta
\]

(First-Stage)

---

3In this particular case, the LATE is the same as the TOT, but this will not generally be true
In this case, $\gamma$ is the difference in D between treatment and control. In the above example, we have $\gamma = 0.5$. The estimate of gamma would be the same if 10 percent of the control group somehow obtained the pill ($D=1$) and 60 percent of the treatment group obtained the pill.

The reduced form is:
\[
Y = \alpha + \delta T + \varepsilon
\]  
(Reduced Form)

The IV estimator is $\rho_{IV} = \frac{\text{Reduced Form}}{\text{First Stage}} = \frac{\delta}{\gamma}$.

Notice that this is the same formula as $LATE = \frac{\text{ITT}}{p(D=1|T=1)}$.

Note that the Reduced form gets us the ITT, so the reduced form can be useful for policy makers in and of itself.

In the case where the instrument and the endogenous regressor are both binary, the IV estimator with no controls is also referred to as the Wald estimator.

Notice that the wald estimator is a simple ratio of mean differences:

\[
\rho_{IV} = \frac{\delta}{\gamma} = \frac{\bar{Y}_{T=1} - \bar{Y}_{T=0}}{\bar{D}_{T=1} - \bar{D}_{T=0}}
\]

**Take away point:**

In instrumental variables, the reduced form gives us the ITT.

We scale up the ITT to get the LATE.

### 6.2.9 Types of people in IV analyses

In order to understand the "Local" part of the LATE, it is necessary to consider 4 groups of people.

1. **Always-takers:** subpopulation with $D=1$ regardless of value of $T$.
2. **Never-takers:** subpopulation with $D=0$ regardless of value of $T$.
3. **Compliers:** subpopulation with $D=1$ if $T=1$ and $D=0$ if $T=0$.
4. **Defiers:** subpopulation with $D=1$ if $T=0$ and $D=0$ if $T=1$.

**Example with weight loss pill:**

1. **Always-takers:** Will find a way to get weight loss pill no matter what.
2. Never-takers: Will never take weight loss pill (in example from section 6.2.6 this is
group A and half of group C).

3. Compliers: Takes the pill if and only if T=1. (in example from section 6.2.6 this is
group B and half of group C).

4. Defiers: Take the pill if and only if T=0. (in the example from section 6.2.6 there are no
defiers)

Instrumental variable regression estimates the average treatment effect for the
compliers.

6.2.10 Example: (Angrist 1990) Veterans Draft Lottery

Suppose we want to understand the impact of being in the army on eventual wages. If we
compare the wages of veterans to non-veterans, we will get a biased estimate of the impact of
army on wages since becoming a soldier is endogenous.

Angrist uses institutional knowledge about the vietnam draft lottery to construct an
instrument. He notes that during the vietnam war, men were randomly given draft numbers
and those with low numbers were drafted first. Those with high draft numbers were never
drafted. Having a low draft number is the instrument for going to war.

Q. —- What is the first stage?
A. —- Regress veteran status on having a low draft number.

Q. —- What is the second stage?
A. —- Regress wages on predicted veteran status from the first stage.

Q. —- What is the reduced form?
A. —- Regress wages on having a low draft number.

Q. —- What is the ITT?
A. —- The ITT is the average impact of being assigned a low draft number on eventual
earnings.

Q. —- Who are the always-takers, never-takers, compliers and defiers?

- Always-takers are those who go to war even without being drafted.
- Never-takers are those who will not go to war, even if they have a low draft number.
  These individuals are either unfit for service, draft dodgers, conscientious objectors, or
  something similar.
• Compliers are those who go to war if they get a low draft number, but don’t go to war if they get a high draft number.

• Defiers probably don’t exist in this context. Defiers are people who sign up for the army if they have a high draft number, but refuse to go if they have a low draft number.

**LATE interpretation**

In order to be able to interpret the IV estimate as the treatment effect for the compliers, we need monotonicity and conditional independence of Z:

1. Z must be conditionally independent of the potential outcomes of D. In other words, the first stage must have a causal interpretation.

2. Z must have a monotonic impact on D.

In the case of a discrete treatment, the monotonicity assumption is the same as assuming that there are no defiers. (Z causes D to increase from 0 to 1 for the compliers. Monotonicity requires that Z not cause D to decrease from 1 to 0 for any individual.)

In most papers, the exclusion restriction assumption is the most dubious and the two assumptions necessary for a LATE interpretation are fairly plausible.

Under these assumptions, the IV estimator tells us the wage impact of joining the army for the compliers. In this case, I think the compliers are not a very unusual group, but still, the compliers are not representative of the average person. In some cases, the complying population for an instrument is unusual, and this should be kept in mind when evaluating the external validity of the estimator.

Note that in this case, the compliers are quite different than the treated individuals since many people volunteer for the army and these people are treated, but not compliers. Thus, in this case, IV estimates a LATE that is likely different than the TOT.

**Results from Vietnam Draft Lottery Study.**

First Stage: Having a low draft number increases veteran status by 16 percentage points. (19 percent vs 35 percent).

Reduced Form: Having a low draft number lowers earnings by around $400. This is the ITT.

Q. — Based on these numbers, can we infer the second stage estimate of the impact of going to the army on earnings?
A. \[ IV = \text{reduced form first stage} \] This means that Army = \[ \frac{400}{0.16} \]. Going to the army causes around a $2400 reduction in earnings.

The logic of IV:

- The reduced form is a regression of earnings on whether you have a low draft number. This regression shows that those with a low draft number have earnings that are $400 less than those with a high draft number.

- Angrist argues that your draft number itself does not impact your earnings, so why is there a correlation between earnings and draft numbers?

- Low draft numbers increase the likelihood of entering the army, so the difference between earnings of high and low draft numbers can be attributed to the impact of joining the army. The ITT is $400.

- But only 35% of those with low draft numbers actually went to the army!? And 19% of those with high draft numbers went to the army.

We can visualize this as before:

```
<table>
<thead>
<tr>
<th></th>
<th>T=0: High draft numbers (unlikely to be drafted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non Vets: D=0  (81%)</td>
<td></td>
</tr>
<tr>
<td>Veterans (19%): D=1</td>
<td></td>
</tr>
<tr>
<td>Non Vets: D=0 (65%)</td>
<td></td>
</tr>
<tr>
<td>Veterans (35%): D=1</td>
<td></td>
</tr>
</tbody>
</table>
```

Looking at the above figures, we can figure out the exact proportion of Always-takers, never-takers and compliers.

- The 19% of veterans in the T=0 group must be always-takers since T=0, but D=1.
  - Since the T=0 and T=1 group have identical characteristics, this means that 19% of the T=1 group are always-takers.

- Since 19% of the T=1 group are always-takers, but 35% of the T=0 group went to army, this means that 16% of this group are compliers.
  - Since the T=0 and T=1 group have identical characteristics, this means that 16% of the T=0 group are compliers.

- The remaining 65% of both groups must be never-takers.
Note that the above analysis rested heavily on the assumption that there are no defiers. Without this assumption we could not conclude that 19% of the T=1 group are always takers.

The above analysis makes clear that the T=0 group and T=1 group are identical except for the 16% of each group who are compliers. For the compliers, their veteran status is different in T=1 vs T=0. Everyone else is unaffected by T.

This means that $0.16 \times \text{LATE} + 0.84 \times 0 = \text{ITT}$.

- The $400 \text{ ITT}$ can be entirely attributed to the compliers. That is, the $400$ difference between those with high and low draft numbers is driven by the 16% that join the army as a result of their low draft number.
- Since 16% is about one-sixth, the compliers must have increased their earnings by about $6 \times 400 = 2400$. The compliers increase of $2400$ is sufficient to increase the whole T=1 group by $400$.
- The $2400$ is the IV estimate of the impact of joining the army on wages.
- This estimate is only relevant for the compliers. It is a LATE.

### 6.2.11 Split-sample IV

This is discussed in further detail in MHE, but the basic point is as follows:

The IV estimate is \[ \text{reduced form} \]

Nothing requires that these two equations are estimated based on the same sample.

Suppose we have a dataset on income and draft numbers, but we don’t know veteran status.

We also have a separate dataset on draft numbers and veteran status.

We can implement split-sample IV in this context.

### 6.2.12 Other covariates; continuous endogenous regressors

For convenience, I have analyzed everything assuming that we have no other covariates. Also, I have assumed that the endogenous variable is binary (0,1). The logic, intuition and implementation of instrumental variables is unchanged when we relax these assumptions.

If Z has no direct impact on Y, we don’t need to include extra controls for consistency, but it will help reduce the residual variation and improve the standard errors.

In some cases, it might be implausible that Z has no impact on Y, but conditional on covariates, the assumption might be more plausible. Just as in OLS, we can control for potential direct reasons that Z impacts Y. If we have $Z \rightarrow Q \rightarrow Y$ this means that Z is a bad instrument for D. However, if we control for Q, Z might be a good instrument.
Including covariate controls to improve consistency

Example: School vouchers might be randomly provided to certain families, but perhaps only poor families are eligible at all to receive vouchers. In this context, receiving a voucher is directly predictive of test score because poor students have lower test scores than richer students. Using school vouchers to instrument for school attendance would be a bad instrument in this case. However, if we can control for family income, then it’s possible that conditional on family income, vouchers are randomly assigned and the instrument would be good.

Continuous endogenous regressors

If D is continuous or the instrument is continuous, this doesn’t change anything.

Example: If we want to study the impact of years of education, we could use distance from your parent’s home to the nearest college as an instrument.

- The reduced form is a regression of wages on the distance from your parents home to the nearest college.
- The first stage is a regression of your years of education on the distance from your home to the nearest college.
  - Note that the first stage estimate not necessarily in the [0,1] interval anymore, but this is ok.
  - The interpretation of the first stage is something like: Increasing the distance to the nearest school by 100 miles increases your years of education by 0.7. (just an example; these numbers are made up)
- Who are the compliers, always takers, never takers?
  - Compliers are those who are induced to get more schooling because of the distance to the nearest school.
  - The term always-taker or never-taker doesn’t make as much sense with a continuous variable, but the key point is that there are some people whose education level is not impacted by the instrument.
- The monotonicity assumption assumes that there is no one induced to go to college because they live far away from college.

6.2.13 IV Implementation in Practice

Include covariates (Xs) in first stage and second stage.

If we have other covariates, we include these covariates as controls throughout the analysis. The first stage, second stage and reduced forms should all control for the other covariates. It is not ok to include the covariates in one stage and then exclude them in another stage of the IV analysis.

Do not “manually” do 2SLS
If you run the first stage and then plug in predicted values to estimate the second stage, you will get correct estimates, but incorrect standard errors. This is because the second stage won’t understand that $\hat{D}$ is estimated. Instead, you should use ivreg in Stata. Just as with OLS, the default standard errors that assume homoskedastic standard errors should rarely be used.

7 Instrumental Variables: Issues, Evaluation, and Testing

7.1 The weak instrument problem

7.1.1 1 endogenous regressor and 1 instrument

An instrument is weak if it has a small relationship with the endogenous regressor. This can be directly tested by looking at the F-statistic on the instruments in the first stage. Essentially, we are testing whether $Z$ impacts $D$, conditional on the covariates.

If you have a single instrument and a single endogenous variable, the weak instrument problem does not create bias per-se.

The most informative way to think about weak instruments is that they are instruments with very few compliers - that is, few people are impacted by the instrument.

Consider how this changes the analysis presented earlier. Suppose only 0.1% of people joined the army due to the draft, e.g. 0.1% of the population are compliers.

We have:

<table>
<thead>
<tr>
<th>Non Vets: D=0 (81%)</th>
<th>T=0: High draft numbers (unlikely to be drafted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Veterans (19%): D=1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Non Vets: D=0 (80.9%)</th>
<th>T=1: Low draft number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Veterans (19.1%): D=1</td>
<td></td>
</tr>
</tbody>
</table>

In this example, the reduced form is the difference between the earnings of T=0 and T=1. Suppose the difference in average wages between the two groups is $40. Just as before, we assume that this entire difference is due to a change in income for the compliers.

But in this case, the compliers make up such a small fraction that in order to generate a $40 increase in the whole average for T=1, the compliers’ income must have increased enormously. To be exact, the IV estimate would be $\frac{40}{0.001} = 40,000$.

If T is truely randomly assigned, and the sample sizes are big enough, the IV estimate is correct here. BUT, if there is any other reason that incomes differ between T=1 and T=0, these differences will be attributed to veteran status.
The notion that “other” differences get attributed to veteran status is always the concern in the IV context. The real problem here is that the weak first stage will make it so that small differences between T=0 and T=1 get blown up 1000 fold.

For example, suppose the compliers only account for $30 of the difference between T=0 and T=1 and the remaining $10 difference is due to some other factor. This will bias our estimate of the impact of joining the army by $10,000!

7.1.2 Multiple endogenous regressors

In most cases, you can only find one instrument, but in certain contexts you might have multiple instruments for the same endogenous regressor. In these cases, you can use identifying variation from both instruments at the same time and the model becomes over identified.

In general, having multiple instruments is a good thing, but if the instruments are weak, this has the potential to create bias, even if the instruments perfectly satisfy the exclusion restriction. This is an issue of bias vs consistency. It can be shown that the bias is an inverse function of the F statistic, and in the worse case, the IV estimate is biased towards the OLS estimate.

In general, IV estimators are biased, but consistent; they don’t perform well in small samples. Even in cases where the sample is large though, you might have issues of small sample bias because the effective sample in instrumental variables is just the size of the complying population.

The weaker the instruments, the fewer compliers there are. If you have many instruments and few compliers, this creates an effective small sample problem.

Unlike failures of the exclusion restriction, the weak instruments problem is fundamentally a technical statistical problem — not one with economic meaning. As such, the fix is a statistical one. I won’t get into details, but it has been shown that Limited Information Maximum Likelihood (LIML) performs well, even when instruments are relatively weak. That said, the more general problem of magnifying violations of the exclusion restriction cannot be solved with a technical fix.

A rule of thumb that many people use is that the F-statistic on the instruments should be larger than 10 to avoid weak instrument bias. Naturally this is not a law and in general the bias is just proportional to $\frac{1}{F}$.

When instruments are very weak, why is IV biased towards OLS rather than biased towards 0?

The intuition is that the IV randomly picks up some of the variation in the endogenous regressor and this is like a random sample of the X variation so we end up using variation that is representative of X (and so we are basically doing OLS). As the sample gets big, the IV will no longer be randomly picking up some of X since it will get closer to the true relationship between Z and X.
EX: What is the impact of education on wages?

Using OLS, we could compare the incomes of those with more education to those with less. To evaluate whether we would get bias, we need to consider the causes of education.

**Causes of education:**

- motivation
- intelligence
- family resources
- random factors

Many of these factors also cause wages, so there will be bias.

Suppose we randomly assign people into group A and group B, but these two groups are just randomly grouped (i.e. there is no treatment being assigned). Call the random assignment variable Z. Z is a very good, very weak instrument.

As the sample size tends to infinity, group A and group B will have the same level of education and income on average.

For small N, group A might randomly have more educated people than group B. If we use our instrument, we are comparing group A to group B. Suppose group A randomly has more education. They will have higher wages as well.

Q. --- What causes group A to have higher education compare to group B?

A. --- *People in group A have more motivation, intelligence or other factors that tend to cause more education.*

The factors that bias OLS are exactly the same as the factors that bias IV in this case since we are basically just comparing educated to uneducated people with no real instrument.

### 7.2 How biased are bad instruments?

The primary reason we get bias in IV is because of a failure of the exclusion restriction. If the instrument has a direct impact on the outcome, this will be wrongly attributed to the impact of D.

To assess how bad the bias is, you have to consider how large of a direct effect that Z has on Y. Note that even if Z impacts Q which impacts Y, this is considered a direct effect.

Bias in IV has the potential to be very large because any bias in the reduced form gets amplified when we divide by the first stage coefficient. This problem is most pronounced when the first stage is very weak, but it is a general concern.
7.3 The LATE interpretation

Remember that IV estimates the average treatment effect for the compliers!

This means that OLS and IV are not directly comparable since they estimate the treatment effect for different populations.

It is not possible to exactly identify who is a complier and who is not a complier, but you can identify the characteristics of the compliers. See the paper by Angrist and Fernandez-Val called ExtrapoLATE-ing: External Validity and Overidentification tests in a LATE Framework.

7.4 Testing your instruments

7.4.1 Hausman Tests and Overidentification Tests

Hausman test compares the OLS estimates of the impact of D to the IV estimates of the impact of D.

Overidentification tests are used when you have multiple instruments. An OverID tests whether the two (or multiple) instruments are getting you the same estimated impact of D.

Neither of these tests is particularly useful in a heterogeneous effects context. The reason is that the hausman test is misleading since OLS estimates the TOT and IV estimates the LATE. The OverID tests are misleading since two different instruments estimate 2 different LATEs, assuming the compliers are different for the 2 instruments.

7.4.2 Test whether Z predicts things that it shouldn’t predict

This is an example of a falsification test.

In the study on the impact of joining the army on wages, Angrist estimates the reduced form using wages prior to the draft (lagged outcome).

7.4.3 An invalid test: \( Y = \gamma Z + \beta D + \varepsilon \)

This test has the right intuition (testing if Z impacts Y conditional on D) but it is a bad test to run.

The problem is that D is endogenous, so controlling for D leads to biased estimates of the impact of Z.

<table>
<thead>
<tr>
<th>Example:</th>
<th>T=0: High draft numbers (unlikely to be drafted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non Vets: D=0 (81%)</td>
<td>Veterans (19%): D=1</td>
</tr>
</tbody>
</table>
Suppose we hold D fixed and examine the relationship between Z and Y. This is a flawed comparison because even though Z is randomly assigned, Z conditional on D is definitely not randomly assigned. This is why our estimate of the impact of Z on Y will be biased.

In this example, even if Z has no impact on Y, we will be find an impact on Y. Holding D=1, regressing Z on Y gives us a comparison of the always-takers (z=0) vs the compliers (z=1). These two groups have very different characteristics and may have different wages even if Z has no direct impact on wage.

If you have 2 instruments, you can run an alternative test where you use 1 instrument to get $\hat{D}$ and then estimate $Y = \gamma Z_2 + \hat{\beta} \hat{D} + \epsilon$. This is like an overID test where you need to assume one Z is valid in order to test the other Z. It’s a good idea to do it, but you don’t really learn that much since both instruments could still be wrong.

### 7.5 How I evaluate instruments

1. Pretend that you don’t know what the authors are trying to show.

2. Consider the reduced form: $Y = Z\delta + \epsilon$

3. If Z causes Y, ask yourself how it is possible for Z to impact Y.

4. If you can think of any reason that Z impacts Y other than through X, the instrument is bad.

### 7.6 How to find instruments

Main sources

1. Laws and institutional factors that create variation in D. This can be interacted with time in a powerful way if the laws change.
   
   (a) Ex. Introduction of a new law requiring that some employers provide health insurance could be an instrument for employers providing health insurance.

2. Geographic factors such as distance
   
   (a) Ex. People who live nearer to a college are more likely to go to college. Distance to nearest school could be an instrument for going to college. This can be interacted with time to make it more compelling

3. Lagged values of D.
   
   (a) Ex. Cities with historical immigration are more likely to have immigrants in time $t$. Historical immigration rates could be an instrument for time $t$ immigration.
These types of instruments rely on the idea that there is not serial correlation in the outcome.

4. Ex. Natural disasters, rain and other weather
(a) Used a lot in development context.

5. Unemployment rate as instrument for employment.

6. “Mistakes” i.e. census forecasting errors in population or administrative errors in who gets placed in treatment.

7. Bartik style instruments that interact national trends with initial conditions and include time and place fixed effects.

In general, all of the above instruments are subject to criticism in most contexts.

7.7 Best practices when writing IV based papers

7.7.1 First stage best practices
Always report the first stage estimates and think about these as you would an OLS regression.

Do the first stage estimates make any sense? Are the magnitudes reasonable?

Think about the fact that your reduced form will be scaled up by \( \frac{1}{\text{first stage}} \). If this scaling factor is very large (accounting for the variable scale) you might be concerned.

Make sure that your F-statistic on the excluded instrument is relatively large. A rule of thumb many people use is F should be larger than 10, but this is obviously not a theorem.

7.7.2 Run the reduced form
Always look at the reduced form. Make sure you see how the IV estimates are a ratio of the reduced form to the first stage estimates.

Think carefully about the interpretation of the reduced form. It is an estimate of the ITT.

To get the IV estimate, you will multiply the ITT by some factor. If the ITT estimate is zero, then the IV estimate will be zero.

7.7.3 Interpret the IV estimate results
You should think about who are the compliers. You can even identify the characteristics of the compliers based on the formula given in MHE.

Compare OLS to IV

- Remember that the reason that you do IV is because you are concerned that OLS is biased. Make sure to consider whether IV may be even more biased.

- Based on the form of the endogeneity in OLS, you should have a guess at the direction of OLS bias.
• Make sure to reconcile the direction of the OLS bias with the difference between the IV and OLS estimates.
  
  – If the omitted variable bias formula tells you that the OLS estimate is biased up, you expect to find that the IV estimate is smaller than OLS.
  
  – **When comparing OLS to IV remember that they estimate different parameters.** IV estimates the LATE, OLS estimates the TOT.
    
    * If IV gets you a different estimate from OLS and the difference contradicts your expectations, it could be that the LATE \(\neq\) TOT.
    
    * The compliers may be very different from the treated individuals. If you have reason to believe that the compliers would be differentially impacted by the treatment, then this can explain why IV gets a different estimate than unbiased OLS.
  
  – Example: Even though we would expect estimates of the income return to education to be biased up, many papers using IV find larger estimates compared to OLS. Some authors argue that in this case, the LATE > TOT because the compliers for their instruments are relatively disadvantaged and may get an unusually large benefit from getting an education.
  
  – Using a LATE \(\neq\) TOT argument to reconcile IV with OLS is sometimes a bit tenuous of an argument since it is hard to know when the compliers would be expected to have a larger or smaller benefit from the treatment.

• While you should try to reconcile your OLS and IV estimates, remember that one strong possibility is that your instrument is failing the exclusion restriction. Just as in OLS, think carefully about what direction you expect the bias to be from failures of the exclusion restriction.

To formally test the difference between OLS an IV, you can run a Hausman test. This is simply a test of whether OLS and IV get different estimates.

The Hausman tests is supposedly a test of whether D is endogenous.

I don’t think the Hausman test is particularly worthwhile for the following reason.

Suppose the OLS and IV estimates are statistically different from one another according to the Hausman test. There are then 3 possibilities:

• The regressor of interest (D) is endogenous

• The LATE \(\neq\) TOT but D is *not* endogenous.
  
  – It could be that OLS correct estimates the TOT and has no bias.
  
  – It could also be that IV correctly estimates the LATE and has no bias.

• The instrument could be failing the exclusion restriction so that IV is biased.

Since there are multiple reasons why one would reject a Hausman test, it seems wrong to interpret a rejection of the Hausman test as evidence of D being endogenous.
The only reason to every report the Hausman test is if you are trying to argue that D is exogenous. In this context, you can show that OLS and IV give similar estimates.

7.8 Examples!

1. Impact of health expenditures on mortality. Instrument is whether a ambulance with a relationship with a high cost hospital picked you up.


4. Impact of sexually transmitted disease prevalence on risky sex behaviors. Instrument: distance from your home to the origin of the AIDS virus

5. Impact of older sister’s education level on younger brothers test scores. Instrument: distance of family to nearest girls school.


8. Impact of adult depression on wages. Instrument: depression measured as a teenager instruments for adult depression.

7.8.1 You can only have 1 first stage!

If we have 2 first stage then we are attributing the impact of Z on Y to two different variables, $D_1$ and $D_2$. When instrumenting for $D_1$, we need to assume that Z only impacts Y through $D_1$. When instrumenting for $D_2$ we need to assume that D only impacts Y though $D_2$. These two assumptions are contradictory.

8 Regression Discontinuity Design (RDD)

The main problem in establishing causality is that the treated individuals D=1 are often different from the untreated individuals D=0.

The RDD design focuses on cases in which D status is determined by an arbitrary cutoff.

When done well, RDD is one of the most convincing empirical methods in terms of internal validity.


The idea behind the RDD design is that some laws or rules lead to discontinuous changes in outcome.

Examples:

- Course grades — 89 gets B, 90 gets A
• Elections — 50.1% leads to winning, 49.9% leads to losing.
• Laws with age cutoffs — Aged 21 can drink, aged 20 cannot legally drink
• Merit scholarship — Students above 95th percentile on SAT get scholarship.

These discontinuous cutoffs lead to people with similar Xs getting different D.

The basic idea is that we assume that people very close to the cutoff have treatment status essentially randomly assigned. This means that we can simply compare outcomes to get an unbiased estimate.

RD terminology:

**Running variable** (also called forcing variable): This is the continuous variable which determines treatment. e.g. age is the running variable in the legal drinking age example.

**Cutoff**: The level of the running variable at which the treatment status changes.

There are two different types of RD designs.

1. **Sharp RD** — Treatment status is completely determined by running variable.
2. **Fuzzy RD** — Treatment status is partially determined by running variable.

### 8.1 Sharp RD

In sharp RD, treatment status is completely determined by the discontinuity.

**Example: Lalive 2008**

Austrian Unemployment benefits experiment: Austrian government in 1988 extended UI benefits for people over 50 living in certain regions of Austria from 30 weeks to 209 weeks.

Causal question: Does the availability of 209 weeks of benefits increase unemployment duration?

Q. —— What is the forcing variable

A. —— Age (or distance to border)

Workers who are 49 only qualify for 30 weeks of unemployment. Workers who are 50 qualify for 209 weeks unemployment.

Since age perfectly predicts the maximum number of weeks unemployment, this is a sharp RD.

The sharpness can be seen in a picture.
Comparing unemployment duration for people aged 49 vs 50 tells us the impact of unemployment benefits on duration.

All we need to assume is that 49 and 50 year olds are otherwise similar.

Average duration for 49 year olds was 13 weeks. For 50 year olds, average duration was 27 weeks.

To make this even more convincing, we can show a picture.

This is very compelling evidence that the policy impacted peoples unemployment duration.

The RD estimate is the vertical distance between the outcomes right at the cutoff.

8.2 Implementing RD
Notation:

\[ \delta: \text{Causal impact} \]
\[ x: \text{running variable} \]
\[ c: \text{The cutoff} \]
\[ \delta(c) = \lim_{x \downarrow c} E(Y_1|X = x) - \lim_{x \uparrow c} E(Y_0|X = x) \]  \hspace{1cm} (19)

If data were perfectly continuous, we could just compare outcomes just to the left and just to the right of the cutoff.

In practice, \( \delta(c) \) needs to be estimated for two reasons.

- Data is always discrete e.g. we cannot observe whether a person is 49.9999.
- Data are noisy. If few people are exactly around the cutoff, we don’t want to base our entire estimate on just 2 people.

In practice, we must estimate the left limit and the right limit.

Suppose we have data on age in years.

In this case, we do not observe anyone who is 49.9 years old.

We need to estimate the left and right limits.

The idea behind RD is simple, but there are many more complicated methods to estimate the left and right limits. In practice, the particular method you use should not be driving the estimates. You should be able to see the discontinuity estimate without running any regression at all.

There are 3 methods commonly used for estimating the discontinuity.

1. Parametric
2. Parametric with excluding observations with X far away from cutoff.
3. Non-parametric
8.3 Parametric Estimation

8.3.1 Linear estimation

We can estimate the left and right limit by linearly extrapolating the trend on the left and the right separately.

The projected outcomes at age 50 are our estimate of the left and right limits.

The easiest way to estimate this gap is as follows:

1. Subtract off the cutoff score from the running variable to center the cutoff at zero. Call the centered at zero running variable S.

2. Estimate a regression: \( Y = \alpha_2 + \gamma_2 S \) For \( S \geq 0 \). Note that \( \alpha_2 \) is the estimate of the intercept at \( S=0 \).

3. Estimate a regression: \( Y = \alpha_1 + \gamma_1 S \) For \( S < 0 \). Note that \( \alpha_1 \) is the estimate of the intercept at \( S=0 \).

We estimate \( \delta = \alpha_2 - \alpha_1 \).

In practice, there is an even easier way.

First, create an indicator for \( Above = 1 \ if S \geq 0, \ Above = 0 \ if S < 0 \). Then simply estimate:

\[
Y = \alpha + \delta Above + \gamma S + \eta (S \ast Above) \tag{20}
\]

Notice that in this equation:

- When \( S \) is just below zero, we have \( Y \approx \alpha \).
- When \( S \) is just above zero, we have \( Y \approx \alpha + \delta \)
The equation allows the slope to be different on the left and the right of the cutoff. The slope on the left is $\gamma$ and the slope on the right is $\gamma + \eta$.

The estimate of the discontinuity is simply $\delta$. Regression provides the standard errors on $\delta$ here!

Sometimes, linearity is not a good assumption and can lead to poor estimates of the discontinuity.

Example: Looking at the picture below, there does not appear to be a discontinuity, but the linear extrapolation method will estimate a discontinuity.

Note: The X axis is the running variable here. In the regression, you have recentered the running variable so that 50 corresponds to 0. In RD figures, you can either center things around zero (since this corresponds to the regression) or you can artificially recenter things back at the original cutoff.

This discontinuity seems to be driven by poorly modeled non-linearity on the right of the cutoff.

In this case, the figure makes it clear that we should be using a more flexible functional form, for example, a quadratic, or a cubic.

8.3.2 Global Polynomial

To do this, we estimate:

$$Y = \alpha + \delta Above + \gamma S + \gamma_2 S^2 + \eta(S * Above) + \eta_2 (S^2 * Above)$$  \hspace{1cm} (21)

Note that for consistency sake you should use the same functional form on the left and the right.

This will yield an estimate closer to zero, and the figure will look like.
We could make the polynomial more and more flexible and potentially get a better and better fit.

The key is to use a polynomial that is sufficiently flexible, but does not ”over fit” the data. You should always look at the picture to get a sense of whether your parametric fit makes sense.

Below is an example, where we could get a very bad estimate if we use a 9th order polynomial. This is made up, but if you have a higher order polynomials (particularly odd polynomials) they could have weird extrapolation issues.

Ideally, your estimate will not be terribly sensitive to the choice of polynomial. If the estimate is sensitive, you want to make sure that your preferred estimate looks reasonable in the figure.

8.4 Parametric with excluding some observations.

In order to get a more reasonable fit, instead of trying to fit a polynomial to the entire range, we can potentially do better by focusing in around the discontinuity.
This is estimated exactly the same as the parametric regressions above, but we exclude observations far from X=c.

### 8.5 Non-parametric estimation

#### 8.5.1 Local Linear Regression

Local linear regression is a more general technique than its use in RD. Despite its name, local linear regression is a non-parametric method for estimating a functional form.

The main reason to do local linear regression is if you really care about the exact functional form. That is exactly the case in RD.

Local linear regression method:

1. At each s, we estimate a linear regression in the neighborhood of s.
2. The size of the neighborhood is called the bandwidth and is denoted by h.
   \[ Y = \alpha_1 + \gamma_1 S \text{ if } S - h \leq s \leq S + h \]
3. We estimate the local linear regression separately on the left and on the right side of the cutoff.
   (a) DO NOT estimate a linear regression with S=0 in the neighborhood \( \pm h \)
   (b) Estimating at S=0 would try to estimate a continuous slope at S=0 which contradicts the idea of RD.

Graphically, if we use a very big bandwidth, local linear regression will look like:
As we narrow the bandwidth, the local linear regression will fit the data better and better.

The above figures are not actual local linear regressions, but just my sketches of the idea,

8.5.2 Local Linear Regression with Kernel smoothing

In order to make the local linear regression match the data more closely, people frequently use kernel weighting in their local linear regression.

Instead of estimating a simple linear regression at each \( s \), we estimate a kernel weighted regression at each \( s \).

We once again separately estimate a series of local linear regressions on either side of the cut off.

\[
Y = \alpha_1 + \gamma_1 S \text{ if } S - h \leq s \leq S + h
\]

In each regression, \( S=s \) gets the most weight. \( S-h \) and \( S+h \) get the least weight.

We can use a variety of different weights, e.g triangle, Epanechnikov, rectangular.
Rectangular is the easiest weight to implement since this simply puts equal weights on all the observations and is equivalent to not using weights at all.

If we use a rectangular kernel (as is recommended by Lee and Lemieux), the local linear estimate becomes equivalent to parametric regression with excluding observations. This is because only the linear regression near the discontinuity actually matters. For presentational purposes, you might show then entire dataset in your figure, but for local linear regression, the only data that matters is the observations around your discontinuity.

8.5.3 A needlessly complicated phrase: local linear regression with a rectangular kernel and bandwidth h

This is a recommended method for estimating RD and the implementation is incredibly simple. We just estimate:

\[ Y = \alpha + \delta \text{Above} + \gamma S + \eta (S \times \text{Above}) \text{ if } -h < S < h \]  

8.5.4 Bandwidth selection

Regression discontinuity is much more transparent than most other methods because there are relatively few choices for the researcher to make.

The one exception is that you need to choose the polynomial order if you are doing global polynomial fitting, or you need to choose the bandwidth and kernel if you are doing local linear regression.

There are formal ways of determining the “optimal” bandwidth, but I think it’s usually better just to fit the discontinuity you judge to be there from the figure. Not everyone agrees with that though.

Identifying discontinuities visually takes some judgement. In most cases it won’t be super clear.

8.5.5 Judging discontinuities in figures

Q. — For each of the following graphs, do the following:

1. Evaluate whether there is a discontinuity or not.
2. Then think about which estimation strategy will best fit it.
3. Think about how robust the estimate will be to different estimation choices.
Based on my judgement, the first 2 graphs have clear discontinuities. The 3rd and 4th arguably have discontinuities and I would argue that they do. The 5th one clearly has no discontinuity.

I base my evaluation on whether it looks like there is a structural shift at the threshold. Not everyone will agree with my judgement on this.

### 8.6 Fuzzy RD

In fuzzy RD, the idea is the same, but the running variable doesn’t perfectly predict treatment.

Example: The impact of health care at birth on academic achievement.

If we run OLS of achievement = health care, this will likely be biased since children who get
health care at birth are sicker than other children.

Bhatawagi (2013 August AER) notes that doctors use a rule of thumb to decide whether to provide certain infant healthcare.

1. Infants born below 1500g should get intervention.
2. Infants born before 32 weeks should get intervention (regardless of weight)

This allows for a regression discontinuity design.

Q. —— What is the running variable?

A. —— *Birthweight*

The idea is to compare infants who weight 1499 to infants who weight 1501

Q. —— Why is this a fuzzy RD?

A. —— *The running variable does not perfectly predict health intervention because doctors use their best judgement in close cases.*

A fuzzy RD context is just like random assignment with imperfect compliance.

Just like IV, there are two stages:

1. First stage: RD analysis of how running variable impacts D (treatment variable)
2. Reduced form: RD analysis of how running variable impacts Y (outcome variable)

What is the RD estimate in the fuzzy case?

\[
\delta = \frac{\lim_{x \downarrow c} E(Y_1 | X = x) - \lim_{x \uparrow c} E(Y_0 | X = x)}{\lim_{x \downarrow c} E(D_1 | X = x) - \lim_{x \uparrow c} E(D_0 | X = x)}
\]  

(23)

Example:
We can see the first stage and second stage graphically.
8.7 RD issues

The key assumption in RD is that individuals on either side of the cutoff are similar. More formally, this is an assumption about continuity of potential outcomes at the cutoff.

The internal validity of RD fails if individuals on one side of the cutoff are systematically different than individuals on the other side.

This can occur for 2 reasons

1. The cutoff is used for other policies as well.
   - Example: For the distance based RD with UI benefits discussed above, it would be problematic if the borders designated county lines, or school districts as well. This would result in people on either side of the cutoff being impacted both by the UI benefits and other policy differences.

2. The more common issue is that there is manipulation of the running variable.
   - Example: Suppose students with a 3.0 GPA get a scholarship and those with a 2.9 GPA lose their scholarship.
This suggests that we could do an RD examining how scholarships impact dropout rates.

This is not terrible, but there could be concerns that those with a 3.0 GPA are very different than those with a 2.9 GPA.

Example: Anti-discrimination law only applies to firms larger than 14

Do we think that firms of 15 people will be similar to firms of 14 people?

### 8.8 RD empirical testing

One of the big appeals to RD is that there are great tests to explore the potential threats to identification.

The most common concern is that agents can manipulate their running variable in order to get treatment or control.

If running variables can be manipulated, this means that treatment status will be endogenous, even in a narrow range around the cutoff.

The empirical tests are very direct and very convincing.

**Empirical tests**

1. Examine the histogram smoothness of the running variable around the discontinuity.

2. Examine the smoothness of many other covariates around the discontinuity.

#### 8.8.1 Histogram test

**Q.** If there is manipulation of the running variable for the firm discrimination example, what would the histogram of firms size look like?

**A.** Very few firms at 15, 16 and many firms at 13, 14

If there is no manipulation, the histogram should be fairly smooth.

Mccrary (2008) has a formal test of whether the histogram appears smooth. Even if you think that there is a possibility of manipulation, if you pass the histogram smoothness test, this is convincing evidence that there is no manipulation of the running variable.

#### 8.8.2 Covariate smoothness test

Showing covariate smoothness in other variables around the discontinuity makes the RD design very convincing.

Example from the study of the impact of hospital interventions on academic performance.

To argue that children on either side of 1500g threshold are similar, we look at their covariates on either side.
8.9 RD is not a before and after comparison

Suppose a policy to reduce pollution is implemented in California. We cannot do regression discontinuity using just before the policy vs just after the policy.

The spirit of RD must theoretically \( \lim_{x \to c} \). In this case, comparing one second before the policy to one second after the policy makes no sense.

If we expand the time frame to look 1 year before vs 1 year after, we no longer are sufficiently local to be able to assume nothing else changed.

If you are in this context, a DID design is more appropriate. (will be discussed in a few weeks).

8.10 RD estimates a LATE

Just like IV, RD estimates a Local Average Treatment Effect (LATE). In this case, the estimate is the average benefit of treatment for people near the cutoff. This is clear since given enough data, the RD estimate can be obtained by just comparing average outcomes just to the left and just to the right of the cutoff. This suggests that only people near the cutoff contribute to the estimate.
Example: RD analysis of returns to remedial education.

When students enter UIC, they need to take a placement exam and if they score too low, they are required to take remedial courses.

Supposing the cutoff is 50, we can compare the graduation rates of students with a 49 to students with a 51 in order to get an RD estimate of the impact of remedial courses on college completion.

Several papers have used this estimation strategy at other schools and found a zero effect of remedial coursework.

Q. Do these results suggest that remedial education should be abandoned?

A. Not necessarily, the RD analysis suggests that people on the margin of 50 get no benefit to remedial education compared to entering directly into college courses. It is possible that these students nearly have the skills necessary for college courses and learn relatively little in the remedial track. This would explain an estimated zero benefit of the remedial coursework. At the same time, it could be that there are very big benefits to remedial coursework for people with very poor preparation (e.g. someone who scores a 20 on the placement exam).

In this context, the RD estimate is still an internally valid estimate of the impact of remedial education on people local to the 50 cutoff. It is important to keep the LATE interpretation in mind when thinking about policy implications.

8.11 Regression Kink Designs
Regression Kink is a similar idea to RD, but there is a discontinuous change in the slope rather than a jump in the level.

Regression kink is a relatively new empirical method (Card and coauthor paper on method just published in Econometrica in 2015).

Various other papers used the method informally before 2015, but still relatively uncommon.

This may mean that relative to other empirical methods, there are more potential projects not yet done that use this method.

Example: Suppose we want to look at the impact of parental income on children’s test scores.

After tax income vs before tax income. The impact of income on child test scores is the opposite of the impact of the amount of taxes paid.

Because of this, I will talk about the impact of taxes paid and the impact of income as basically the same concept.
For simplicity, I will assume in all of the following that people’s earnings do not respond endogenously to the taxes they will pay.

In general, as earnings increase, test scores rise for 3 reasons:

- Partly because of the causal effect of earnings
- Partly because of unobservables correlated with earnings
- Partly because as earnings rise, taxes rise. This relationship has sharp changes in the slope as shown below.

Notice that taxes never discontinuously jump but the derivative discontinuously jumps.

In this exaggerated example, the slope is 1 once you get to earnings of around 8. This means that for the 9th and 10th dollar earned, the tax rate is 100% and income doesn’t increase at all.

Running, Outcome and Endogenous regressor

- Running variable is before tax earnings.
- Outcome variable is test scores
- S’s of taxes paid is D (the first stage outcome)

Suppose we see that the relationship between test scores and before tax earnings shows similar kinks as shown below.
We should notice several things in the above graph.

- In the above graph, we see that in general test scores rise with earnings. This may reflect any of the forces discussed above.
  - Even at earnings levels 8.9 and 10, test scores still rise a little with earnings even though total income isn’t increasing much in this region.
    - This likely reflects the fact that higher earnings is correlated with things like ability that will genetically be passed on to children and improve their test score.
    - This is what I meant earlier by an unobservable correlated with earnings.

- The sharp changes in slope are unlikely to be the unobservable correlates or the direct causal impact of earnings since these don’t change sharply at the kinks.
  - The key insight is that these changes in slope are driven by a causal impact of taxes.
  - My discussion above acts as if the underlying relationship between earnings and test scores would have been linear (same slope everywhere). This need not be the case. The key question is whether there are discontinuous changes in the slope.
  - If taxes have no causal impact on test scores, we should see smoothly changing slope.

Exaggerated example: Consider the first stages and reduced form shown below. There is a single kink.

This is an extreme example where taxes go from 0 to 50% once you earn $5.

- Based on the reduced form, we see that in the region below $5 earnings, test score rises at a rate of 1 points per $1 in earnings.
  - This is the unobservable correlates of earnings effect and there is no taxation effect.
- Based on the reduced form, we see that in the region above $5 earnings, test score rises at a rate of 0.5 points per $1 in earnings.
  - This is due to the unobservable correlates of earnings effect plus the fact that taxes rises to 50 cents for every $1 increase in earnings.
In other words, a $1 increase in earnings will increase test scores, but at the same time, it increases taxes by 50 cents per dollar.

- The increase in test scores will not be as large as it would have been had there been no taxes.

• What exactly is our estimate of the impact of taxes on test scores then?
  - The test/earnings slope changes by 0.5 because we went from tax rate of 0 to tax rate of 0.5.
  - That suggests that a 0.5 change in the tax rate generates a 0.5 change in the test/earning slope.
  - This suggests that moving the tax rate from 0 to 1 (0% taxes to 100% taxes) would result in a 1 point reduction on the test per dollar earned.
  - In other words, the impact of 1 extra dollar in taxes is 1 point worse the test.

• The exact estimate is the ratio of the reduced form change in slope to the first stage change in slope. In this case 0.5/0.5 = 1.

This intuition is a little tricky; the math can help strengthen this intuition.

\[
\frac{\partial \text{Test}}{\partial \text{Earn}} = \frac{\partial \text{Test}}{\partial \text{Earn}|\text{Tax}} + \frac{\partial \text{Test}}{\partial \text{Tax}} \frac{\partial \text{Tax}}{\partial \text{Earn}} 
\]

This equation shows that the slope between Test and earnings can be thought of as 3 components. Notice that on the LHS, \( \frac{\partial \text{Test}}{\partial \text{Earn}} \) refers to the total effect of changes in earnings on test scores, but on the RHS, \( \frac{\partial \text{Test}}{\partial \text{Earn}|\text{Tax}} \) refers to the effect of earnings on test scores, holding fixed taxes.

We can consider the slope between test and earnings in two sections of earnings where the tax rate changes. Call the first region, region 1 and the second region, region 2.

\[
\frac{\partial \text{Test}_1}{\partial \text{Earn}_1} = \frac{\partial \text{Test}}{\partial \text{Earn}|\text{Tax}} + \frac{\partial \text{Test}}{\partial \text{Tax}} \frac{\partial \text{Tax}_1}{\partial \text{Earn}_1} \\
\frac{\partial \text{Test}_2}{\partial \text{Earn}_2} = \frac{\partial \text{Test}}{\partial \text{Earn}|\text{Tax}} + \frac{\partial \text{Test}}{\partial \text{Tax}} \frac{\partial \text{Tax}_2}{\partial \text{Earn}_2} 
\]

In the above two equations, I only put subscripts on things that are different in the 2 regions.

• The tax rates are different \( \frac{\partial \text{Tax}_1}{\partial \text{Earn}_1} \neq \frac{\partial \text{Tax}_2}{\partial \text{Earn}_2} \)

• The total impact of earnings on test scores could be different since the relationship between earnings and taxes is different.

If we take the difference between the equations for the two periods we get:
\[
\frac{\partial \text{Test}_1}{\partial \text{Earn}_1} - \frac{\partial \text{Test}_2}{\partial \text{Earn}_2} = \frac{\partial \text{Test}}{\partial \text{Tax}} \frac{\partial \text{Tax}_1}{\partial \text{Earn}_1} - \frac{\partial \text{Test}}{\partial \text{Tax}} \frac{\partial \text{Tax}_2}{\partial \text{Earn}_2}
\]

\[
\frac{\partial \text{Test}_1}{\partial \text{Earn}_1} - \frac{\partial \text{Test}_2}{\partial \text{Earn}_2} = \frac{\partial \text{Test}}{\partial \text{Tax}} \left[ \frac{\partial \text{Tax}_1}{\partial \text{Earn}_1} - \frac{\partial \text{Tax}_2}{\partial \text{Earn}_2} \right]
\]

Simplifying yields our estimate of the impact of taxes on test scores based on the RK model.

\[
\frac{\partial \text{Test}}{\partial \text{Tax}} = \frac{\frac{\partial \text{Test}_1}{\partial \text{Earn}_1} - \frac{\partial \text{Test}_2}{\partial \text{Earn}_2}}{\frac{\partial \text{Tax}_1}{\partial \text{Earn}_1} - \frac{\partial \text{Tax}_2}{\partial \text{Earn}_2}}
\]

- Numerator is change in slope estimated by reduced form.
- Denominator is change in slope estimated by first stage.

In practice, we get these estimates based on estimating the same regression as we had for the RD, but we are interested in a different variable.

For the RD we estimated:

\[
Y = \alpha + \delta \text{Above} + \gamma S + \eta (S \ast \text{Above})
\]  

(25)

The discontinuity estimate is $\delta$. For the RK, we are interested in the change in the slope, given by $\eta$.  

90
9 Panel data

9.1 Review: Dummy variables and FWL

Recall that instead of multivariable OLS, we can do things in two steps using FWL.

\[ Y = \gamma D + \beta X + \epsilon \]  

1. regress D on X and capture the residuals.
2. regress Y on the residuals.

Thinking of things in this way is particularly useful when X is a dummy variable. Suppose X=1 is female and X=0 is male. The first step in FWL is:

\[ D = \alpha + \eta X + \nu \]  

Note that \( \alpha \) is the male average of D and \( \eta + \alpha \) is the female average of D.

For women we have: \( \nu = D - (\eta + \alpha) \).

For men we have: \( \nu = D - (\alpha) \).

This example demonstrates that the residual \( \nu \) is simply the demeaned values of D.

In the second step of FWL, we simply regress Y on the demeaned values of D.

This provides very important intuition: Controlling for a dummy variable is the same as demeaning all the other covariates according to that dummy variable.

Example: We want to know the impact of height on weight, holding fixed gender.

We can estimate this three different ways:

1. Multivariate OLS:

\[ Weight = \beta Height + \eta Female + \epsilon \]

2. Demean Height by gender, then run bivariate OLS.

\[ Weight = \beta (Height_i - \text{average height for gender}) + \epsilon \]

3. Demean Height by gender, demean weight by gender. Run bivariate OLS.

\[ Weight - \text{averageweight for gender} = \beta (Height_i - \text{averageheight for gender}) + \epsilon \]

Remember this intuition moving forward.

Definitions:
**Within Variation:** Variation in D and Y that occurs holding the person fixed.

**Between Variation:** Variation in D and Y that occurs across people.

In the plot below, I have labeled the within variation and the between variation for height.

If nobody changes their union status, then there is no within variation and all between variation.

If everyone is in a union one period and then leaves the union, there is no between variation in D.

Q. — Do you think there is usually more between or within variation in most contexts?

A. —* Most settings have more between variation since there are bigger differences between different people than between the same person in 2 time periods*

### 9.2 Between and Within Variation

Example:
Panel data on a representative sample of firms from 1970 to 2010.

What is the impact of having more female board members on firm profits?

To think about this question, we want to think about the causes/correlates of D.

Q. — What causes a company to have more female board members

1. Later years will have more female board members
2. Smaller firms might have more female board members
3. More progressive firms have more female board members
4. etc..
For simplicity, suppose we have just 3 firms. We can plot the data.

Q. —- Is there more between or within variation in this example?

A. —- *Looks like a bit more between variation, but they are similar in size*

Q. —- Based on these two firms, do female board members help or hurt profits?

A. —- *Overall, it appears that female board members hurt profits since the slope is downward, but this masks important variation*

If we use ONLY between variation, then we will focus exclusively on variation across firms. In practice, this is done by averaging all the data by firm and running a regression using the averages (the number of observations will be N instead of NT).

Graphically, we would be doing:

**Between Variation**

If we use ONLY within variation then we focus entirely on data within each firm and we ignore differences across firms. This is what fixed effects does (demeaning D and Y).
Graphically, we would be doing:

**Within Firm Variation**

Q. —— Does it seem like female board members help or hurt? Is the between or the within variation better?

A. —— *Not obvious. You will get very different conclusions with between vs within, but it’s not clear which you should choose.*

The within variation could be driven by aggregate time changes, e.g. perhaps firm profits have just gone up over time as have proportion female:

I would not want to conclude that women help firm profits based solely on this type of variation.

### 9.3 Different levels of fixed effects

In this example, we can include year fixed effects instead of firm fixed effects.

\[ Y_{it} = \beta D_{it} + \gamma X_{it} + \delta_i + \epsilon_{it} \]
This means that we are demeaning by year and are then running ols on the demeaned data.

We could plot this by removing the time effects. For simplicity I have only plotted 2 years of data.

**Within Time Variation**

![Graph showing within time variation](image)

Notice that in the above graph, the upward trend caused by the time variation is completely gone.

**Q.** Does time fixed effects eliminate the between or within firm variation?

**A.** We still are using all of the between firm variation here. We are using some of the within firm variation, but not the part explainable by time. We are only using within time variation.

Focusing on within time variation makes sense here since we don’t want to conflate social change over time with the impact of female board members. Using all of the between firm variation worries me since firms that are more progressive may have less profits (perhaps because they are too nice)

### 9.4 Multiple fixed effects

As if this isn’t getting complicated enough!

We now add time AND firm fixed effects at the same time.

\[ Y_{it} = \beta D_{it} + \gamma X_{it} + c_i + \delta_t + \epsilon_{it} \]

This is too complicated to precisely show graphically, but there is a bastardized way to show the graph that gets at the intuition. Remember, the analysis that follows is not literally what multiple fixed effects does, but it is a useful way to think about it.
We take out each fixed effect sequentially. It doesn’t matter the order.

First we take out the time fixed effects and we get:

**Within Time Variation**

![Graph showing within time variation](image)

Now, using this plot, we take out the firm fixed effects by demeaning this data by time.

![Graph showing demeaned data](image)

We use the remaining variation to determine the relationship of interest.

Note that we are using the intersection of within time variation and within firm variation.

**Another way to think about this**

**We are using within firm variation that is not explained by aggregate differences across years.**

The reason that the above is not a precise description of what the double fixed effect is doing is because when we demean by time, we are demeaning the firm fixed effects as well. Thinking about this is confusing. In practice, the intuition I described above is how you should think about things – just don’t estimate things that way!
When you use Stata to estimate a model with 2 levels of fixed effects, you can directly include one set of dummy variables, and use xtreg or areg to deal with the other dimension of fixed effect. Another option is the user written program: felsdvreg. This program is ideal if you have 2 sets of fixed effects, both with very high dimension.

There is not that much variation left! In order to get consistent estimates, we need to assume that the remaining variation is clean.

In other words, we still need to assume CIA, but now we are holding fixed a huge amount of the variation.

9.5 Is there any variation left? (perfect colinearity)

9.5.1 Firm by Year fe vs Firm and Year fe

The above analysis was based on firm and year fe. This means that we are including dummy variables for each year and each firm. With 40 years of data and 1000 firms, we have 1,040 fixed effects. Including firm-by-year fixed effects means something very different. This would include a separate dummy variable for every year-firm combination, e.g. 40,000 fixed effects. With firm-year fixed effects, we only use variation that occurs within a firm-year combination. This demeans the annual data and we end up with every point sitting at zero – no variation left!

9.5.2 Too many fixed effects

Suppose we are estimating a model with firm and year fixed effects. Which of the following controls can we include:

- Industry fixed effects
- Firm size
- National anti-discrimination law
- Illinois anti-discrimination law

9.5.3 Observable controls

From a mechanical perspective, fixed effects are no different than other controls. As such, usually we include observable covariates together with fixed effects. Which observables might we want to control for in a model with firm and year fixed effects?

- Firm size (though this might be a partly bad control)
- Publicly traded
  - Can only be included if firms change from privately held to public
  - Might also be thought of as a bad control.
9.6 Thinking about identifying variation: Example 2

In each panel setting, we have within and between variation. You need to think carefully about whether there is any clean variation and then you try to control for all the other variation. There is no procedure for this process. Every context will be different.

If the within variation seems clean, do fixed effects. If the between variation is clean, aggregate the data.

**Example 2: Does taking an advanced placement (AP) course in economics during high school improve college graduation?**

Suppose we have data on 500 suburban high schools and 1000 students per school for 11 years (2000-2010). This is a 2 dimensional panel in which N=500 and T\(_1\)=11 years and T\(_2\)=1000 students. Even though T\(_2\) is larger than N here, we should conceptually think of doing asymptotics in N here since as we get more data, number of students per school is relatively fixed, but N can go towards \(\infty\).

Suppose that in 2005 many high schools start offering AP economics.

**Causal Question:** Impact of taking AP economics course on college graduation.

Sources of Variation in who takes AP courses:

1. Only the smartest students at each high school take AP courses
2. Only certain high schools offer AP economics.
   - Some high schools offer other AP courses.
   - Some high schools offer no AP courses.
   - Some high schools begin offering AP economics in the year 2005.
3. Only the most motivated students at each school take AP courses.
4. etc.

OLS would estimate:

\[
\text{Graduate}_{ist} = \gamma \text{AP}_{ist} + \epsilon_{ist}
\]

(28)

Having lengthy subscripts is common in multi-level panels. In this case, we are indexing individual \(i\) at school \(s\) in time \(t\).

The OLS estimate is a means comparison between the graduation rates of people who take AP courses and people who don’t.
Q. — What types of people are we comparing?

A. — *OLS compares the smartest people at the schools in years when AP economics being offered to all other students.*

If we pretend we have just 2 schools, we can think of 4 groups.

<table>
<thead>
<tr>
<th>Smart, motivated kids who take AP</th>
<th>Unmotivated kids who don’t take AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group (A)</td>
<td>Group (B)</td>
</tr>
</tbody>
</table>

School 1 with AP offered

<table>
<thead>
<tr>
<th>Smart, motivated kids who don’t take AP</th>
<th>Unmotivated kids who don’t take AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group (C)</td>
<td>Group (D)</td>
</tr>
</tbody>
</table>

School 2 with no AP economics offered

OLS compares Group A to the average of Group B, C, and D.

It seems clear that unconditional OLS will be biased without other controls. Even if we can control for something like 10th grade test scores, I wouldn’t believe the estimates. Since we can’t easily control for motivation.

Q. — Could we include person fixed effects to control for motivation?

A. — *No, within each student there is no variation of whether he graduates college.*

Q. — Could I include school-by-year fixed effects to control for other factors about a school that influence both AP course offerings?

A. — *Yes, this is possible since there is still variation left over since within a school some students take AP economics and some don’t.*

**Within School Estimator**

Include school-by-year fixed effects:

\[
Graduate_{ist} = \gamma AP_{ist} + \omega_{ist} + \epsilon_{ist}
\]  

(29)
Notice the subscript on $\omega$ is *really* important. It designates which fixed effect we are including.

**Q.** — What is the identifying variation in this context?

**A.** — *The variation that will be used is caused by the difference between unmotivated kids and motivated kids within each school that has AP. Schools that don’t offer AP economics don’t at all contribute to the analysis at all.*

Comparing Group A to Group B seems even worse than comparing group A to B,C and D. *The bias here is probably even worse than from simple ols!*

**Between school-year Estimator**

In this context, the between estimator uses only variation across schools. To do this, we need to aggregate Y and D to the school level.

Calculate average graduation rates by school, and average fraction taking AP economics. We then estimate:

$$
\text{Graduate}_{st} = \gamma \text{AP}_{st} + \epsilon_{st}
$$

Notice that the subscripts here are s,t so that the identifying variation is at at the school-year level.

The type of identifying variation here is comparing group A+B vs C+D.

**Q.** — What is the problem with using the between variation here?

**A.** — *There are two distinct problems*

- Schools that offer AP economics may be better schools in many ways.
- Among schools that offer AP economics, if a larger fraction of students take AP economics this suggests that a larger fraction of students are motivated. Comparing across schools just aggregates the motivation problem here.

**Q.** — Which seems cleaner, the within variation or the between variation?

**A.** — *In this context, both types of variation seem pretty dirty, but probably within is more dirty.*

**9.6.1 Isolating the clean variation**

The discussion that follows, is much more general than the panel data context. You should usually be thinking about empirical analysis in this way.

Steps:
1. Think about a clean comparison.
2. Figure out how to aggregate or add controls to create a regression that performs that comparison.

The source of clean variation usually comes from changes. In this example, the source of cleaner variation will be the schools who introduce AP economics in 2005.

Q. —— What is a relatively clean comparison in this context?
Multiple Choice:

A. Compare students who take AP economics in 2005-2010 to students at the same school from 2000-2004.
B. Compare students enrolled in a school-year that offers AP economics to students enrolled in a school-year that doesn’t offer AP economics.
C. Compare students enrolled at a time when AP economics isn’t offered to students enrolled in the same school when AP economics is offered.

Comparison A

Comparison A is clearly dirty since the AP students in 2005-2010 are better than the typical student from 2000-2004.

Q. —— How can we implement comparison A?
If we wanted to run a regression to perform comparison A we could run OLS at the individual level with school fixed effects. Notice that as written, the comparison excludes non-AP students in 2005-2010. This exclusion makes sense, since these students are particularly different from AP students. To exclude these students, from the comparison, we simply include a dummy variable to flag these students.

\[ \text{Graduate}_{ist} = \gamma \text{AP}_{ist} + \delta \text{NoAP} \text{and AP offered} + \omega_s + \varepsilon_{ist} \] (31)

The means comparison we want is given by \( \gamma \), i.e. the difference between those who take AP and those who did not take AP (and didn’t have the option to).

Comparison B

Comparison B could be thought of as the between estimator, so we simply aggregate to the school level.

Another way to implement comparison B is to use instrumental variables. The instrument is whether AP economics was offered.
Reduced form:
\[ \text{Graduate}_{ist} = \delta \text{Offered}_{ist} + \epsilon_{ist} \] (32)

First stage:
\[ \text{AP}_{ist} = \gamma \text{Offered}_{ist} + \nu_{ist} \] (33)

The IV estimate is \( \rho = \frac{\delta}{\gamma} \). This approach is using pure between school-year variation.

Note that in this context, the ITT is the impact on overall college graduation rates by beginning to offer AP economics.

**Comparison C**

Comparison C seems relatively clean to me. So long as the types of students going to a school don’t change over time, students enrolled in a school in 2003 will be similar to students enrolled in a school in 2007.

Q. —— How can we implement comparison C?

**This comparison uses no within school-year variation, and uses no between school variation. We use within school, between school-year variation.**

We can implement this using IV with fixed effects.

Reduced form:
\[ \text{Graduate}_{ist} = \delta \text{Offered}_{ist} + \omega s + \epsilon_{ist} \] (34)

First stage:
\[ \text{AP}_{ist} = \gamma \text{Offered}_{ist} + \omega s + \nu_{ist} \] (35)

The IV estimate is \( \rho = \frac{\delta}{\gamma} \). This approach is using pure between school-year variation, but is controlling for school fixed effects.

**9.6.2 Addressing remaining biases**

When we implement comparison C, there are several potential sources of bias.

1. Schools make other changes in the year they begin to offer AP economics.
2. Schools begin to offer AP economics because of a change in students. e.g. a particularly motivated cohort.
3. A statewide improvement in college graduation rates exists from 2000 to 2010.

**State-wide changes** We can address the statewide improvement by including year fixed effects. Note that year fixed effects can only be added because some schools introduced AP economics and other schools did not. The year fixed effects take out the average college attendance for each cohort for the whole state.
School-by-year fixed effects could not be included since the IV specification would have no
variation within school-year.

**Changing students** We could control for observable student characteristics, to address
concern 2. This will only deal with observable changes though.

My preference is to simply do comparison C and then show suggestive tests regarding whether
concern 2 is a problem.

For example, if we measure student predetermined covariates \( X \), we could estimate:

\[
AP_{offered_{ist}} = X\beta + \omega s + \epsilon_{ist} \tag{36}
\]

If any of the \( \beta \) coefficients are big, this suggests that AP offering is a function of student
demographics – even controlling for a school fixed effect.

**Contemporaneous policy changes** There is really nothing to do to rule out other school
specific contemporaneous policy changes.

Ultimately, we will need to make the assumption that controlling for school, whether AP is
offered or not is random. This is essentially CIA assumption. In this case, it’s a bit more
plausible.

**9.7 What is the identifying variation? Is it clean?**

**9.7.1 Complex collinearity problems**

Suppose we are estimating equation (34) but including year fixed effects and school fixed
effects.

Q. —— Is the model identified under the following situations?

- Every school changed in 2005 from no AP to AP.
  
  The model will not run. There is no within year variation in whether a school offers
  AP and we are not using any of the within school-year variation in whether a student
takes AP.

- Some schools change in 2005 and some schools change in 2006, but all schools change
  in one of these 2 years.

  In 2005 there is variation and in 2006 there is variation in whether AP is offered. In
  all other years, there is no variation, so the estimates of the impact of AP are completely
  based on the year 2005 and 2006.

- Some schools change in 2005, and some schools never change AP status.

  We are using variation from all of post-2005 years (since there is cross-sectional
  variation in AP status in each of these years). The schools that never change AP status
  are used to identify the year fixed effects and the schools that change AP status
  contribute to the estimates of the impact of AP.
Sometimes it is really hard to figure out whether there is variation left over. In these cases, I suggest created a small toy dataset and exploring it with stata to see if there is identification. The problem with using a larger dataset is that mistakes in the data and unusual observations mean you rarely get true perfect colinearity. In your toy dataset you will be able to see whether there is perfect colinearity.

Example: Suppose we try to include firm and industry fixed effects in estimating the impact of female board members on profits. This model has perfect colinearity, but probably there is some firm that somehow switched industries during the time period which breaks the colinearity. As such, if you run the model in Stata on real data, you will not know that the industry fixed effects are collinear. Your industry estimates will be based entirely on the weird firm that switches industries. That is very bad)

9.8 Classical measurement error in fixed effects

Recall that if D has classical measurement error, the bias is given by:

$$\hat{\beta} = \beta \frac{\text{var}(X_T)}{\text{var}(X_T) + \text{var}(\nu)}$$

Where \(\text{var}(X_T)\) is the variance of the true value of the covariate. The severity of the measurement error is based on \(\text{var}(\nu)\).

If we implement fixed effects, we will partial out much of the true variation in \(X_T\), but the fixed effects will not be able to explain any of the variation in \(\nu\).

This means the measurement error will lead to much more severe attenuation.

Example: Impact of union on wages

Suppose almost no people actually change their union status, but there is a small amount of measurement error in union.

If we run a model with individual fixed effects, we will remove all the across person variation and be left with only within person variation. Most of this variation is actually measurement error in this example, so we will be looking at the impact of measurement error on wages.

The problem can also be seen if we think of doing first-differences instead of fixed effects.

\[
Y_{i1} = \beta D_{i1} + \nu_{i1} + c_i + \mu_{i1} \\
Y_{i2} = \beta D_{i2} + \nu_{i2} + c_i + \mu_{i2}
\]

When we do the first differencing, much of the true D will cancel out but \(\nu_{i1} - \nu_{i2}\) will have even larger variance than the regular measurement error.

9.9 Fixed effects vs Aggregation vs Random effects vs Pooled OLS

Fixed Effects uses only within variation.

Aggregation uses only between variation.
Pooled OLS uses a weighted average of the within and between variation.

Random effects uses a different weighted average of within and between variation.

Using random effects or pooled OLS is the same as assuming that both the between and the within variation are clean. Stated another way, both random effects and pooled OLS assume that $D_i \perp c_i|X_i$.

While OLS and random effects are somewhat similar, the weights are very different.

OLS uses all of the variation and treats within and between variation equally.

Random effects is a weighted average of within and between estimators and the weights are based on the inverse of the residual variation.

Whenever there is a lot of variation between people and not much variation within people, random and fixed effects will be about the same. This is because random effects will put the most weight where the residual variation is smallest.

When there is not a lot of variation between people and a lot of variation within people, random and fixed effects will still be about the same since there is not a lot of between variation, so there can’t be much bias. (they will both be similar to OLS). Essentially, we have mostly within variation in this context.

Fixed effects only gets you something when there is a large amount of correlated between variation AND there is an even larger amount of within variation. The between variation will create bias, and the large within variation ensures that the random effects puts more weight on the dirty between variation.

In general, the reason to do random effects is for efficiency and one rarely sees it done in microeconometrics. This is because rarely do you want to assume that $D_i \perp c_i|X_i$.

### 9.10 Clustering Standard Errors

In simple OLS, we usually use robust standard errors which allows for heteroskedasticity but doesn’t allow for serial correlation. This is a problem with panel data where serial correlation is inevitable.

The idea behind clustering is as follows:

Suppose we are looking at the impact of union on wages:

$$Wages_{it} = \beta \text{Union}_{it} + \epsilon_{it}$$

Since the subscripts are i and t, I know that I have a panel of N workers and T, time periods.

There are 2 ways to add more sample size:
1. Add new people

2. Add more years per person

Adding more sample, by (1) provides a lot of new information since we observe the relationship between union and wages for a new set of people.

Adding more sample by (2) doesn’t provide very much information since for most people, their union and wage status doesn’t vary much over time.

Simple OLS, doesn’t understand the difference between sample added with way 1 vs 2.

In the extreme, if you had 1000 total observations and they are all for the same person, you essentially only have 1 observation. OLS doesn’t know this though and it will treat all 1000 observations as new, independent information. OLS makes this mistake because when using robust standard errors we assume $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$

Clustering relaxes this assumption and allows for arbitrary correlation within each cluster.

If we cluster at the person level in the union regression, OLS will allow for serial correlation within each person. This means that when it observes a new year of data for a person already in the sample, it will not consider this truly new information if it is very similar to past observations for that person.

So long as the within cluster serial correlation is positive, clustering the standard errors will tend to increase the standard errors. The reason is that it essentially reduces your sample size. If there is negative serial correlation within clusters, clustered standard errors could be smaller than OLS, but in practice this is relatively rare.

Mathematically:

**clustered standard errors assumes:**

- no serial correlation across clusters, but allows for serial correlation within clusters.
- clustered standard errors are only consistently estimated as the number of clusters tends towards infinity ($k \to \infty$).

Clustered standard errors assumes $\varepsilon \varepsilon'$ is block diagonal. If the first cluster had 2 observations and the second cluster had 3 observations, it would take the form:

$$
\begin{bmatrix}
\sigma_1^2 & \sigma_{12} & 0 & 0 & 0 \\
\sigma_{21} & \sigma_2^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\
0 & 0 & \sigma_{43} & \sigma_4^2 & \sigma_{45} \\
0 & 0 & \sigma_{53} & \sigma_{54} & \sigma_5^2 \\
\end{bmatrix}
$$
9.10.1 Level of clustering

In order to cluster the standard errors, you must determine the level of clustering. To do this, you have to think about what types of new observations provide new information and what types of observations just reconfirm information you basically already knew.

The answer to the above question will depend on the level of identifying variation.

Cluster at the level of identifying variation

Example: AP courses

There are many levels at which we could cluster:

- student
- classroom
- year
- school
- school-year

When we estimate a model with school-year fixed effects, the identifying variation is comparing students in the same school-year to one another. This means the student is the level of identifying variation. Since our dataset assumed only one observation per student, clustering is unnecessary in this context.

When we estimate our preferred model with school fixed effects and a dummy variable for whether the school offers AP each year, the identifying variation is comparing the whole school in 2004 to the whole school in 2006. This means the level of identifying variation is the school-year level and this is the level at which we should cluster.

In our preferred model, if we add new students to a particular school and year, we get more information about that school and year, but we don’t get new identifying variation. When we add a whole new school, we get a lot more identifying variation since we can now compare across years in that school. When we add another year of data, we get more identifying variation.

Note that the clustering level is NOT the same as the fixed effects level.

Q. —— Why might we want to include fixed effects and cluster at the same time?

A. —— Suppose we will use variation across time within a school. We can include school fixed effects, but we still expect that there is serial correlation within a school because students are grouped into classrooms

107
10 Differences-in-differences

10.1 Basic Idea in Algebraic Form

Differences-in-differences is a an improved version of the before/after estimator.

Example:
In 1980 Castro allowed 125,000 immigrants to leave cuba. They all went to Miami and about 1/2 stayed there.

David Card studied the impact of this immigration influx on unemployment and wages of natives.

10.1.1 Before/After

Compare unemployment in 1979 to unemployment in 1981. If we are interested in longer run impacts, we could compare 1979 to 1985 or even later.

In the graph above, we see that unemployment increases between 1979 and 1981. This could be the result of immigration, but it could also be the impact of a national recession or any number of other things.

The DID design does a before and after design, but uses a control group. In his study, he compares unemployment changes in miami to unemployment changes in Atlanta, Los Angeles, Houston and Tampa, Florida.

The DID gets it’s name because it is the difference of 2 differences.

\[
\text{Difference 1} = Y_{Miami}^{1981} - Y_{Miami}^{1979}
\]

\[
\text{Difference 2} = Y_{Houston}^{1981} - Y_{Houston}^{1971}
\]

\[
\text{DID} = \text{Difference 2} - \text{Difference 1}
\]

In this case, the first difference is simply the change in unemployment between 1979 and 1981 in Miami.

The reason the DID works can be easily seen algebraically:
The unemployment rate in a particular city in a particular year is given by:

\[
U_{miami}^{1979} = \alpha_{miami} + \delta_{1979} \\
U_{miami}^{1981} = \alpha_{miami} + \delta_{1981} + \beta \\
U_{Houston}^{1979} = \alpha_{Houston} + \delta_{1979} \\
U_{Houston}^{1981} = \alpha_{Houston} + \delta_{1981}
\]

\(\beta\) is the impact of that the mariel boat lift had on the Miami unemployment rate. \(\alpha\) is a city fixed effect and \(\delta\) is the time fixed effect.

We have

\[
\text{Difference 1} = \delta_{1981} - \delta_{1979} + \beta \\
\text{Difference 2} = \delta_{1981} - \delta_{1979} \\
\text{DID} = \beta
\]

As the algebra above shows, the DID analysis relies on the time dummies (\(\delta\)) being the same for each city. In other words, we need to assume that the change in Houston from 1979 to 1981 is a reasonable counterfactual for the change that would have happened in Miami.

DID does not need to assume that Houston and Miami are comparable! The key point is that the Miami fixed effect and the Houston fixed effect cancel out in the first set of differencing, so they can be different.

Mathematically, we can write our assumption as:

\[\delta_{Miami}^{1981} - \delta_{Miami}^{1979} = \delta_{Houston}^{1981} - \delta_{Houston}^{1979}\]  \(37\)

Note that \(\delta_{Miami}^{1981}\) is not the same as \(\gamma_{Miami}^{1981}\). \(\delta_{Miami}^{1981}\) is what would have happened in Miami in 1981 if there was no immigration.

10.2 Graphs

We can see the DID analysis graphically. In fact, almost all DID papers are presented graphically.
In the graph above, Miami is the triangle and Houston is the circle (these numbers are made up). The key assumption is that the red dot is where Miami would have been, had there been no immigration. The blue dashed line on the far right is the difference in difference estimator.

In the graph above, I have shown only 4 points to emphasize the simplicity of the DID design. It is literally estimable with 4 data points. Unlike estimation of fixed effects, you do not need micro-level data here.

More traditionally, the graph is drawn as follows:

The DID estimate is the gap between the counterfactual 1981 Miami unemployment and the actual 1981 Miami unemployment.

Since the dashed counterfactual line is entirely based on what happened in Houston, the critical assumption here is that the change between Houston and Miami would have been the same in the absence of the immigration in Miami. This is called the common trend assumption and it is equivalent to the assumption written in equation 37.
10.3 Testing Common Trend Assumption

Unlike in OLS and IV, there is a fairly straight-forward and fairly convincing test of the key DID assumption.

Assuming we have more than 2 years of data, we can implement the DID analysis on years in which there was no policy change.

\[
\begin{align*}
U_{\text{miami}}^{1978} &= \alpha_{\text{miami}} + \delta_{1978} \\
U_{\text{miami}}^{1979} &= \alpha_{\text{miami}} + \delta_{1979} \\
U_{\text{Houston}}^{1978} &= \alpha_{\text{Houston}} + \delta_{1978} \\
U_{\text{Houston}}^{1979} &= \alpha_{\text{Houston}} + \delta_{1979}
\end{align*}
\]

Note that this DID is almost identical to what we did earlier, but there is no \( \beta \) term.

\[
\begin{align*}
\text{Difference 1} &= \delta_{1979} - \delta_{1978} \\
\text{Difference 2} &= \delta_{1979} - \delta_{1978} \\
\text{DID} &= 0
\end{align*}
\]

For the DID analysis, we need to assume that Miami would have followed Houston between 1979 and 1981 if it wasn’t for the immigrants. We cannot directly test this assumption, but we can test whether we see a zero effect where we expect to see a zero effect.

The most compelling evidence in favor is if Miami and Houston had the same trends prior to 1979.

10.3.1 Graphical test

Testing the common trend assumption is very simple when looking at a graph. We should see Houston and Miami move together until 1980 and then diverge. Note that Miami and Houston have very different unemployment rates the entire time, but based on the graph below, the DID assumption looks very good.
If the trends prior to the treatment year look very different, it doesn’t mean the DID assumption couldn’t be satisfied, but it makes it very unlikely that the assumption is satisfied.

Below is an example where the trends differ prior to the treatment year. I would not want to assume that the red dashed line is a what would have happened in Miami had immigration not occurred.

**Important Point:** The common trend assumption is not an assumption about the trends before the treatment year. The common trend assumption is only an assumption about what would have happened in the treated state if there had been no treatment. Looking for common trends pre-treatment is an indirect test of the common trends assumption.

### 10.4 Implementation through fixed effects

We could implement the simple 4 points difference in difference by manually computing the DID and bootstrapping to get a standard error.

A much simpler way, is to use a regression.

The DID regression is implemented by including time and city fixed effects. Suppose we only have 2 years (1979 and 1981) and we have 2 cities.

\[ Y_{it} = \alpha + \gamma \text{Miami}_{it} + \delta 1981 + \beta 1981 \times \text{Miami} + \epsilon \]

**Q.** What do \( \alpha \), \( \gamma \), \( \delta \), and \( \beta \) each correspond to?

**A.** When Miami, 1981 and 1981*Miami are all zero, we are at the observation for Houston in 1979. Using similar logic, we can get at all the parameters.

Note that based on the above equation we have:
\[ U_{\text{Houston}}^{1979} = \alpha \]
\[ U_{\text{Houston}}^{1981} = \alpha + \delta \]
\[ U_{\text{Miami}}^{1979} = \alpha + \gamma \]
\[ U_{\text{Miami}}^{1981} = \alpha + \gamma + \delta + \beta \]

A. — \( \alpha \) is the unemployment rate in Houston in 1979.

A. — \( \gamma \) is the difference between the unemployment rate in Miami and Houston in 1979.

A. — \( \delta \) is the difference between the unemployment rate in Houston in 1981 compared to 1979

A. — \( \beta \) is the DID

The DID estimator is simply the coefficient on the interaction term. The standard error on the interaction term is the standard error on the DID estimator.

The general regression equation to estimate DID is:

\[ Y_{it} = \alpha + \gamma \text{Location} + \delta \text{After} + \beta \text{After} \times \text{Location} + \varepsilon \]

10.5 Why DID fails

There are several reasons why DID can fail.

1. Common trend assumption fails.
   This is the most common concern and it can be tested for indirectly by looking at the pretrend.

2. Simultaneous treatment effect
   If something else happens at the same time in Miami, it will bias estimates. The pretrend test will not detect this problem.

3. Contamination
   If Houston is impacted by Miami’s labor market this will bias estimates (usually biases estimates towards zero). The pretrend test will not detect this problem.

4. Policy endogeneity
   If the policy was implemented in response to something, this will probably bias estimates. For example, if Chicago passes a new anti-violence law, you might be concerned that the laws passage was done because chicago anticipated that crime would rise a lot next year without the policy. If NYC doesn’t implement the same anti-violence law, this might mean that NYC has more reason to be optimistic. The pretrend test could detect this problem, but it could also miss it. If the policy is implemented based on trends in previous years, we will see that the pretrend test fails.
10.6 Multiple dimension DID (state-year fe)

Example: Suppose we have 2 schools. School 1 starts offering AP economics in 2006; the other school never offers AP economics.

This is a perfect instance to do a DID design.

The estimating equation would be:

\[ Y_{st} = \alpha + \gamma_{School1_{st}} + \delta_{After2005_{st}} + \beta_{After2005 \times School1_{st}} + \epsilon_{st} \]

Notice that the ”After x Location” interaction term is simply an indicator for treatment status.

Now, suppose we have hundreds of schools. Some start offering AP economics in various years, some never offer AP, some schools always offer AP.

Conceptually, the design is now more complicated, but the implementation is the same:

\[ Y_{st} = \alpha + \gamma_{s_{st}} + \delta_{t_{st}} + \beta_{OfferingAP_{st}} + \epsilon_{st} \]

In the above equation, we have school fixed effects, year fixed effects and the offer AP dummy.

This type of design is very commonly done with US states.

Suppose each state implemented an increased minimum wage in different years. We can estimate the impact of these laws on any outcome by using the extended DID framework:

\[ Y_{st} = \alpha + \gamma_{s_{st}} + \delta_{t_{st}} + \beta_{HigherMinWage_{st}} + \epsilon_{st} \]

Above, we have year fixed effects, state fixed effects and an indicator for whether a particular year and time has had a higher minimum wage. Importantly, the higherminwage is a 0 in years prior to the minimum wage increase and it is a 1 in years after the minimum wage increase. It is not simply a 1 in the year of the policy change and a zero in all other years.

Notice, that we could run this model, either with individual or state-year level data. As written above, we are running it with state-year level data. This means that if we have 10 years of data, we have 50*10=500 observations total. If we had individual level data we would estimate:

**DID with Individual level data**

\[ Y_{ist} = \alpha + \gamma_{s_{ist}} + \delta_{t_{ist}} + \beta_{HigherMinWage_{ist}} + \epsilon_{ist} \]

In the individual level equation, the identifying variation is still the same as before, but now we potentially have millions observations instead of 500.

The standard errors will get way smaller, but the coefficient estimates will be identical whether we aggregate or not.

While using individual data is fine, it is wrong to just use the robust standard errors from this regression.
Q. —— What level should we cluster the standard errors at?

A. —— According to an article by Bertrand and Mullainathan, you should cluster at the state level. Conceptually, it makes more sense to cluster at the state-year level since that is the level of identifying variation, but their article shows that state clustered errors perform better.)

10.7 DID extensions: multi-valued treatments; covariates

10.7.1 Including covariates

With individual data, or with multiple states, we can include covariates in the regression. If we have person level covariates (e.g. age, education), it won’t change the estimates, but it might reduce the standard errors.

If we have multiple states and we include state-by-time covariates (e.g. other policy change in state and time), this can impact our estimates.

The reason that only state-by-time controls will impact the estimates is because the identifying variation is at the state-year level.

10.7.2 Continuous treatments

If the treatment is multi-valued (e.g. level of minimum wage) it is not possible to show the DID graph since there is no clear control group. We can still implement the flavor of the DID analysis by estimating the same DID equation.

DID with multi-valued treatment

\[ Y_{ist} = \alpha + \gamma_s + \delta_t + \beta \text{MinWageLevel}_{st} + \varepsilon_{ist} \]

10.7.3 Common Trends Test with many treatment and control groups

With 50 states and many different policy changes, it is more difficult to create a graph to show the pre-trends assumption. Assuming the treatment variable is an indicator, We can test it mathematically as follows:

\[ Y_{ist} = \alpha + \gamma_s + \delta_t + \sum_{j=-m}^{n} \beta_j D_{st,j} + j + \varepsilon_{ist} \]  \hspace{1cm} (38)

\( D_{st,j} \) is an indicator for whether the treatment got switched on in year \( t \).

The above regression calculates the DID estimate and does the test all at the same time. The summation includes \( m \) state-specific dummies prior to the policy change and \( n \) state-specific dummies after the policy change.

The coefficients on the \( n \) dummies after the policy change show the impact of the policy (allowing for differential impacts 1 year after, 2 years after, etc..)

The coefficients on the \( m \) lagged dummies should all be zero if the pretrends are the same.
It is nice to show the coefficients (and their confidence intervals) plotted on a graph. A made up example that shows an ideal scenario is shown below.

![Graph showing coefficients and confidence intervals]

Notice that in the above, all the estimates prior to the policy change are indistinguishable from zero and we don’t see a consistent pattern. The 1 year impact can is about 1 percentage point and the long run impact is about 1.5 percentage points.

**10.8 DIDID**

DIDID is the same idea as DID, but you add another dimension of differencing.

In order to implement the DIDID, you need 2 different types of control group.

Ex:

Suppose welfare only is provided to poor mothers. Suppose that Illinois has an expansion of welfare in 2000 and Ohio does not have an expansion. We are interested in the impact on labor force participation.

We could do a DID analysis:

Illinois before and after 2000 VS Ohio before and after 2000.

The first Diff is state, the second Diff is time.

\[
U_{IL}^{1999} = \alpha_{IL} + \delta_{1999} \\
U_{IL}^{2001} = \alpha_{IL} + \delta_{2001} + \beta \\
U_{OH}^{1999} = \alpha_{OH} + \delta_{1999} \\
U_{OH}^{2001} = \alpha_{OH} + \delta_{2001}
\]

One example that would bias estimates is if something else happened in Illinois in 2000 that didn’t also happen in Ohio. For example, maybe in the year 2000, a major chicago company went bankrupt, depressing demand in Illinois.
If we calculate the DID estimate, it will conflate beta and bankruptcy.

To address this, we could do a DIDID.

\[ \text{Diff1=State} \]
\[ \text{Diff2=Time} \]
\[ \text{Diff3=Women with kids vs Women without kids} \]

A DIDID is the difference between 2 DID designs.

DID1: Do DID with State and Time just using women with kids.

DID2: Do DID with state and time just using women without kids.

If the original DID design was good, the DID2 estimate should be zero since people without kids aren’t impacted by the welfare law.

Only women with kids qualify for the welfare expansion, but women without kids should be impacted by “other” changes that occur in 2000 in Illinois.

The estimates are:

\[ \text{DID1} = \beta + \omega_{\text{Bankruptcy}} \]
\[ \text{DID2} = \omega_{\text{Bankruptcy}} \]
\[ \text{DIDID} = \beta \]

In the above, the key assumption is that women with kids and women without kids were equally impacted by the bankruptcy.

Q. —- What can still bias estimates here?

A. —- *Estimates will be biased if there is something else that impacted women with kids in Illinois 2000 that didn’t impact women without kids in Illinois in 2000*

Notice that if none of the following will create bias:

- Something else happens in 2000
- Women with kids are different than women without kids
- Something else happened in Illinois in 2000
- Something else happened to women without kids in 2000
10.8.1 DIDID Implementation

We run a regression just like the DID regression.

\[
Y_{ist} = \alpha_0 + \alpha_1 Illinois + \alpha_2 After2000 + \alpha_3 mother\varepsilon_{ist} \\
+ \alpha_4 mother \times Illinois + \alpha_5 mother \times After2000 + \alpha_6 After2000 \times Illinois \\
+ \beta mother \times After2000 \times Illinois + \varepsilon_{ist}
\]

As was done with the DID, we can work through exactly why the \( \beta \) corresponds to the DIDID.

We can calculate DID1 and DID2 to get the following:

**DID1:**

\[
Y_{OH, no kids}^{2001} - Y_{OH, no kids}^{1999} = \alpha_2 \\
Y_{OH, mother}^{2001} - Y_{OH, mother}^{1999} = \alpha_2 + \alpha_5 \\
DID1 = \alpha_5
\]

**DID2:**

\[
Y_{IL, no kids}^{2001} - Y_{IL, no kids}^{1999} = \alpha_6 + \alpha_2 \\
Y_{IL, mother}^{2001} - Y_{IL, mother}^{1999} = \alpha_6 + \alpha_2 + \alpha_5 + \beta \\
DID2 = \alpha_5 + \beta
\]

The simple DID could be biased if \( \alpha_5 \) is non-zero. This would correspond to a situation where something happened to mothers in 2000 that didn’t also affect non-mothers. This would bias estimates of the DID since it would be a contemporaneous shock to the treatment group.

The DIDID estimate solves this problem. When we take DID2 - DID1 we cancel out the mother specific shocks, so long as the factors affecting mothers in Ohio are the same as the
factors affecting mothers in IL. Importantly, even if mothers in Ohio are very different than mothers in IL so long as the changes experienced between 1999 and 2001 are similar, the DIDID estimates will be unbiased.

10.8.2 DIDID example: Many Tennessee residents without children lost public health care

Causal question: How did losing health insurance impact employment? (The idea is that people will start working in order to qualify for health insurance).

Do the before trends here seem parallel to you? The trends are not exactly parallel, but not too far off.

Looking at the above example, we can graphically see exactly what the DID estimate is.

In the above graph, the yellow line is the DID estimate. This comes from assuming that the change for Tennessee would have been basically flat without the program change (since the other southern states were basically flat).

In this example, the pretrends are somewhat similar, but it appears that Tennessee was on a slightly downward trajectory compared to the other states. Based on this, it seems like the DID
estimate will actually understate the impact of the policy change. DID assumes that tennessee would have followed the other states trend between 2005-2007, but my guess is tennessee would have actually been lower (red dot). BUT, that is just a guess. I would never want to base any estimates on that guess because it is based entirely on extrapolation of the Tennessee pretrend. It doesn’t incorporate the fact that other stuff may have happened in the post period nationally.

In this example, he adds another level of differencing by noting that people with children weren’t impacted by the health care law change.

The assumption in the DIDID is that the difference in trend between southern and non-southern states is the same for those with no children as it is for those with children.

**Q. —— Does it look like this assumption holds here?**

**A. —— It looks like the pre-trend for those with no children in Tennessee looks similar to the pretrend for those with children in Tennessee. This suggests it would be fine to do a simple DID focused only on before vs after for those with children. The DIDID assumption doesn’t seem to hold perfectly, but it isn’t far off. Doing the DIDID is useful because it rules out contemporaneous shock stories that wouldn’t be detected by the common pretrend test. We can see in the figure exactly what the 2 DID estimates would be and the DIDID estimate as well.

The key point is that we need to get a counterfactual for what would have happened in Tennessee.

**10.9 DID is IV**

Just as RD was a type of IV, in many cases DID is a type of IV as well.

The reason DID is an IV estimator here is that not everyone who lived in a southern state and didn’t have children was impacted by the law.

The gaps we saw above were the reduced form estimates. The true impact of losing your
health insurance must be even bigger since the aggregate effect is driven by a subset of the population.

The reduced form:

\[
Y_{ist} = \alpha_0 + \alpha_1 \text{Tennessee} + \alpha_2 \text{After2005} + \alpha_3 \text{haskids} \epsilon_{ist} \\
+ \alpha_4 \text{haskids} \times \text{Tennessee} + \alpha_5 \text{haskids} \times \text{After2005} + \alpha_6 \text{haskids} \times \text{Tennessee} \\
+ \delta \text{haskids} \times \text{After2005} \times \text{Tennessee} + \epsilon_{ist}
\]

The First Stage

\[ \text{Haspublichealthinsurance} = \alpha_0 + \alpha_1 \text{Tennessee} + \alpha_2 \text{After2005} + \alpha_3 \text{haskids} \epsilon_{ist} \\
+ \alpha_4 \text{haskids} \times \text{Tennessee} + \alpha_5 \text{haskids} \times \text{After2005} + \alpha_6 \text{haskids} \times \text{Tennessee} \\
+ \gamma \text{haskids} \times \text{After2005} \times \text{Tennessee} + \epsilon_{ist} \]

We can see the first stage in a graphs. Just as for the reduced form, we can estimate things either as a DID or a DIDID.

First Stage DID

![Figure 2. Share Publicly Insured](image)

Note: This figure presents the share of CPS respondents ages 21–64 without an advanced degree who report being covered by public health insurance in Tennessee versus other Southern states. The figure presents means by two-year cells. See text for details.

First Stage DIDID

![Figure 3. Share Publicly Insured, Triple Difference](image)

Note: This figure presents the share of CPS respondents ages 21–64 without an advanced degree who report being covered by public health insurance in Tennessee versus other Southern states. The figure presents means by two-year cells. See text for details.
Just as in IV, the estimator is \( \text{reduced form} \) \( \text{first stage} \)

**Q.** What exactly is the instrument here?

**A.** *Tennessee X after 2005 X has kids*

To estimate it in Stata, you can run `ivregress` where you control for all of state, year dummies and the instrument for having health insurance using “Tennessee X after 2005 X has kids”.

The results shown below show the reduced form and the first stage in the first 2 columns.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Has Public Health Insurance</td>
<td>Employed</td>
<td>Employed and Working &lt;20 hours per week</td>
<td>Employed and Working 20-35 hours per week</td>
</tr>
<tr>
<td>Tennessee × Post</td>
<td>- 0.046</td>
<td>0.025</td>
<td>- 0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>2005</td>
<td>(0.004)</td>
<td>(0.004)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td></td>
<td>[0.000]</td>
<td>[0.000]</td>
<td>[0.396]</td>
<td>[0.643]</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.871</td>
<td>0.867</td>
<td>0.392</td>
<td>0.418</td>
</tr>
</tbody>
</table>

**A. Difference-in-Difference Estimates**

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Has Public Health Insurance</td>
<td>Employed</td>
<td>Employed and Working &lt;20 hours per week</td>
<td>Employed and Working 20-35 hours per week</td>
</tr>
<tr>
<td>Tennessee × Post</td>
<td>- 0.073</td>
<td>0.046</td>
<td>0.002</td>
<td>0.018</td>
</tr>
<tr>
<td>2005 × No Children</td>
<td>(0.006)</td>
<td>(0.010)</td>
<td>(0.006)</td>
<td>(0.005)</td>
</tr>
<tr>
<td></td>
<td>[0.000]</td>
<td>[0.000]</td>
<td>[0.757]</td>
<td>[0.002]</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.952</td>
<td>0.941</td>
<td>0.665</td>
<td>0.824</td>
</tr>
</tbody>
</table>

**B. Triple-Difference Estimates**

<table>
<thead>
<tr>
<th></th>
<th>Mean of dep. variable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>0.705</td>
</tr>
<tr>
<td></td>
<td>0.037</td>
</tr>
<tr>
<td></td>
<td>0.097</td>
</tr>
</tbody>
</table>

### 11 Missing data

#### 11.1 Types of missing data

- Data missing by construction (wages for unemployed people, age of death for people still alive, wages at 40 for young workers)
- Data missing institutionally (If data comes from 1 state, I don’t observe outcomes for people who leave the state)
- Data missing due to mistakes in reporting (people with negative ages, people with GPA’s outside of the possible range)
- Data missing due to attrition (Some people can’t be found for followup surveys (usually not random))
- Data missing due to non-response on survey (people may skip particular questions)
- Data missing at random
etc...

Q. —— How might data be missing at random?

A. —— *Typos in data entry, merging issues might generate random missings, random non-response, new questions not asked to everyone*

Non-response is rarely random since it’s probably correlated with preferences or ability.

Important to understand how missing data will influence estimates.

Types of missing data:

- Data missing at random
- Missing data systematically related to D
- Missing data systematically related to Y
- Missing data systematically related to D and Y

For each of the above, we run a simulation to understand how our estimates are influenced by missing data.

Results from simulation:

- Data missing at random
  - Will not bias estimates.
- Missing data systematically related to D
  - Will not bias estimates, but will yield estimate that differs from ATE or TOT. We get estimated effect for those with non-missing data.
- Missing data systematically related to Y
  - May or may not bias estimates. If the value of Y causes you to be missing, this will probably bias estimates so long as D has a causal impact on Y. This is because changing D causes both Y to change and probability of missing to change.
- Missing data systematically related to D and Y
  - Will bias estimates

Just as with omitted variables, it is important to be able to sign the direction of the bias caused by missing data.

**Example: PSID**

PSID interviews original sample members and their offspring every 2 years regarding income and many other household characteristics.
Research question: Does parent receipt of welfare impact own income measured at 30?

For simplicity, assume that we randomly assign which parents gets welfare. PSID is well suited to study this question, but there will be missing data.

Reasons for missing data on parent or own income.

- Income will be missing if you are not yet 30.
- Income are missing for those that are unemployed
- Some people don’t respond to income questions
- Some people cannot be found in later surveys.
  - Kids who move out of the house are more likely to be lost.
  - Kids who are no longer in contact with parents are more likely to be lost
  - Kids who move to a cave and meditate are likely lost.
  - Families who move a lot get lost

Q. —— Which of these missing data generate bias? What is the direction of the bias?

- Income are missing for those that are unemployed.
  - This creates bias that can be signed under reasonable assumptions.
  - Suppose welfare causes long-run income to rise and unemployment to fall. (i.e. welfare benefits long-run outcomes). This means that welfare causes some people who would be missing (and not contribute to estimates) to be present. Under the reasonable assumption that those who were unemployed will now have relatively low income, this makes the benefits of welfare biased down.
  - Suppose welfare causes long-run income to fall and unemployment to rise. This is just the opposite case and estimates will be biased up (towards zero).

- Income will be missing if you are not yet 30.
  - Might bias estimates if there are particular types of cohort effects. Suppose the later cohorts have higher incomes because of improving educational quality. This alone will not bias estimates, even though the missing is correlated with Y. The reason is that treatment cannot cause Y to change in such a way that causes it to be missing.
  - Do we get bias if each subsequent cohort has higher income than the cohort before it AND the later cohorts are more likely to have D=1 than the earlier cohorts? NO! even though it really seems like we would. In this case, missing is related to Y only through cohorts and missing is related to D only through cohorts. Since cohorts doesn’t impact the relationship between D and Y (constant effects over time), we do not get bias.
• Some people don’t respond to income questions
• Some people cannot be found in later surveys.
  – Kids who move out of the house are more likely to be lost.
  – Kids who are no longer in contact with parents are more likely to be lost
  – Kids who move to a cave and meditate are likely lost.
  – Families who move a lot might get lost

11.2 Addressing missing data

There are many different procedures to try to correct for missing data, and I am not that familiar with most of them.

I tend to just try to think carefully about the direction any missing data is likely to bias my estimates (if at all).

11.2.1 Sample selection bias

Example: This is a very common type of missing data. Suppose we randomly assign people to private high school or public high school.

We are interested in how private high schools impact high school dropout and high school test scores. Students that dropout don’t take the tests.

Suppose we find that private high school decreases dropout by 10 percentage points but doesn’t impact high school test scores.

The high school dropout estimate is unbiased because we had random assignment.

The high school grades estimates is likely biased due to sample selection because we only observe test scores for those that don’t dropout.

Initially, those at private and public high schools have the same potential outcomes since it was randomly assigned.

Presumably, dropouts are the lowest performing students at each type of school.

Reducing dropout means that there are more of the lower achieving students still around at the private high schools. This means the set of persisting students at private schools is unlikely to have the same potential outcomes as those at public schools.

Q. — What direction will the bias be?

A. — I think it is very clear that the sample selection will bias our estimate of the impact of private schools on test scores negatively. The fact that we found no impact on high school test scores probably suggests that private schools help test scores
11.3 Addressing missing data

Suppose 80% of public school students graduate and 90% of private school students graduate. One procedure would be to try to infer the outcomes of the extra 10% that are missing in the public schools.

11.3.1 Quantile regression at the median

It can be hard to guess the tests scores of the 10% of dropouts, but it seems reasonable to assume that they are in the bottom 10%.

For OLS, it matters a lot if they are very low test scores or just low test scores.

For quantile regression at the median, so long as they are in the bottom half, it doesn’t matter where they are exactly.

We can do median regression and impute all the dropouts as having zero test scores.

11.3.2 Lee bounds

This procedure is easy to implement and intuitive, but it gives you fairly large bounds.

Instead of guessing about what the test scores would have been of the dropouts, we can instead artificially drop 10% of the private school students that did not dropout.

Suppose, the dropouts at the public schools would have been the bottom 10% of the class. In that case, we should drop the bottom 10% of the private school students and then run the regression.

Suppose, the dropouts at the public schools would have been the top 10% of the class. In that case, we should drop the top 10% of the private school students and then run the regression.

These two extreme assumptions give us bounds on the possible effect. Either the 10% of dropouts are at the top of the class, the bottom of the class or somewhere in between.

It is possible to