MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 1

Describing Distributions: Expectation and Moments

1 Univariate distributions

- Dive in: X_i denotes a random variable; it has a probability distribution. The subscript i reminds us that we see this for a particular person, place, time, or thing. Statisticians go to great lengths to make the idea of a probability distribution precise, but we will cut to the chase:
  - The *distribution* of X_i is the relative frequency of the values it can assume in the population over which it's defined. The *population* of interest contains all of the possible units we might see. Populations can be concrete (line the U.S. population up randomly: X_i is the age of the ith person in line) or imagined (X_i is the outcome of the ith coin toss when a coin is tossed repeatedly).
  - We use samples to learn about populations. A *sample* includes a specific number of chosen (sampled) units, in which case X_i describes something about one of these units. Sample units are indexed by $i = 1, ..., n$, where $n$is the sample size.
  - Tricky point: X_i is random variable ... until it's not. If I tell you, say, that we observe $X_i = 6$, then it's no longer random (it's the number 6).

1.1 Expectation of a random variable

- discrete r.v.: $X_i \in \{x_1, ..., x_J\}$

$$E(X_i) = \sum_j x_j p(x_j)$$

where x_j is one of $j = 1, ..., J$ values that X_i can take on and $p(x_j)$ is the probability that $X_i = x_j$

  - Consider Bernoulli X_i

- continuous r.v.

$$E(X_i) = \int_{-\infty}^{\infty} t f(t) dt$$

where $f(t)$ is the probability density function (*pdf*) of X_i (not "PDF", yo)

  - Consider uniform X_i

Expectation is the population analog of a sample average.

- Learn the lingo
  - The expectation of a random variable is a *parameter* that describes its distribution. My people often label parameters in Greek, sometimes writing μ_X for $E(X_i)$. We also say: "μ_X is the population mean of X_i". When it's clear what you're talking about, ditch the subscript and just write μ.
  - Draw a sample of n *observations* of r.v. X_i; We *estimate* $E(X_i)$ using the sample mean:
$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i$$
  - When it's important to keep track of sample size, we write $\bar{X}_n$ (e.g., when using the law of large numbers)

Exercise Show that in a sample of size n, the sample mean of a discrete r.v. satisfies:
$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i = \sum_j x_j \hat{p}(x_j),$$
where $\hat{p}(x_j)$ is the sample proportion with $X_i = x_j$.

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_j \mathbb{I}\{j\} x_j$

$= \sum_j \frac{\mathbb{I}\{j\}}{n} x_j = \sum_j x_j \hat{p}(x_j)$

1.2 Moments

- The rth population moment of random variable X_i is defined as $E(X_i^r)$
- The rth central population moment of random variable X_i is $E[(X_i - E(X_i))^r]$
- The moments of X_i characterize its distribution
  - The mean, $E(X_i)$, is a *first moment*, sometimes said to be a measure of location
  - The variance, a *second moment*, measures the dispersion of X_i around the mean:
$$\sigma_X^2 = V(X_i) = E[(X_i - E(X_i))^2] = E[(X_i - \mu_X)^2]$$
- The sample variance, s_X^2, replaces expectations with sample averages:
$$s_X^2 = \frac{1}{n}\sum (X_i - \bar{X})^2$$
(sometimes we divide by $n-1$)
- The 3rd central moment measures *skewness*, the extent to which a distribution is asymmetric; the 4th central moment measures *kurtosis*, or the likelihood of tail events (usually in comparison with tail probabilities for a Normal distribution). We're mostly concerned with first and second moments.

1.3 Expectation and variance: rules and properties[1]

1. *Expectation of linear functions.* Let $Z_i = a + bX_i + cW_i$ for constants a, b, c and random variables X_i and W_i. Then

$$E(Z_i) = a + bE(X_i) + cE[W_i]$$

Law of the unconscious statistician

$$\mathrm{E}[g(X)] = \sum_x g(x) f_X(x)$$

The proof uses the law of the unconscious statistician, which tells us how to evaluate the expectation of a function of a random variable.

2. *Ways to write variance*

$$\sigma_X^2 \equiv E[(X_i - E(X_i))^2] = E(X_i^2) - E(X_i)^2$$

Proof:

$$\sigma_X^2 = E[(X_i - \mu_X)^2] = E[X_i^2 + \mu_X^2 - 2X_i\mu_X] = \ldots$$

Can also write variance as $E[X_i(X_i - \mu_X)]$

$E[X_i^2] + (E[X])^2 - 2(E[X])^2$

$= E[X_i^2] - (E[X])^2$

(now use #1). How many ways to write variance? Three, (3), (iii)!

- Did we miss one?

3. *Mean-squared error (*MSE*).* Suppose you'd like to predict the realization of random variable X_i. You get one chance: your prediction is a constant. The MSE of X_i around any constant c is the expectation of squared mistakes:

$$\begin{aligned} MSE_X(c) &= E(X_i - c)^2 = \sigma_X^2 + (c - \mu_X)^2 \\ &= variance + bias^2 \end{aligned}$$

Proof:

$$(X_i - c)^2 = [(X_i - \mu_X) + (\mu_X - c)]^2 = (X_i - \mu_X)^2 + (\mu_X - c)^2 + 2(X_i - \mu_X)(\mu_X - c)$$

Now take expectations and use #1.

4. *Setting $c = \mu_X$ minimizes $MSE_X(c)$.* Proof: use #3. We can think of μ_X as the minimum MSE (MMSE) predictor of X_i. Sounds like machine learning! Statistics has long been concerned with prediction. The subspeciality of machine learning focuses on this.

5. Properties 2-4 also hold in samples, e.g., $s_X^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2$

$E[(X_i - \mu_X)^2 + (\mu_X - c)^2 + 2(X_i - \mu_X)(\mu_X - c)] = variance + (\mu_X - c)^2 + 2(\mu_X - c)E[X_i - \mu_X]$

$= variance + bias^2$

[1] Expect daily use: brush your teeth with these.

2 Bivariate distributions (understanding relationships)

2.1 Joint moments

- An $r+s$ joint moment is $E(X_i^r Y_i^s)$
- An $r+s$ joint central moment is $E[(X_i - \mu_X)^r (Y_i - \mu_Y)^s]$

The big ones for us are:

Covariance $C(X_i, Y_i) = E[(X_i - \mu_X)(Y_i - \mu_Y)]$

(note that $r+s=2$, so this is a joint *second* moment)

Correlation $\rho_{XY} = \frac{C(X_i,Y_i)}{\sigma_X \sigma_Y} \in [-1, 1]$

Cov and corr measure the extent of linear relationship between X_i and Y_i (more below).

2.2 Conditional expectation

Conditional expectation is expected value of Y_i when X_i is fixed at a particular value.

- discrete r.v.: $Y_i \in \{y, ..., y_J\}$

$$E(Y_i|X_i = x) = \sum_j y_j f_2(y_j|X_i = x)$$

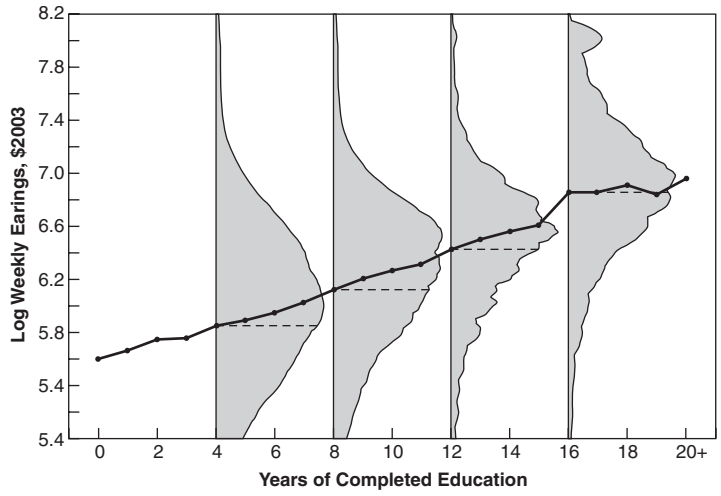
- continuous r.v.

$$E(Y_i|X_i = x) = \int_{-\infty}^{\infty} t f_2(t|X_i = x) dt$$

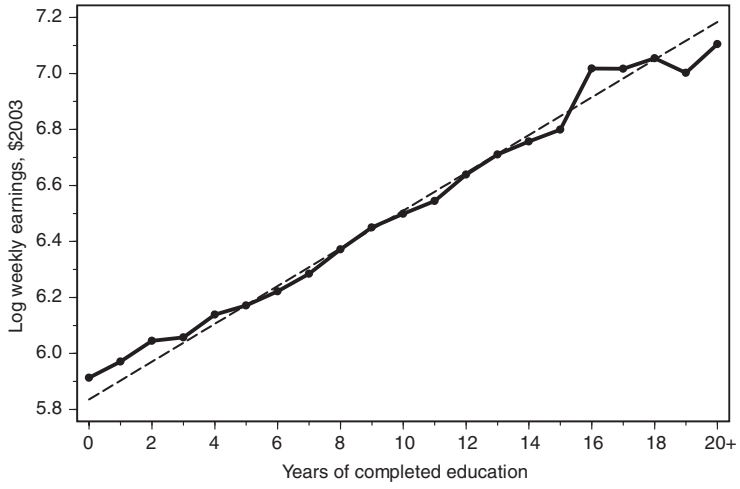
CEF The *Conditional Expectation* Function $E(Y_i|X_i)$ shows how mean Y_i varies as a function of X_i, without specifying which value of X_i we have in mind.

$E(Y_i|X_i)$ is a random variable and therefore has a distribution (unless and until we pick a particular value of X_i: $E(Y_i|X_i = 6)$ is a constant.

- Schooling and wages: conditional distributions



- Schooling and wages: CEF and the regression line that fits it



- Interesting fig ... what do these conditional means *mean*? An important question . . . one we'll visit and revisit in the weeks to come

2.3 Covariance and CEF: rules and properties[2]

1. *Ways to write covariance*

$$C(X_i, Y_i) = E[(X_i - \mu_X)(Y_i - \mu_Y)] = E(X_i Y_i) - \mu_X \mu_Y$$

This means that if $E(X_i)$ or $E(Y_i) = 0$ then $C(X_i, Y_i) = E(X_i Y_i)$. Also,

If μ_X or $\mu_Y = 0$ $\quad C(X_i, Y_i) = E[Y_i(X_i - E(X_i))] = E[X_i(Y_i - E(Y_i))]$

[2]Very important! *How* important, you ask? Very.

How many ways to write covariance? Three; 3; iii! Or maybe four.

- Lingo: Random variables that are uncorrelated, that is, $C(X_i, Y_i) = 0$, are said to be *orthogonal*

2. *Covariance of linear combinations of r.v.s.* Suppose

$$Z_{1i} = a_1 + b_1 X_i + c_1 Y_i$$
$$Z_{2i} = a_2 + b_2 X_i + c_2 Y_i$$

Then

$$C(Z_{1i}, Z_{2i}) = b_1 b_2 V(X_i) + c_1 c_2 V(Y_i) + C(X_i, Y_i)(b_1 c_2 + c_1 b_2)$$

3. *Variance of sums and differences*

$$V(X_i + Y_i) = V(X_i) + V(Y_i) + 2C(X_i, Y_i)$$
$$V(X_i - Y_i) = V(X_i) + V(Y_i) - 2C(X_i, Y_i)$$

Show this using #2, above or work out longhand.

- The variance of a sum of uncorrelated r.v.s is the sum of their variances

4. *Correlation measures the extent of linear relationship.* Suppose $Y_i = a + bX_i$ for some constants a and b, then $\rho_{XY} = 1$ when $b > 0$ and $\rho_{XY} = -1$ if $b < 0$. The intermediate case is:

$$Y_i = a + bX_i + e_i,$$

where $E[X_i e_i] = 0$. In general, $-1 \leq \rho_{XY} \leq 1$. Much more on this later.

5. *The Law of Iterated Expectations* (LIE). For any r.v.s, Z_i and X_i,

$$E(Z_i) = E[E(Z_i|X_i)]$$

In other words, "the marginal mean is the mean of the conditional means." Proof: See pages 31-32 in MHE and 14.32 Pset 1. Note: Z_i might be a function of other r.v.s, say X_i and Y_i.[3]

6. Properties of a CEF residual

i) $E[(Y_i - E(Y_i|X_i))X_i] = 0$

Think of $E(Y_i|X_i)$ as a predictor for Y_i using information on X_i (e.g., predict wages using schooling). Whatever's left over is uncorrelated with X_i. In fact, we can say something even stronger:

ii) $E[(Y_i - E(Y_i|X_i))g(X_i)] = 0,$

for *any* function, $g(X_i)$. Prove this using the LIE.

[3] Very very important. Make sure you can use the LIE. There's no expectation we won't iterate!

Proof of ①

$MSE = E[(y - m(x))^2] = E[(y - E[y|x] + E[y|x] - m(x))^2] = E[(y - E[y|x])^2 + (E[y|x] - m(x))^2 + 2(y - E[y|x])(E[y|x] - m(x))]$

$= E[(y - E[y|x])^2] + E[(E[y|x] - m(x))^2] + 2E[(y - E[y|x])(E[y|x] - m(x))]$ — 0 using #6

$= E[(y - E[y|x])^2] + E[(E[y|x] - m(x))^2]$

Independent of choice of $m(x)$ — Equals 0 when $m(x)$ is chosen to be $E[y|x]$

2.4 Bonus Properties

More to know 'bout the CEF

1. $E(Y_i|X_i)$ is the minimum MSE predictor of Y_i given X_i (Prove this using #6 above).

Proof of ANOVA

Expectation of conditional variance + Variance of conditional expectation

$V[Y] = E[Y^2] - E[Y]^2$

i) $E[Y^2] = E[E[Y|X]^2]$ using LIE

$= E[V[Y|X] + E[Y|X]^2]$ using variance definition

ii) $E[Y]^2 = E[E[Y|X]]^2$

$E[Y^2] - E[Y]^2$

$= E[V[Y|X] + E[Y|X]^2] - E[E[Y|X]]^2$

$= E[V[Y|X]] + E[E[Y|X]^2] - E[E[Y|X]]^2$

$= E[V[Y|X]] + V[E[Y|X]]$

2. *Analysis of variance (ANOVA)*

$$\sigma_Y^2 = E[\sigma_{Y|X}^2] + V[E(Y|X)] \quad (1)$$

where $\sigma_{Y|X}^2$ is defined as $E\{(Y_i - E[Y_i|X_i])^2|X_i\}$. Equation (1) is called the analysis of variance (ANOVA) formula. The ANOVA formula is interpreted as follows: $\sigma_{Y|X}^2$ is "within-X_i" variance; i.e. variance in Y_i given X_i, while $V[E(Y_i|X_i)]$ is "between-X_i" variance, i.e., the variance in the CEF of Y_i given X_i (note that because X_i is random, $\sigma_{Y|X}^2$ and $E(Y_i|X_i)$ are also random). Total variance is therefore the sum of within-group and between-group variance.

Bounding probabilities using Chebyshev's Inequality

Pafnuty L. Chebyshev (b. May 16, 1821, Zhukovsky District, Kaluga Oblast, Russia) showed that for any random variable, X_i, and any positive constant, c:

$$P(|X_i - \mu_X| \geq c\sigma_X) \leq 1/c^2.$$

In other words, the probability that X_i is more than c standard deviations from its mean is less than $1/c^2$.

- This inequality made Pafnuty popular with his neighbors because it allowed them to bound the probability of extreme events, mostly disasters of various kinds
- We don't use this awesome inequality every day, but it's good to know and a good exercise to show. We'll use it later to prove the *Law of Large Numbers*.

(Prove these bonus properties for extra credit on Pset 1)

Proof of Chebyshev's Inequality:

Let $Y = (X - \mu_X)^2$ $E[Y] = E[(X - \mu_X)^2] = \sigma_X^2$

$P(|X - \mu_X| \geq c\sigma_X) = P((X - \mu_X)^2 \geq c^2\sigma_X^2) = P(Y \geq c^2\sigma_X^2)$

Using Markov's Inequality:

$P(Y \geq c^2\sigma_X^2) \leq \frac{E[Y]}{c^2\sigma_X^2} = \frac{\sigma_X^2}{c^2\sigma_X^2} = \frac{1}{c^2}$

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 2

Sampling Distributions and Statistical Inference

1 Populations and samples

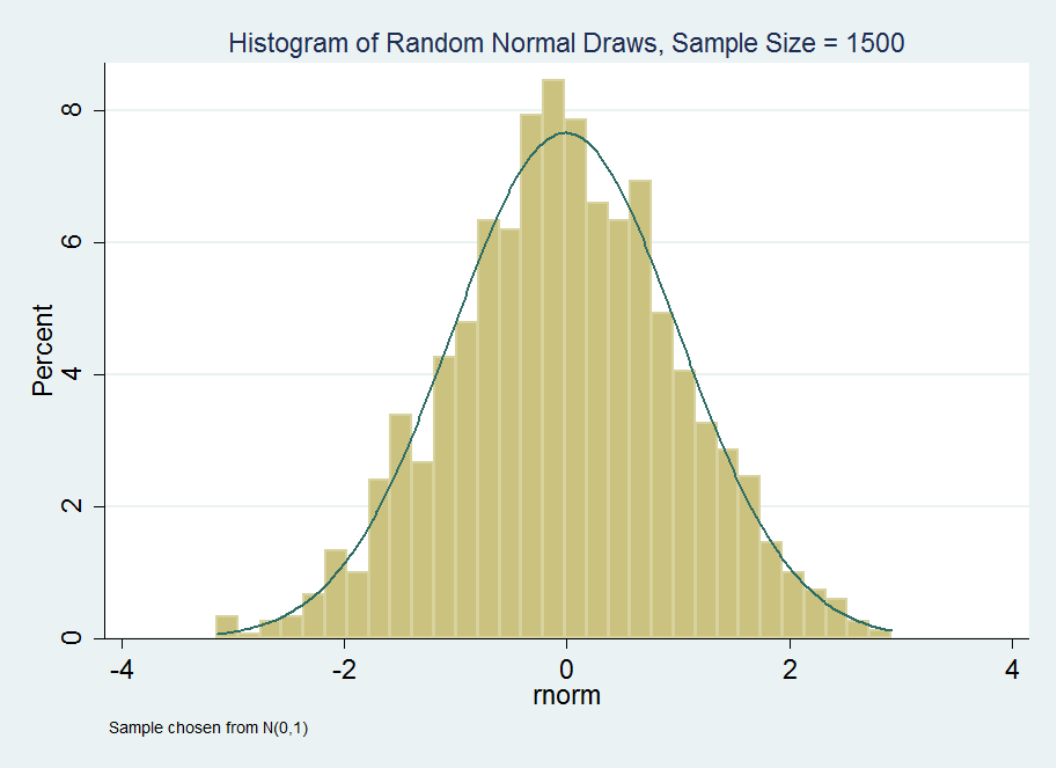
1.1 Ideas

We're interested in features of a population, like the population of economics majors (Are there really fewer women than men studying this awesome subject? We use statistical methods to evaluate the evidence)

- Moments and functions of moments – mean, variance, covariance, regression coefficients of various kinds – are the lens through which we see populations

We learn about population parameters by drawing samples. The most interesting samples are drawn from real data. But sometimes we sample from computer-generated theoretical distributions, just to see what the resulting samples look like.

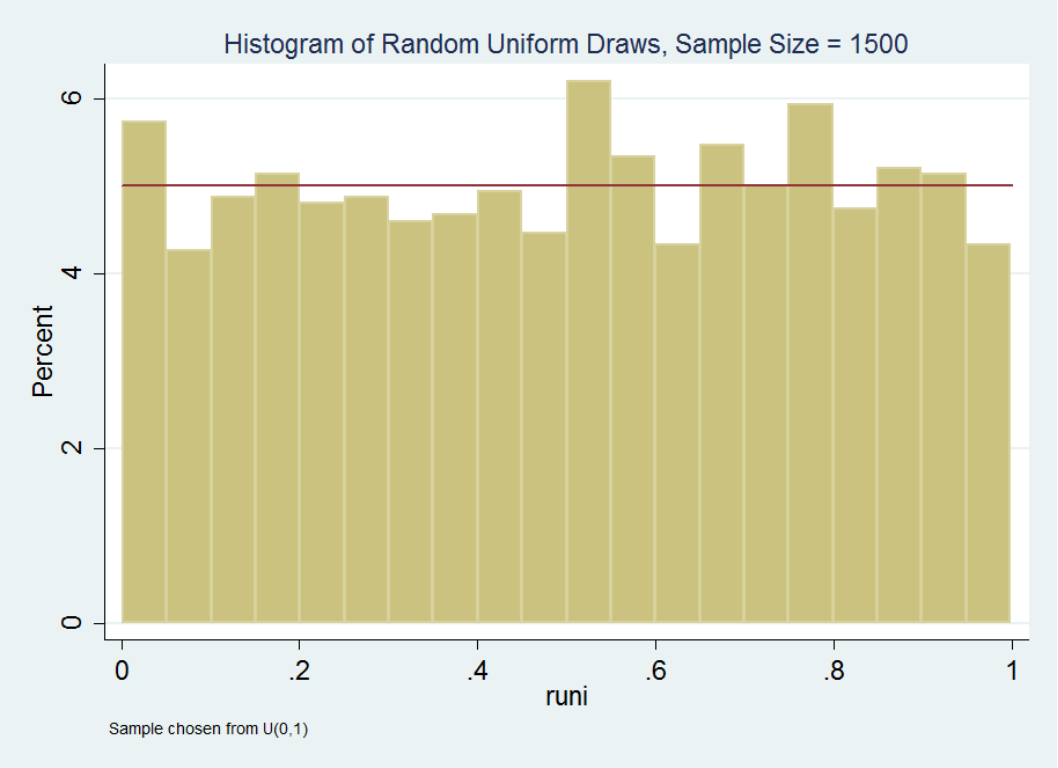
- Sampling from a Normal distribution (density and histogram below)



Q. What's this distribution a good model for?

- The Normal distribution is a *thing*: it's got a formula, it's *parametric*.
  - We write: $X_i \sim N(\mu, \sigma^2)$. What are the parameters of a Normal distribution?

- Sampling from a Uniform distribution defined over $[0, 1]$ (density and histogram below)

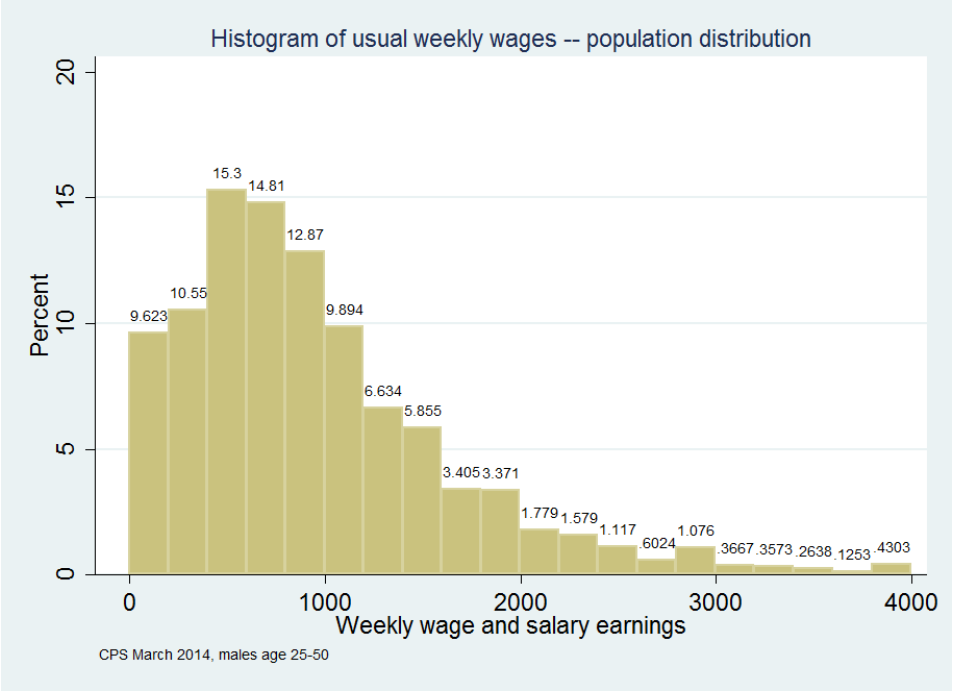


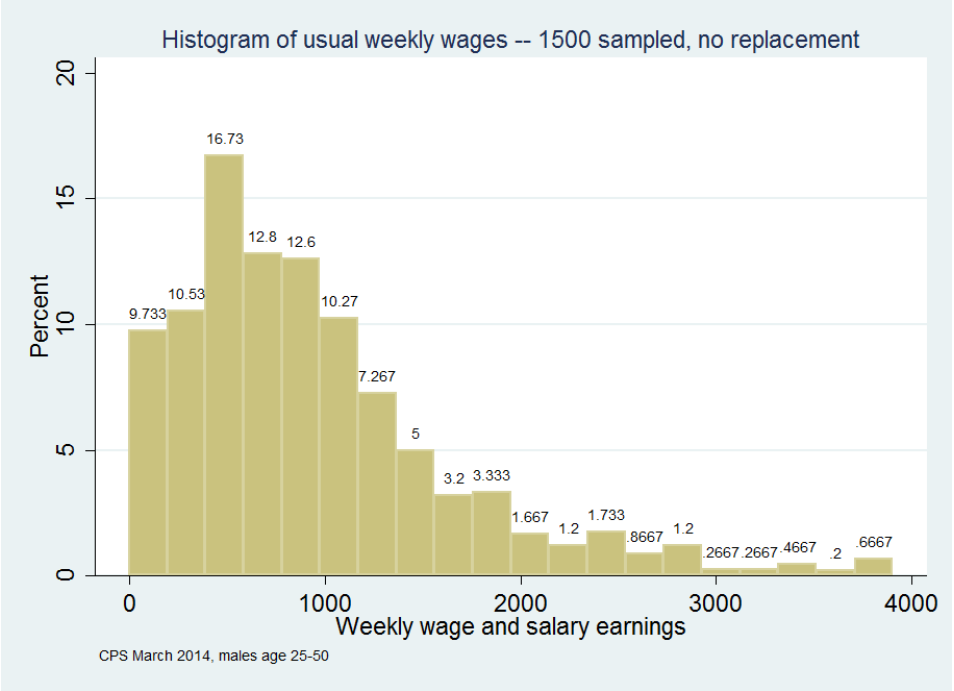
Q. What's a uniform distribution a good model for?
Q. Why isn't the histogram pictured here perfectly flat?

- A uniform distribution is parametric - what are its parameters?

- Other often-encountered parametric distributions: Dummy variables for things switched on or off (also called a Bernoulli distribution); Binomial and Poisson for count data; exponential and log-Normal for continuous non-negative r.v.s; multinomial for categorical data

- We'll often explore statistical properties by sampling real-world distributions. Our m.o. here is to treat a large sample as a notional "population," and then draw smaller samples from this. The larger sample defines a realistic *empirical distribution*.
  - Consider the Current Population Survey (CPS), a large sample of US households (roughly 60,000 interviewed every month). With CPS samples playing the role of "population," I've drawn a much smaller sample of 1500.
  - Here's population and sample data on the usual weekly wage, which is computed by asking workers in outgoing CPS rotation groups to report how much they usually earn for whatever time period seems most natural, and converting their reports to a weekly equivalent.

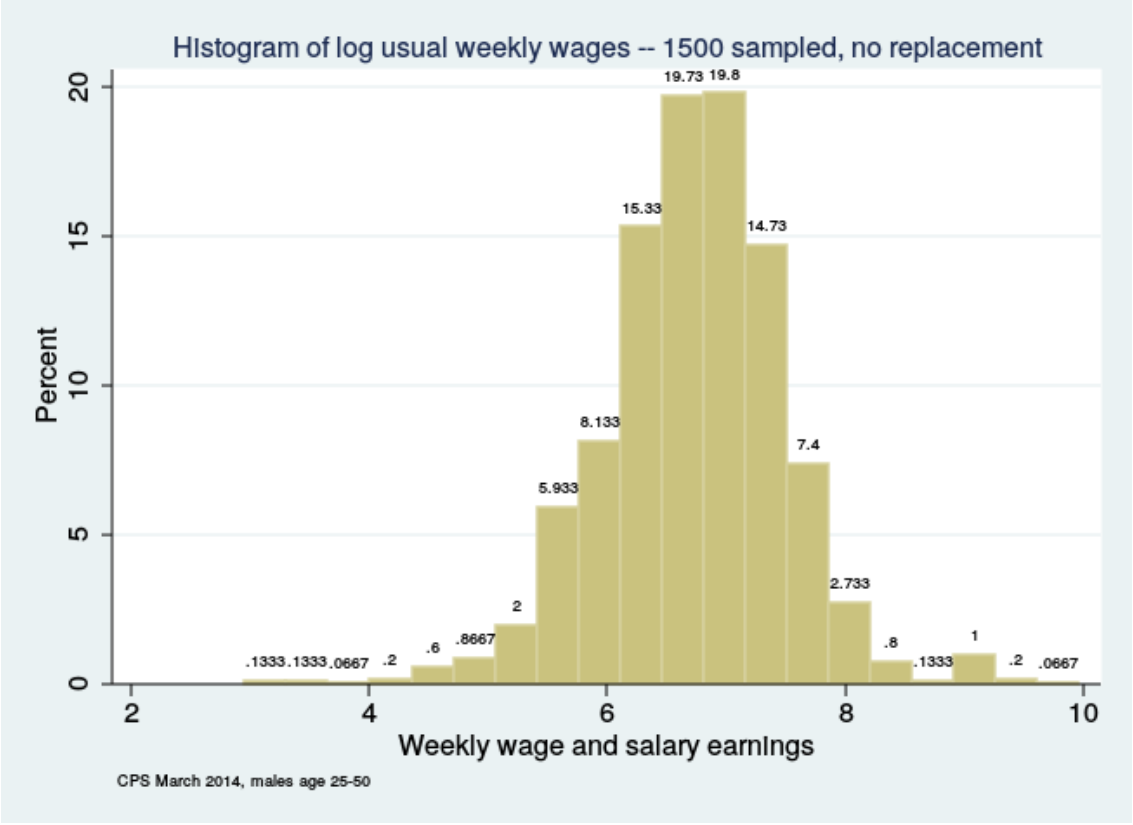




A few stats:

	N	Mean	s.d	Min	Max
UWW ("population")	53,455	939.81	690.05	0	3996.6
UWW (random sample)	1500	943.96	714.68	0	3904.4

- The CPS is used to track and study the US economy. The federal Bureau of Labor Statistics uses the CPS to compute the unemployment rate. Each month's CPS data set is a sample, but we can sample from the CPS as if it's a population (for teaching purposes, or to get a smaller and more manageable data set, or as part of a *sampling experiment*, in which we study sampling distributions of sample statistics and econometric estimators).
- We distinguish random sampling from the construction of an *extract*. Often, we're interested in a particular population group, say nonwhite female college graduates. To study this population, we'll take as many sample observations in this category as we can get. Selecting a particular subsample of interest yields an extract.
- BTW, wages love to be logged:



- We'll log more time with logs later. Q. Why are logs for this kind of data?

2 Sampling distributions

2.1 The Logic of Statistical Inference

1. Draw a sample from the population of interest
2. Use the this sample to support conclusions about the population from which the sample was drawn. These conclusions come in two forms:
   (a) Evaluate (test) hypotheses about population parameters (are men and women equally likely to major in economics?)
   (b) Measure (estimate) population parameters as best we can, while quantifying the statistical uncertainty inherent in these estimates (what's the gender gap in econ major rates?)
3. Execute (a) and (b) by quantifying *sampling variance*. Sampling variance summarizes the randomness in *sampling distributions*, that is, distributions of sample statistics obtained in repeated samples.

- Statistical inference differs from causal inference
- Statistically significant differences need not reflect causal effects (though sometimes they do)

2.1.1 Expectation and variance of the sample mean

Draw a random sample $\{X_i; i = 1, ..., n\}$. *Random sampling* means the observations are independent. For example, if we're sampling Course 14 majors and recording their gender (coded as a Bernoulli or dummy random variable), the fact that the first observation is female doesn't change the probability that the second is female.

Write the sample mean as

$$\bar{X} = \frac{1}{n}\sum_i X_i$$

Though we sample only once, the statistical behavior of $\bar{X}$ in repeated samples – it's *sampling distribution* – is easy to describe.

The mean and variance of the sample mean are:

$$E[\bar{X}] = \mu_X = E[X_i] \tag{1}$$

$$V(\bar{X}) = \frac{\sigma_X^2}{n} = \frac{V(X_i)}{n} \tag{2}$$

Prove it!

- Equation (1) says that the sample mean is an *unbiased* estimator of the population mean. Equation (2) says that variance of a sample mean declines with sample size at rate $\frac{1}{n}$.
- The ratio $\frac{\sigma_X}{\sqrt{n}}$ is called the *standard error of the sample mean* (Square root of sampling variance)
- The *standard error* of $\bar{X}$ is distinct from the *standard deviation* of the underlying X_i, though they're clearly related
  - SE ($\frac{\sigma_X^2}{n}$) measures the statistical *precision* of $\bar{X}$, and declines with n. SD (σ_X) measures the dispersion of X_i and is a feature of the distribution of X_i.
  - Q. Does σ_X change as a function of sample size? No.

Proof of (1)
$E[\bar{X}] = E\left[\frac{1}{n}\sum_i X_i\right]$
$= \frac{1}{n} E\left[\sum_i X_i\right] = \frac{1}{n}\sum_i E[X_i]$
$= \frac{1}{n} n E[X_i] = E[X_i]$

Proof of (2)
$V(\bar{X}) = V\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2} V\left(\sum_i X_i\right) = \frac{1}{n^2}\sum_i V(X_i) = \frac{1}{n^2} \cdot n V(X_i) = \frac{V(X_i)}{n}$

Chebyshev's Inequality

$$P(|X_i - \mu_X| \geq c\sigma_X) \leq 1/c^2.$$

2.1.2 The Law of Large Numbers

The LLN says that the probability that the sample mean is close to the corresponding population mean approaches one as the sample size increases. How close? As close as you like. The odds that the sample mean is very close to the population mean are high in large samples. How large? Hard to say, but often we think we have enough data for the LLN to be relevant.

Write $\bar{X}_n$ for the sample mean in a random sample of size n. Sample size is particularly important in this context, so we keep the n subscript. The LLN says

$$\lim_{n\to\infty} P\left\{\left|\bar{X}_n - \mu_X\right| \geq \epsilon\right\} = 0,$$

for any number, ϵ, no matter how small. Equivalently, we write:

$$\underset{n\to\infty}{plim}\, \bar{X}_n = \mu_X,$$

where *plim* denotes a probability limit. Casino owners and insurance companies rely on the LLN. Why is the LLN reliable? Because it's a theorem. Pset 1 asks you to prove *Chebyshev's inequality*, which generates the LLN as a consequence.

2.1.3 Sampling distributions: two ways

We know the expectation of $\bar{X}$ and its variance. We've also seen how precision grows as sample size increases. Still, we strive for more. We want to know the *sampling distribution* of $\bar{X}$ and not just it's mean and standard error.

We describe sampling distribution in two frameworks:

- Normal distribution theory. When the data are Normally distributed:
  - the sample mean is Normally distributed
  - the (appropriately scaled) sample variance has a chi-square distribution
  - the ratio of the centered sample mean to the estimated standard error of the mean has a t distribution
  - appropriately scaled ratios of sample variances from the same population have an F distribution
- Asymptotic distribution theory. When the data are distributed according to almost any distribution, then, in large samples:
  - sample moments are likely to be close to the corresponding population moments (the LLN)
  - the ratio of the centered sample mean to the estimated standard error of the mean is *approximately* distributed standard Normal t-stat is ~ normal
  - appropriately scaled ratios of sample variances have a chi-square distribution

Asymptotic distribution theory relies on the *central limit theorem* (CLT) as well as the LLN.

3 Normal Distribution Theory

The Normal distribution is a symmetric, bell-shaped distribution that offers a good a model for data that are, well, kinda normal. In practice, our data are often abnormal, but this matters less than you might think. Since many theoretical results are easily derived when the data are Normally distributed, Normal distribution theory provides a valuable reference point for the more general *asymptotic* (large-sample) framework that doesn't rely on having Normally distributed data.

3.1 The Normal distribution

- If X is Normal, we can write $X \sim N(\mu_X, \sigma_X^2)$

$$Z = (X - \mu_X)/\sigma_X \sim N(0,1) \qquad \text{Standard Normal r.v.}$$

$$P(Z \leq z) = \Phi(z) =_{-\infty} \int^z \phi(t)dt \qquad \text{Std Normal } \textit{cdf}$$

$$\text{where } \phi(t) = (1/\sqrt{2\pi})exp\{-t^2/2\} \qquad \text{Std Normal } \textit{density}$$

  $\Phi(z)$ is tabulated in books and by computers

3.2 Standardizing the sample mean

- Because linear combinations of Normal random variables are themselves Normally distributed, the transformed variable

$$Z = \frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \tag{3}$$

  has a *standard Normal distribution,* when the data are Normal. We write $Z \sim N(0,1)$

- Transforming any random variable by subtracting its mean and dividing by its standard deviation is said to *standardize* it

3.3 t-statistics

- Sample means standardized according to (3) are standard Normal. Alas, σ_X is unknown and must be estimated. Let Let $s_X^2 = \frac{\sum_i (X_i - \bar{X})^2}{n-1}$. (Note that s_X^2 is the unbiased sample variance estimator). We have

$$\begin{aligned} T_n &= [\bar{X} - \mu_X]/[\sigma_X/\sqrt{n}] \div (\sum_i (X_i - \bar{X})^2/\sigma_X^2(n-1))^{1/2} \\ &= \frac{\bar{X} - \mu_X}{s_X/\sqrt{n}} \\ &\sim t(n-1) \end{aligned}$$

  because it has a t distribution, we refer to T_n as a t-statistic. This result uses the fact that with Normal data, Z is standard Normal, $W = \sum_i (X_i - \bar{X})^2/\sigma_X^2 \sim \chi^2(n-1)$, and $Z/[W/n-1]^{\frac{1}{2}} \sim t(n-1)$ when Z is independent of W, as is so (if you don't believe me, look it up and see).

4 An Asymptopian Alternative: The CLT on Toast

4.1 Sampling under Normality

Suppose your data (X_i) are Normally distributed, so that the statistic $T_n \sim t(n-1)$ in a sample of size n. Then,

$$\lim_{n \to \infty} P(T_n \leq c) = \Phi(c),$$

where $\Phi(c)$ is the standard Normal cdf evaluated at c. When sampling from a Normal distribution, if the sample size is large enough, we can use standard Normal tables for t-tests [See entries in a t-table for $t(\infty)$].

4.2 Non-Normal data

The LLN applies whether or not the moments we're interested in were constructed using Normally distributed data. The LLN is reassuring (its the law!) but not enough for statistical inference. The *Central Limit Theorem* (CLT) is our ace in the hole.

Suppose that $\bar{X}_n$ and s_X are the sample mean and standard deviation from an i.i.d. sample of n observations on random variable X_i – not Normally distributed – with mean μ_X. Then it remains true that

$$\lim_{n\to\infty} P(T_n \leq c) = \Phi(c),$$

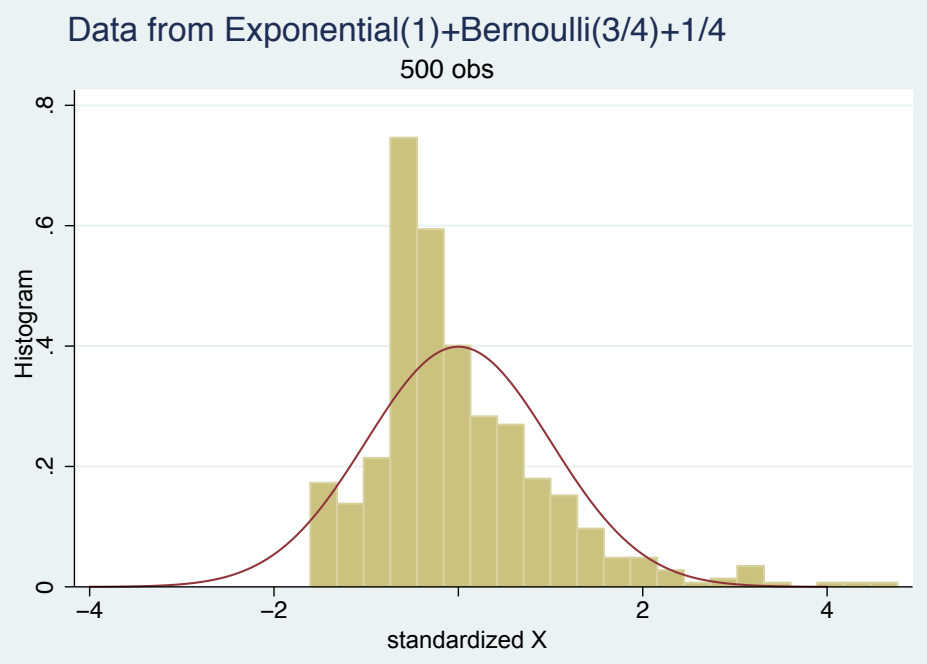
so the limiting distribution of the t-statistic is standard Normal even when the data are not Normally distributed.

This remarkable fact is called the *Central Limit Theorem*.

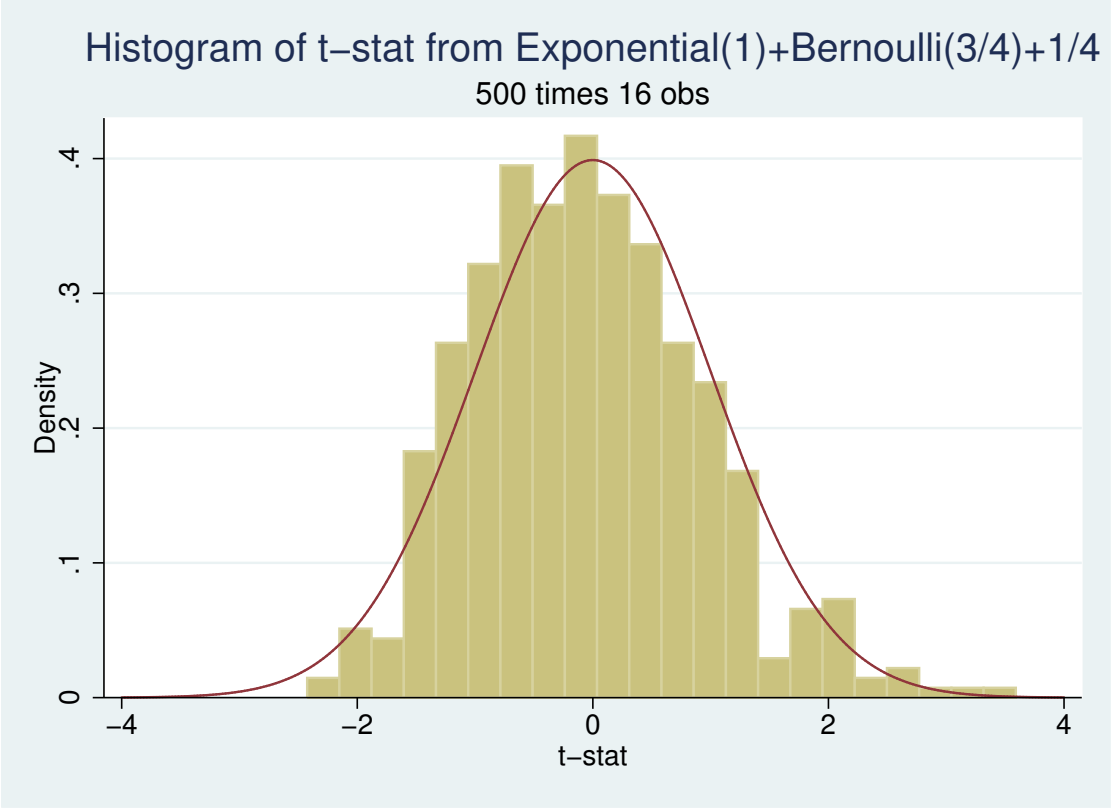
- Recap:
  - $E(\bar{X}_n) = \mu_X$ regardless of sample size or the distribution of the underlying data
  - The variance of $\bar{X}_n$ is $\frac{\sigma_X^2}{n}$ regardless of sample size or the distribution of the underlying data
  - The LLN promises that, in large samples, it's very likely that $\bar{X}_n$ is close to μ_X
  - The CLT says that if the sample is large enough, the distribution of T_n should be close to standard Normal, no matter the underlying data distribution
- With few exceptions, the CLT applies reliably to sample moments and functions of sample moments, including differences in means, regression coefficients, and other econometric estimators we'll meet later in term. As a rule, the CLT says

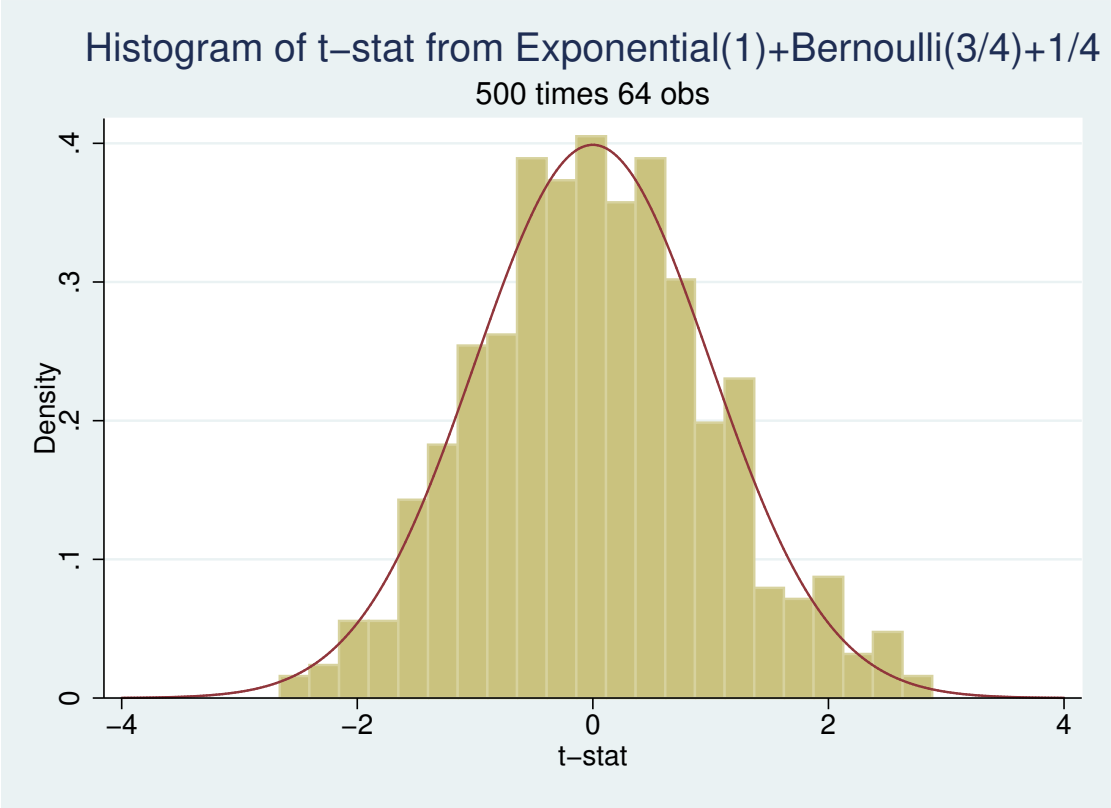
$$\frac{estimate - plim(estimate)}{SE(estimate)} \sim_a N(0,1)$$

  where $\sim_a$ means "approximately" or "asymptotically" distributed.
  - Masters of 'metrics mostly live and work in asymptopia
- Monte Carlo evidence for the CLT



– Sampling from an ab-Normal distribution





5 Testing with t

- We may want to test:

$$H_0: \quad \mu_X = \mu_0$$
$$H_1: \quad \mu \neq \mu_0$$

- If $\mu_X = \mu_0$, then

$$T_n = \frac{\bar{X} - \mu_0}{s_X/\sqrt{n}} \sim t(n-1)$$

If T_n is unusually large, that is, surprising given what we expect from a $t(n-1)$ distribution, we reject H_0. Of course, big t-stats just happen sometimes; we might wrongly reject.

False positive

- *Type I error*: reject a true null. The probability of Type I error is also called the *significance level or size* of a test. Sometimes this is denoted by α and we talk about an $\alpha - level$ test.

False negative

- *Type II error*: accept a false null. The probability of a Type II error is sometimes denoted β. *Statistical power* is $1 - \beta$. Low-powered tests reject rarely–even when the null is false (low power is the bane of an empiricist's existence)

Remember the two types of error by thinking about a criminal courtroom. In US jurisprudence, the null hypothesis is innocence, you're "innocent until proven guilty." Prosecutorial evidence weighs against innocence, perhaps causing the jury to reject the null and convict. But sometimes the evidence is misleading and juries convict wrongly. This is a Type I error. On the other hand, in the face of evidence of guilt that's too weak to banish reasonable doubt, juries fail to convict, even when the defendant is guilty. Letting a guilty man go free is a Type II error.

We face a trade-off between the the two types of errors, that is between significance and power. It's standard practice among empiricists to choose test size (often 5%) and hope for high power given this choice. Other testing approaches adjust the probability of the two types of error as a function of sample size.

In practice, we don't often follow orthodox testing theory, but rather use elements of the hypothesis testing machinery as suits the project at hand.

5.1 Comparing Means

Few standalone facts are of interest to us: *everything interesting is relative*. We're into comparisons in a big way, comparing treatment and control groups in randomized trials, for example. We'd like to know whether the random samples from two populations weigh against the null hypothesis that treatment and control populations are the same, in which case, we say that there's a *treatment effect*. Sometimes, our comparisons are merely descriptive. We might, for example, compare men and women to see if they differ (I'm told that they do, but we must check to be sure).

- Given observations on

1st sample: X_{1i} for $i = 1, ..., n_1$	$\bar{X}_1$ is the sample mean	μ_1 is the population mean
	s_1^2 is the sample variance	σ_1^2 is the population variance
2nd sample: X_{2i} for $i = 1, ..., n_2$	$\bar{X}_2$ is the sample mean	μ_2 is the population mean
	s_2^2 is the sample variance	σ_2^2 is the population variance

- We want to test:

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 \\ H_1: &\quad \mu_1 \neq \mu_2 \end{aligned}$$

  Assuming the data are Normal, we have:

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right])$$

  Where does the variance formula,

$$\left[\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right], \tag{4}$$

  come from? This is the formula for the *sampling variance of a difference in sample means*. The square root of this is the corresponding *standard error*.

- Under the null hypothesis, H_0:

$$(\bar{X}_1 - \bar{X}_2)/([\sigma_1^2/n_1] + [\sigma_2^2/n_2])^{1/2} \sim N(0, 1)$$

  Assuming also that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, [Do not do this if assuming quantities are different]

$$Z = (\bar{X}_1 - \bar{X}_2)/[\sigma(\frac{1}{n_1} + \frac{1}{n_2})^{1/2}] \sim N(0, 1)$$

  and

$$T_n = (\bar{X}_1 - \bar{X}_2)/[s(1/n_1 + 1/n_2)^{1/2}] \sim t(n_1 + n_2 - 2)$$

  where $s^2 = [\sum_i (X_{1i} - \bar{X}_1)^2 + \sum_j (X_{2j} - \bar{X}_2)^2]/(n_1 + n_2 - 2)$ is the pooled variance estimate.[1]

[1] This comes from $s^2 = \frac{1}{n_1+n_2-2}[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]$, where s_j^2 uses an $n_j - 1$ denominator.

- Even with Normal data, when pop variances differ, the distribution of T_n is only approximately $t(n_1 + n_2 - 2)$. But we still call it a $t-$*statistic*. In practice, variances are unknown, so we use the empirical t-statistic:

$$T_n = (\bar{X}_1 - \bar{X}_2)/([s_1^2/n_1] + [s_2^2/n_2])^{\frac{1}{2}}$$

This is the version with separate variance estimates, s_1^2 and s_2^2.

  - T_n can be compared with critical values for the $t(n_1 + n_2 - 2)$ distribution or a standard Normal. The distinction between standard Normal and t-distributions usually matters little.
  - Anyway, from a large-sample, asymptopian point of view, we expect T_n to be standard Normal under the null hypothesis regardless of whether the data are non-Normal and the variances are estimated or known.

- A little more lingo:

$$([s_1^2/n_1] + [s_2^2/n_2])^{\frac{1}{2}}$$

is said to be the *estimated standard error* of a difference in means, to distinguish it from the theoretical SE, the square root of expression (4).

5.2 Inference for experimental effects in randomized trials

- 'Metrics masters Angrist, Lang, and Oreopoulos (2009) were disturbed by rumors of lazy, low-performing college students. They decided to see whether students who won't boost effort for love (of learning), might do it for money
- ALO randomly allocated 1,656 full-time freshmen at a Toronto-area college to four groups, three treated and a control:
  - The SSP (250 students): treated with peer advising and the opportunity to attend facilitated study groups
  - The SFP (250 students): treated with the opportunity to win merit scholarships based on grades
  - The SFSP (150): treated with extra services and the opportunity to win merit scholarships
  - The controls consisted of everyone in the original 1,656 not randomly selected for a treatment
- Did cash incentives and/or enhanced services boost college achievement? For whom and by how much?

- Here's the award scheme

Appendix

Student Fellowship Program Award Schedule

	Award amount		
High school GPA quartile	$1,000	$2,500	$5,000
0–25th percentile	2.3 (C+)	2.7 (B−)	3.0 (B)
25th–50th percentile	2.7 (B−)	3.0 (B)	3.3 (B+)
50th–75th percentile	3.0 (B)	3.3 (B+)	3.7 (A−)

Notes: Eligibillity was determined by the student's best four courses. Half of SFP/SFSP participants were offered the opportunity to qualify for a $2,500 award.

- Descriptive statistics and check for balance

TABLE 1—DESCRIPTIVE STATISTICS

		Contrasts by treatment status				
	Control mean (1)	SSP v. control (2)	SFP v. control (3)	SFSP v. control (4)	*F*-stat (all=control) (5)	Obs. (6)
Administrative variables						
Courses enrolled as of fall 2005	4.745 {1.370}	−0.053 [0.095]	0.015 [0.095]	−0.158 [0.118]	0.702 (0.551)	1,656
No show	0.054	0.002 [0.016]	−0.030 [0.016]*	0.020 [0.019]	1.852 (0.136)	1,656
Completed survey	0.898	−0.018 [0.022]	−0.010 [0.022]	−0.051 [0.028]*	1.228 (0.298)	1,656
Student background variables						
Female	0.574	−0.006 [0.036]	0.029 [0.035]	−0.005 [0.045]	0.272 (0.845)	1,571
High school GPA	78.657 {4.220}	0.170 [0.308]	0.238 [0.304]	−0.018 [0.384]	0.276 (0.843)	1,571
Age	18.291 {0.616}	−0.054 [0.045]	−0.033 [0.044]	0.026 [0.056]	0.752 (0.521)	1,571
Mother tongue is English	0.700	0.017 [0.033]	0.009 [0.033]	0.049 [0.041]	0.495 (0.686)	1,571
Survey response variables						
Lives at home	0.811	−0.040 [0.030]	0.009 [0.030]	−0.004 [0.038]	0.685 (0.561)	1,431
At first choice school	0.243	0.024 [0.034]	0.060 [0.033]*	0.047 [0.042]	1.362 (0.253)	1,430
Plans to work while in school	0.777	0.031 [0.032]	−0.066 [0.031]**	0.037 [0.040]	2.541 (0.055)	1,431
Mother a high school graduate	0.868	0.015 [0.026]	−0.021 [0.026]	−0.045 [0.033]	1.040 (0.374)	1,431
Mother a college graduate	0.358	0.053 [0.037]	−0.020 [0.036]	−0.052 [0.046]	1.487 (0.216)	1,431
Father a high school graduate	0.839	0.025 [0.028]	0.008 [0.027]	−0.017 [0.035]	0.416 (0.741)	1,431
Father a college graduate	0.451	0.021 [0.038]	−0.001 [0.037]	−0.024 [0.048]	0.216 (0.885)	1,431
Rarely puts off studying for tests	0.208	0.031 [0.032]	0.031 [0.031]	0.107 [0.040]***	2.534 (0.055)	1,431
Never puts off studying for tests	0.056	−0.019 [0.016]	−0.016 [0.016]	−0.032 [0.021]	1.206 (0.306)	1,431
Wants more than a BA	0.556	0.052 [0.038]	−0.029 [0.037]	0.073 [0.048]	(1.752) (0.155)	1,431
Intends to finish in 4 years	0.821	−0.008 [0.030]	−0.006 [0.029]	−0.063 [0.037]*	(0.942) (0.419)	1,431

Notes: Standard deviations are shown in braces in column 1. Standard errors are reported in brackets in columns 2–4. *p*-values for *F*-tests are reported in parentheses in column 5. The last column shows the number of nonmissing observations.

• Did anyone notice?

TABLE 3—PROGRAM SIGN-UP AND USE OF SERVICES

	Signed up for STAR		Received SSP services		Met with/emailed an advisor		Attended FSGs	
	Basic controls (1)	All controls (2)	Basic controls (3)	All controls (4)	Basic controls (5)	All controls (6)	Basic controls (7)	All controls (8)
Panel A. All								
Offered SSP	0.519	0.549	0.238	0.255	0.204	0.217	0.106	0.118
	[0.032]***	[0.034]***	[0.028]***	[0.029]***	[0.026]***	[0.028]***	[0.020]***	[0.021]***
Offered SFP	0.863	0.867						
	[0.022]***	[0.022]***						
Offered SSP	0.762	0.792	0.412	0.431	0.383	0.397	0.131	0.139
and SFP	[0.036]***	[0.036]***	[0.041]***	[0.044]***	[0.041]***	[0.043]***	[0.029]***	[0.031]***
Observations	1,571	1,431	1,571	1,431	1,571	1,431	1,571	1,431
Panel B. Men								
Offered SSP	0.447	0.464	0.194	0.206	0.145	0.149	0.096	0.107
	[0.049]***	[0.052]***	[0.039]***	[0.042]***	[0.035]***	[0.038]***	[0.029]***	[0.032]***
Offered SFP	0.792	0.806						
	[0.040]***	[0.040]***						
Offered SSP	0.705	0.708	0.298	0.291	0.282	0.270	0.115	0.112
and SFP	[0.058]***	[0.065]***	[0.058]***	[0.063]***	[0.057]***	[0.061]***	[0.042]***	[0.046]**
Observations	665	594	665	594	665	594	665	594
Panel C. Women								
Offered SSP	0.571	0.605	0.273	0.287	0.251	0.264	0.113	0.124
	[0.043]***	[0.044]***	[0.038]***	[0.040]***	[0.037]***	[0.040]***	[0.027]***	[0.029]***
Offered SFP	0.912	0.908						
	[0.024]***	[0.026]***						
Offered SSP	0.800	0.835	0.506	0.532	0.466	0.489	0.146	0.155
and SFP	[0.046]***	[0.043]***	[0.056]***	[0.058]***	[0.056]***	[0.058]***	[0.040]***	[0.042]***
Observations	906	837	906	837	906	837	906	837

Notes: The table reports regression estimates of treatment effects on the dependent variables indicated in column headings. Robust standard errors are reported in brackets. The sample is limited to students registered for at least two courses as of November 1 with data on the relevant set of controls. "Basic controls" include sex, mother tongue, high school grade quartile, and number of credits enrolled. "All controls" includes basic controls plus responses to survey questions on procrastination and parents education.

* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

- Impact!

TABLE 5—TREATMENT EFFECTS ON FIRST YEAR OUTCOMES IN THE SAMPLE WITH FALL GRADES

	SFP by type			Any SFP		
	All (1)	Men (2)	Women (3)	All (4)	Men (5)	Women (6)
Panel A. Fall grade						
Control mean	64.225	65.935	62.958	64.225	65.935	62.958
	(11.902)	(11.340)	(12.160)	(11.902)	(11.340)	(12.160)
SSP	0.349	−0.027	0.737	0.344	−0.014	0.738
	[0.917]	[1.334]	[1.275]	[0.917]	[1.332]	[1.274]
SFP	1.824	0.331	2.602			
	[0.847]**	[1.233]	[1.176]**			
SFSP	2.702	−0.573	4.205			
	[1.124]**	[2.010]	[1.325]***			
SFP (any)				2.125	0.016	3.141
				[0.731]***	[1.164]	[0.972]***
Observations	1,255	526	729	1,255	526	729
Panel B. First year GPA						
Control mean	1.805	1.908	1.728	1.797	1.885	1.731
	(0.902)	(0.908)	(0.891)	(0.904)	(0.910)	(0.894)
SSP	0.073	0.011	0.116	0.071	0.008	0.116
	[0.066]	[0.107]	[0.082]	[0.066]	[0.107]	[0.082]
SFP	0.010	−0.110	0.086			
	[0.064]	[0.103]	[0.084]			
SFSP	0.210	0.084	0.267			
	[0.092]**	[0.162]	[0.117]**			
SFP (any)				0.079	−0.042	0.147
				[0.056]	[0.095]	[0.073]**
Observations	1,255	526	729	1,255	526	729

Notes: The table reports regression estimates of treatment effects on full grades and first-year GPA computed using the full set of controls. Robust standard errors are reported in brackets. The sample is limited to students registered for at least two courses as of November 1 with data on the relevant set of controls and at least one fall grade. The last three columns report estimates from a model that combines the SFP and SFSP treatment groups into "SFP (any)."

* Significant at the 10 percent level.
** Significant at the 5 percent level.
*** Significant at the 1 percent level.

- Well, maybe.
  - See Angrist, Oreopoulos, and Tyler, "When Opportunity Knocks, Who Answers? New Evidence on College Achievement Awards," *J. Human Resources* 49 (Summer 2014) for an update
  - As at the movies, research sequels often disappoint

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 3

Confidence Intervals

1 Confidence intervals for means

With Normally distributed data,

$$T_n = [\bar{X} - \mu_X]/[s_X/\sqrt{n}] \sim t(n-1),$$

where s_X is the sample standard deviation of X_i. And statistic T_n (called a "t-statistic") is also asymptotically standard Normal regardless of the distribution of X_i.

Let c_α denote the critical value for a two-sided α-level test. For example, when $\alpha = .05$, $c_\alpha \approx 2$. This critical value brackets the t-statistic with probability $1 - \alpha$:

$$P[-c_\alpha \leq T_n \leq c_\alpha] = 1 - \alpha$$

Rearranging, we have:

$$P[\bar{X} - c_\alpha(s_X/\sqrt{n}) \leq \mu_X \leq \bar{X} + c_\alpha(s_X/\sqrt{n})] = 1 - \alpha$$

- $s_X/\sqrt{n}$ is the *estimated standard error of the sample mean,* $\bar{X}$, in a sample of size n
- We say that

$$[\bar{X} - c_\alpha(s_X/\sqrt{n}),\ \bar{X} + c_\alpha(s_X/\sqrt{n})]$$

provides an "$(1-\alpha)$% confidence interval for μ_X".

- A 95% c.i. brackets μ_X 95% of the time. *Hey, what time is that?* When we draw repeated samples.
- But we don't draw repeated samples, we draw only one! Still, this probabilistic statement tells us how likely the interval is to *cover* the pop mean were we to conduct such a sampling experiment. Like standard errors, confidence intervals quantify *sampling variance.*
- Confidence intervals are random; Parameters (like μ_X) are fixed.
- In practice, we often ballpark the c.i.
  - Rule of thumb for a 95% confidence interval: sample mean plus/minus two standard errors (because if $\alpha = .05$, $c_\alpha = 1.96$ for Normal)

2 Confidence intervals for differences in means

- The t-statistic for a *difference* in means is

$$T_n = ([\bar{X}_1 - \bar{X}_2] - [\mu_1 - \mu_2])/([s_1^2/n_1] + [s_2^2/n_2])^{1/2} \sim t(n_1 + n_2 - 2),$$

where $T_n \sim N0, 1)$ for large n

- This comes from the fact that the *standard error of a difference in means* is

$$s^* = ([s_1^2/n_1] + [s_2^2/n_2])^{1/2}$$

so,

$$P[(\bar{X}_1 - \bar{X}_2) - c_\alpha s^* \leq \mu_1 - \mu_2 \leq (\bar{X}_1 - \bar{X}_2) + c_\alpha s^*] = 1 - \alpha,$$

where c_α is the $t(n_1 + n_2 - 2)$ critical value for a two-sided α-level test.

- The rule of thumb interval for a difference in means is $(\bar{X}_1 - \bar{X}_2) + / - 2s^*$.

3 Confidence intervals for treatment effects in ALO (2009)

TABLE 5—TREATMENT EFFECTS ON FIRST YEAR OUTCOMES IN THE SAMPLE WITH FALL GRADES

	SFP by type			Any SFP		
	All (1)	Men (2)	Women (3)	All (4)	Men (5)	Women (6)
Panel A. Fall grade						
Control mean	64.225	65.935	62.958	64.225	65.935	62.958
	(11.902)	(11.340)	(12.160)	(11.902)	(11.340)	(12.160)
SSP	0.349	−0.027	0.737	0.344	−0.014	0.738
	[0.917]	[1.334]	[1.275]	[0.917]	[1.332]	[1.274]
SFP	1.824	0.331	2.602			
	[0.847]**	[1.233]	[1.176]**			
SFSP	2.702	−0.573	4.205			
	[1.124]**	[2.010]	[1.325]***			
SFP (any)				2.125	0.016	3.141
				[0.731]***	[1.164]	[0.972]***
Observations	1,255	526	729	1,255	526	729
Panel B. First year GPA						
Control mean	1.805	1.908	1.728	1.797	1.885	1.731
	(0.902)	(0.908)	(0.891)	(0.904)	(0.910)	(0.894)
SSP	0.073	0.011	0.116	0.071	0.008	0.116
	[0.066]	[0.107]	[0.082]	[0.066]	[0.107]	[0.082]
SFP	0.010	−0.110	0.086			
	[0.064]	[0.103]	[0.084]			
SFSP	0.210	0.084	0.267			
	[0.092]**	[0.162]	[0.117]**			
SFP (any)				0.079	−0.042	0.147
				[0.056]	[0.095]	[0.073]**
Observations	1,255	526	729	1,255	526	729

Notes: The table reports regression estimates of treatment effects on full grades and first-year GPA computed using the full set of controls. Robust standard errors are reported in brackets. The sample is limited to students registered for at least two courses as of November 1 with data on the relevant set of controls and at least one fall grade. The last three columns report estimates from a model that combines the SFP and SFSP treatment groups into "SFP (any)."

* Significant at the 10 percent level.

** Significant at the 5 percent level.

*** Significant at the 1 percent level.

Ponder these questions, grasshopper:

- What determines c.i. width?
- Why are narrow intervals preferred to wide?
- Whats the connection between c.i. width and statistical power?

4 Monte Carlo c.i. coverage rates for immigrant-native wage diffs

(pretty good coverage, yo)

```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   16.1   Copyright 1985-2019 StataCorp LLC
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  401606257740
         Licensed to:  Josh Angrist
                       MIT Department of Economics

Notes:
      1.  Unicode is supported; see help unicode_advice.
      2.  Maximum number of variables is set to 5000; see help set_maxvar.
      3.  New update available; type -update all-

. do "/Users/joshangrist/Documents/teaching/14.32/SP2021/1432apps/LN3.do"

.
.
.         cap log close

.         set more off

.
.         log using LN3.log, replace
-------------------------------------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  /Users/joshangrist/Documents/teaching/14.32/SP2021/1432apps/LN3.log
  log type:  text
 opened on:  17 Feb 2021, 12:12:56

.
.         /*
>         14.32 Lecture Note 3
>         c.i. coverage for immigrant-native wage gap
>         */
.
.         // Insheet data from 2016 ACS PUMS
.         import delimited "ss16pusa.csv", clear
(284 vars, 1,623,216 obs)

.
.         // Sample of interest is men aged 40-49
.         keep if sex==1 & age>=40 & age <=49
(1,526,769 observations deleted)

.
.         // Restrict sample to men who worked 50-52 weeks
.         keep if wkw==1
(24,305 observations deleted)

.
.         // Generate usual hourly earnings
.         lab var wagp "Raw annual earnings"

.         lab var wkhp "Usual hours per week"

.         gen uhe = wagp/(50*wkhp)

.         quietly sum uhe, d
```

```
25 .         replace uhe=. if uhe>`r(p99)'
   (819 real changes made, 819 to missing)

26 .         lab var uhe "Usual hourly earnings (truncated at 99pct)"

27 .         gen loguhe=log(uhe)
   (4,963 missing values generated)

28 .         lab var loguhe "Log hourly earnings"

29 .
30 .         // Code up immigrants
31 .         gen immig=(nativity==2)

32 .         lab var immig "Foreign born"

33 .
34 . ttest loguhe, by(immig)

   Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
0	52,300	3.300842	.0031697	.724887	3.294629	3.307055
1	14,879	3.137078	.0064108	.7819809	3.124512	3.149644
combined	67,179	3.264571	.002859	.7410341	3.258967	3.270175
diff		.1637641	.0068562		.1503259	.1772022

```
   diff = mean(0) - mean(1)                                      t =  23.8855
Ho: diff = 0                     degrees of freedom =    67177

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000

35 .
36 . gen popmu_native=r(mu_1)

37 . gen popmu_immig=r(mu_2)

38 . gen popdiff=popmu_native-popmu_immig

39 .
40 .     // Keep only key variables in a newdata set called "workingextract"
41 .
42 .     keep wagp loguhe uhe wkhp immig age popmu_native popmu_immig popdiff

43 .         save workingextract, replace
   file workingextract.dta saved

44 .
45 . summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agep	72,142	44.6125	2.841859	40	49
wagp	72,142	77170.25	83305.79	0	714000
wkhp	72,142	44.70114	10.13632	1	99
uhe	71,323	31.93241	27.94947	0	201.5789
loguhe	67,179	3.264571	.7410341	-6.437752	5.306181
immig	72,142	.2256106	.4179867	0	1
popmu_native	72,142	3.300842	0	3.300842	3.300842
popmu_immig	72,142	3.137078	0	3.137078	3.137078

```
        popdiff |     72,142     .163764           0    .163764    .163764

46 . keep if _n==1
   (72,141 observations deleted)

47 . keep popdiff

48 . save expresults, replace
   file expresults.dta saved

49 .
50 . /* draw samples of N=100 500 times compute means, diff, SEdiff */
51 .
52 .    forvalues s=1/500 {
     2.         quietly use workingextract, clear
     3.         quietly bsample 100
     4.         quietly ttest loguhe, by(immig)
     5.         quietly gen mu_nat=r(mu_1)
     6.     quietly gen mu_imm=r(mu_2)
     7.         quietly gen diff=mu_nat-mu_imm
     8.         quietly gen numNat=r(N_1)
     9.         quietly gen numImm=r(N_2)
    10.         quietly gen SEdiff=r(se)
    11.         quietly gen hi95 = diff + (SEdiff*1.96)
    12.         quietly gen lo95 = diff - (SEdiff*1.96)
    13.         quietly gen cover95=0
    14.         quietly replace cover95 = 1 if lo95<=popdiff &  popdiff<=hi95
    15.         quietly keep if _n==1
    16.         quietly append using expresults
    17.         quietly save expresults, replace
    18.    }

53 .
54 .     keep popdiff mu_nat mu_imm diff numNat numImm SEdiff hi95 lo95 cover95

55 .     drop if missing(diff)
   (1 observation deleted)

56 .
57 .         /* sampling experiment results */
58 .
59 .         list popdiff diff numNat numImm SEdiff lo95 hi95 cover95 if _n<=5
```

	popdiff	diff	numNat	numImm	SEdiff	lo95	hi95	cover95
1.	.163764	.2083545	70	24	.1697101	-.1242773	.5409862	1
2.	.163764	.0158756	70	24	.2155199	-.4065435	.4382946	1
3.	.163764	-.0699334	74	21	.2135841	-.4885582	.3486914	1
4.	.163764	-.0496817	72	20	.1604921	-.3642462	.2648828	1
5.	.163764	.2745914	66	23	.1837659	-.0855898	.6347727	1

```
60 .         summarize diff lo95 hi95 cover95
```

Variable	Obs	Mean	Std. Dev.	Min	Max
diff	500	.1503984	.19436	-.3974841	.7256689
lo95	500	-.2158426	.2081041	-.8727461	.3615561
hi95	500	.5166393	.1942899	-.0447708	1.115047
cover95	500	.942	.2339775	0	1

```
61 .
62 .         log close
       name:  <unnamed>
        log:  /Users/joshangrist/Documents/teaching/14.32/SP2021/1432apps/LN3.log
```

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 4

Causality, Experiments, and Potential Outcomes

1 Casual vs Causal Effects

In an argument that's far from casual, Americans debate the causal effects of health insurance. Does health insurance affect health and/or health care costs? The view that insurance is beneficial on both counts motivated the 2010 Affordable Care Act, known also as Obamacare. No less important, bad health insurance is responsible for Walter White's metamorphasis from high school chemistry teacher to Heisenberg the Major Meth Dealer.

The Affordable Care Act imposed tax penalties on the uninsured, but it remains true that some Americans are covered and some aren't (The 2017 Tax Cuts and Jobs Act ended the individual mandate). This brings us to the question at the heart of MM Chapter 1:

> Are the insured healthier than *they* would have been had they not been insured?

Implicit in this question is a "what if" comparison. The answer is not obvious: after all, anyone can go to the emergency department in an hour of need (federal law requires the ED to treat all comers). This might be coverage enough.

The insured are indeed substantially healthier than the uninsured. But perhaps this just tells us something about the people who are lucky enough to have access to cheap health care coverage (like public sector workers) or rich enough to pay for it (like MIT faculty). The insured may differ from the uninsured for reasons besides their insurance.

Formal notation for *potential outcomes* makes causal questions precise. For each person, indexed by i, we define two possibilities:

- Health of person i when i is insured: Y_{1i}
- Health of person i when i is uninsured: Y_{0i}

The causal effect of insurance on person i is: The difference in a person i's health when insured vs when uninsured

$$Y_{1i} - Y_{0i}.$$

We never see this individual-level causal effect because, in any given data set and at any point in time, i is either insured or not. Still, we can hope to measure the *average treatment effect* (ATE) of insurance, an *average causal effect*:

$$E[Y_{1i} - Y_{0i}].$$

We might also consider the average causal effect of insurance on the insured:

$$E[Y_{1i} - Y_{0i}|D_i = 1], \quad \text{If insured had been uninsured, how would their health have differed.}$$

where D_i is a dummy variable equal to 1 for the insured. This parameter is called the *effect of treatment on the treated* (TOT). Parameter ATE tells us whether insurance benefits all in the population of interest, on average, while TOT tells us whether those in the insured population benefit (on average) from their coverage.

1.1 Selection bias

Research on causal effects often starts with TOT. This can be written

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

TOT compares the health of the insured, $E[Y_{1i}|D_i = 1]$, with *their* health when uninsured, $E[Y_{0i}|D_i = 1]$. Now, $E[Y_{1i}|D_i = 1]$ is easy to estimate in a random sample, but $E[Y_{0i}|D_i = 1]$ is *never* seen.

- $E[Y_{0i}|D_i = 1]$ is said to be *counterfactual*

You don't see the uninsured health of insured people

The challenge of measuring counterfactuals emerges in a comparison of health between insured and uninsured, which can be written:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0]. \tag{2}$$

Note that the Y_i's in (2) have lost 0 and 1 subscripts because we're referencing *observed outcomes* as opposed to *potential outcomes*. We're also ignoring the fact that in practice we make comparisons using sample means and not expectations; for the moment, the distinction between populations and samples is a detail.

Observed and potential outcomes are related. Specifically, we have

$$Y_i = Y_{0i}(1 - D_i) + Y_{1i}D_i \tag{3}$$

In other words, we see Y_{0i} for the uninsured and Y_{1i} for the insured:

$$\begin{aligned} &E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ &\quad = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0]. \end{aligned} \tag{4}$$

It stands to reason that the difference in average outcomes between insured and uninsured in equation (2) tells us something about the average causal effect we're after in equation (1). But not necessarily what we most want to know. Re-arranging equation (4), we get:

Difference in observed outcomes:

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\ &= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] + \{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]\} \\ &= \underbrace{E[Y_{1i} - Y_{0i}|D_i = 1]}_{TOT} + \underbrace{\{E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]}_{selection\ bias} \end{aligned}$$

The health w/o insurance of those insured − health w/o insurance of those uninsured

- The difference in average health between the insured and uninsured is the causal effect of insurance on the insured (TOT) plus the term in curly brackets.
  - This important term is called *selection bias*. It reflects the fact that the Y_{0i}'s of the insured may differed from those of the uninsured, on average.

1.2 Insured and otherwise in the NHIS

- Measuring health on a five point scale, the insured feel a lot better! Check out MM Table 1.1, constructed from the 2009 National Health Interview Survey
- The statistical significance of gaps in health by insurance status is not in doubt
- Statistical inference is easy: with estimates and standard errors in hand, you're good to go!
  - What do these differences in means mean?
  - Selection bias lurks in all such casual comparisons

TABLE 1.1
Health and demographic characteristics of insured and uninsured couples in the NHIS

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Notes: This table reports average characteristics for insured and uninsured married couples in the 2009 National Health Interview Survey (NHIS). Columns (1), (2), (4), and (5) show average characteristics of the group of individuals specified by the column heading. Columns (3) and (6) report the difference between the average characteristic for individuals with and without health insurance (HI). Standard deviations are in brackets; standard errors are reported in parentheses.

- Causal effect or selection bias?
- Panel B should worry those invested in causal claims: when it comes to comparisons by insurance status, *ceteris* is not *paribus*, and the differences here aren't subtle
  - Why is covariate imbalance important?
  - What does this imbalance suggest for the HI/no-HI contrast in mean Y_0, that is, for $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0]$? This quantity is ≥ 0 (Selection bias)

1.3 Random assignment eliminates selection bias

When insurance coverage is randomly assigned, as in a clinical trial or lottery, selection bias disappears. Suppose that D_i is determined by a coin toss: heads you're covered; tails you're not. By virtue of random assignment, the insured and uninsured in this experiment are similar in every way except their insurance status. Most importantly, when D_i is randomly assigned, the insured and uninsured have the same *potential* outcomes:

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

Hypothetical treated outcomes same regardless of actual treatment

Consequently,

$$\begin{aligned} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] & \\ = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] & \quad (5) \\ = E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 1] & \quad (6) \\ = E[Y_{1i} - Y_{0i}|D_i = 1] & \end{aligned}$$

The step from line (5) to line (6) explains the centrality of randomized trials in social science and clinical research: random assignment eliminates selection bias.

In a simple RCT, we also have:

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

In a randomized experiment where everyone does what they're assigned to do, the average causal effect on the treated, $E[Y_{1i} - Y_{0i}|D_i = 1]$, is the same as the population average causal effect, $E[Y_{1i} - Y_{0i}]$. This consequence of randomization is important, but it's not as important as the elimination of selection bias. (And more complicated experiments need not have the feature that ATE=TOT.)

1.4 Capturing causal effects without random assignment

Randomized research designs represent a sometimes-unattainable ideal. Masters of 'metrics therefore develop and implement empirical methods that reduce or eliminate selection bias in settings where random assignment is prohibitively expensive, time-consuming, impractical, or unethical. Even so, the experimental ideal disciplines our thinking. The first question to be answered is always thus:

What's the experiment you'd like to do?

2 A Healthy Debate

2.1 The RAND HIE

- Dateline 1974: Kung Fu enters its 3rd season; The RAND Health Insurance Experiment begins
  - In the 1970s, the RAND Corporation randomly assigned about 6,000 people (who agreed to drop their own insurance) to experimental insurance plans that required either no cost-sharing, a modest deductible, or imposed 25%, 50% or 95% coinsurance rates on subscribers, capped at a maximum annual payment of $1000.
- The next table shows RAND descriptive statistics and checks for balance
  - We look at four groups requiring different levels and types of cost sharing: (i) the catastrophic coverage plan approximates a no-insurance state; the (ii) deductible and (iii) coinsurance plans provided partial coverage; (iv) free care is what it sounds like.

TABLE 1.3
Demographic characteristics and baseline health in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
A. Demographic characteristics					
Female	.560	−.023	−.025	−.038	−.030
		(.016)	(.015)	(.015)	(.013)
Nonwhite	.172	−.019	−.027	−.028	−.025
		(.027)	(.025)	(.025)	(.022)
Age	32.4	.56	.97	.43	.64
	[12.9]	(.68)	(.65)	(.61)	(.54)
Education	12.1	−.16	−.06	−.26	−.17
	[2.9]	(.19)	(.19)	(.18)	(.16)
Family income	31,603	−2,104	970	−976	−654
	[18,148]	(1,384)	(1,389)	(1,345)	(1,181)
Hospitalized last year	.115	.004	−.002	.001	.001
		(.016)	(.015)	(.015)	(.013)
B. Baseline health variables					
General health index	70.9	−1.44	.21	−1.31	−.93
	[14.9]	(.95)	(.92)	(.87)	(.77)
Cholesterol (mg/dl)	207	−1.42	−1.93	−5.25	−3.19
	[40]	(2.99)	(2.76)	(2.70)	(2.29)
Systolic blood pressure (mm Hg)	122	2.32	.91	1.12	1.39
	[17]	(1.15)	(1.08)	(1.01)	(.90)
Mental health index	73.8	−.12	1.19	.89	.71
	[14.3]	(.82)	(.81)	(.77)	(.68)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table describes the demographic characteristics and baseline health of subjects in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors are reported in parentheses in columns (2)–(5); standard deviations are reported in brackets in column (1).

- Impact?

TABLE 1.4
Health expenditure and health outcomes in the RAND HIE

	Means	Differences between plan groups			
	Catastrophic plan (1)	Deductible – catastrophic (2)	Coinsurance – catastrophic (3)	Free – catastrophic (4)	Any insurance – catastrophic (5)
A. Health-care use					
Face-to-face visits	2.78 [5.50]	.19 (.25)	.48 (.24)	1.66 (.25)	.90 (.20)
Outpatient expenses	248 [488]	42 (21)	60 (21)	169 (20)	101 (17)
Hospital admissions	.099 [.379]	.016 (.011)	.002 (.011)	.029 (.010)	.017 (.009)
Inpatient expenses	388 [2,308]	72 (69)	93 (73)	116 (60)	97 (53)
Total expenses	636 [2,535]	114 (79)	152 (85)	285 (72)	198 (63)
B. Health outcomes					
General health index	68.5 [15.9]	−.87 (.96)	.61 (.90)	−.78 (.87)	−.36 (.77)
Cholesterol (mg/dl)	203 [42]	.69 (2.57)	−2.31 (2.47)	−1.83 (2.39)	−1.32 (2.08)
Systolic blood pressure (mm Hg)	122 [19]	1.17 (1.06)	−1.39 (.99)	−.52 (.93)	−.36 (.85)
Mental health index	75.5 [14.8]	.45 (.91)	1.07 (.87)	.43 (.83)	.64 (.75)
Number enrolled	759	881	1,022	1,295	3,198

Notes: This table reports means and treatment effects for health expenditure and health outcomes in the RAND Health Insurance Experiment (HIE). Column (1) shows the average for the group assigned catastrophic coverage. Columns (2)–(5) compare averages in the deductible, cost-sharing, free care, and any insurance groups with the average in column (1). Standard errors are reported in parentheses in columns (2)–(5); standard deviations are reported in brackets in column (1).

- Unlike Table 1.1, the comparisons in Table 1.4 carry causal weight. The question of their statistical significance is therefore similarly freighted
- We see little here to suggest insurance *causes* the insured to be healthier

2.2 HI on the Oregon Trail

- Elderly Americans get publicly provide health insurance through Medicare, while many of the poor are covered through Medicaid
  - In 2008, Oregon's Medicaid agency offered coverage to about 30,000 otherwise uninsured low-income adults who didn't qualify for Medicaid by the usual rules. These 30,000 were chosen by lottery from about 75,000 applicants.
  - This just in from Portlandia ...

TABLE 1.5
OHP effects on insurance coverage and health-care use

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Administrative data				
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)
Any hospital admissions	.067	.005 (.002)		
Any emergency department visit			.345	.017 (.006)
Number of emergency department visits			1.02	.101 (.029)
Sample size	74,922		24,646	
B. Survey data				
Outpatient visits (in the past 6 months)	1.91	.314 (.054)		
Any prescriptions?	.637	.025 (.008)		
Sample size	23,741			

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on insurance coverage and use of health care. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

- Hey, where's my health divided?

TABLE 1.6
OHP effects on health indicators and financial health

	Oregon		Portland area	
Outcome	Control mean (1)	Treatment effect (2)	Control mean (3)	Treatment effect (4)
A. Health indicators				
Health is good	.548	.039 (.008)		
Physical health index			45.5	.29 (.21)
Mental health index			44.4	.47 (.24)
Cholesterol			204	.53 (.69)
Systolic blood pressure (mm Hg)			119	−.13 (.30)
B. Financial health				
Medical expenditures >30% of income			.055	−.011 (.005)
Any medical debt?			.568	−.032 (.010)
Sample size	23,741		12,229	

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on health indicators and financial health. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

- Health insurance makes household *finances* healthier
- As in Table 1.4, the statistical significance of reduced health expenditures in Panel B carries causal weight: that's the miracle of random assignment
- Masters of 'metrics know to distinguish between:
  - random sampling, which facilitates statistical inference about populations using data from samples
  - random assignment, which supports causal inference, i.e., comparisons of potential outcomes free of selection bias)

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 5

Intro to Multivariate Regression

1 Matchmaker, Matchmaker

We're often interested in the dependence of one variable, Y_i, on another variable, X_{1i}, in a scenario where the connection between Y_i and X_{1i} reflects the fact that X_{1i} is correlated with another variable, X_{2i}, that also predicts Y_i.

- The association between health (Y_i) and health insurance (X_{1i}) in the NHIS might be explained by the higher schooling (X_{2i}) of the insured
- We use *multivariate regression* to control for such confounding factors
- Regression advances us on the path to *ceteris paribus* comparisons
- This mitigates and perhaps even eliminates selection bias

Regression mitigates selection bias by conditioning on a set of possibly confounding variables so as to hold them constant. To keep things simple, suppose that X_{1i} is Bernoulli. "Holding things constant" in this case means replacing the unconditional comparison,

$$E[Y_i \mid X_{1i} = 1] - E[Y_i \mid X_{1i} = 0],$$

with conditional comparisons,

Look at how Y differs when X_{1i} differs, with X_{2i} held constant

$$E[Y_i \mid X_{1i} = 1, X_{2i} = x] - E[Y \mid X_{1i} = 0, X_{2i} = x]. \tag{1}$$

In other words, we look at the CEF of Y given X_{1i}, *conditional on* $X_{2i} = x$. In Pset 1, for example, you're asked to compare the health of the insured and uninsured conditional on college graduation status.

- Conditional comparisons of this sort are often said to be (not necessarily causal) "effects" of X_{1i}, computed while *matching* on values of X_{2i}.
  - Matching on X_{2i} ensures that those contributing to the comparison of averages across values of X_{1i} have the same value of X_{2i} (at least)
  - Matching needn't yield 100% *ceteris paribus* comparisons to be useful and interesting
  - Even when the effect of X_{1i} is unlikely to be causal (as in the differences by ethnicity discussed below), the fact that conditioning on X_{2i} changes the size of the effect of X_{1i} contributes to our understanding of it

- Define the conditional comparison:

$$\delta(X_{2i}) \equiv E[Y_i \mid X_{1i} = 1, X_{2i}] - E[Y_i \mid X_{1i} = 0, X_{2i}].$$

  Note that $\delta(X_{2i})$ is a function of X_{2i} and so has the same number of possible values as does X_{2i}.

1.1 Multivariate Regression Makes Me a Match

- The controls, X_{2i}, often take on many values (either because there are many things to be controlled or because the individual controls take on many values, like SAT scores in the public-private college comparisons discussed below and in *MM* Chapter 2).
  - This threatens to overwhelm us with a multitude of conditional comparisons.
- Regression solves this problem by fitting a linear model with a single conditional effect, while also generating the standard errors needed to do statistical inference for this effect.

Regression is a many-splendored thing. I introduce it by assuming the CEF of Y_i given X_{1i} and X_{2i} is linear:

$$E[Y_i \mid X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{2}$$

Defining $\varepsilon_i = Y_i - E[Y_i \mid X_{1i}, X_{2i}]$, we can write:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

where

$$E[\varepsilon_i \mid X_{1i}, X_{2i}] = 0.$$

The LIE therefore implies that β_0, β_1, β_2 satisfy:

Residual uncorrelated with either regressor

$$E[Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}] = E[\varepsilon_i] = 0 \tag{3}$$
$$E[(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) X_{1i}] = E[\varepsilon_i X_{1i}] = 0 \tag{4}$$
$$E[(Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i}) X_{2i}] = E[\varepsilon_i X_{2i}] = 0 \tag{5}$$

Re-arranging terms shows this to be 3 linear equations in 3 unknowns:

$$\beta_0 + \beta_1 E[X_{1i}] + \beta_2 [X_{2i}] = E[Y_i] \tag{6}$$
$$\beta_0 E[X_{1i}] + \beta_1 E[X_{1i}^2] + \beta_2 E[X_{2i} X_{1i}] = E[Y_i X_{1i}] \tag{7}$$
$$\beta_0 E[X_{2i}] + \beta_1 E[X_{1i} X_{2i}] + \beta_2 E[X_{2i}^2] = E[Y_i X_{2i}] \tag{8}$$

The values of $\beta_0, \beta_1, \beta_2$ that solve this system define the *multivariate regression* of Y_i on X_{1i} and X_{2i}.

- Regression residuals are surely uncorrelated with the regressors that made them (why?)
- What if the CEF is nonlinear?
  - As explained in the appendix to MM Chpt 2 (and detailed in MHE Chpt 3 and Pset 2), multivariate regression provides a best-in-class linear *approximation* to any CEF
  - An important consequence of this approximate awesomeness is that regression is an *automatic matchmaker* (we'll "prove" this by computer)
- With a single regressor, X_{1i}, we can write the linear CEF as $E[Y_i \mid X_{1i}] = \alpha + \beta X_{1i}$. In this case, the solution to equations (3)-(5) can be shown to be:

$$\beta = COV(Y_i, X_{1i}) / V(X_{1i})$$ Covariance divided by variance of regressor
$$\alpha = E[Y_i] - \beta E[X_{1i}]$$

We'll use these *bivariate regression* formulas shortly (be sure you can derive this simple special case from (3)-(5))

Derive α

$E[Y - \alpha - \beta X] = 0$

$\alpha = E[Y] - \beta E[X]$

Derive β

$E[(Y - \alpha - \beta X) X] = 0 \quad E[XY - \alpha X - \beta X^2] = 0 \quad E[XY] - \alpha E[X] - \beta E[X^2] = 0$

$E[XY] - \beta E[X^2] - (E[Y] - \beta E[X]) E[X] = E[XY] - E[X]E[Y] + \beta E[X]^2 - \beta E[X^2] = 0$

$Cov(X,Y) - \beta V(X) = 0 \quad \beta = \frac{Cov(X,Y)}{V(X)}$

2 Regression Talk

- Regression casts random variables in one of three roles
  - a variable to be explained, the *dependent variable*, often denoted by Y_i
    - In many of our examples, wages, grades, health, and test scores are the dependent variables
    - Dependent variables are sometimes called *outcome variables*, especially when the regressor of interest is a treatment dummy
  - *independent variables*, also known as *regressors*, of which, there are usually two types
    - *the regressor of interest*, like a treatment dummy in an experiment, a dummy for health insurance, years of education, or college characteristics
    - *control variables* that help us interpret the coefficient on the regressor of interest, perhaps making it more likely that this is a causal effect
- Sometimes we lump all regressors together, labeling them with generic symbol X_i, which most often denotes a group (or vector) of regressors
- We sometimes use different symbols to distinguish the regressor of interest from control variables. We might, for example, use D_i for treatment status, with controls denoted by W_i. The vector X_i then contains both the focal regressor, D_i, and the controls, W_i.

3 Ordinary Least Squares

We *estimate* regression parameters with sample analogs. For example, to estimate bivariate regression parameters, we use:

$$\hat{\alpha} = \bar{Y} - \bar{X}_1\hat{\beta}$$
$$\hat{\beta} = s_{X_1Y}/s^2_{X_1}$$

Compare to

$$\beta = COV(Y_i, X_{1i})/V(X_{1i})$$
$$\alpha = E[Y_i] - \beta E[X_{1i}]$$

} The actual slope & intercept in CEF

These *Ordinary Least Squares* (OLS) estimators of α and β seem natural and indeed have good statistical properties.

- Traditional 'metrics texts derive OLS as the solution to a sample least squares problem (hence the name). Here's the traditional story in a nutshell:
  - Given observations on a pair of random variables: $\{(Y_i, X_{1i}); i = 1, \ldots, n\}$, you'd like to model Y_i as a linear function of X_{1i}
  - How should you pick the slope and intercept? Minimizing the sample sum of squared errors,

$$\widehat{SSE}_{Y|X}(a, b) = \sum_i (Y_i - a - bX_{1i})^2,$$

    generates $\hat{\alpha}$ and $\hat{\beta}$, above
  - Prove this at home
- OLS estimators, like sample means, are random and subject to sampling variance. We quantify this sampling variance with the tools of *regression inference*.

Derive $\hat{\alpha}$

$\widehat{SSE}_{Y|X}(a,b) = \sum_i (Y_i - a - bX_{1i})^2$

$\frac{d}{da} = \sum 2(Y - a - bX)$

$= 2\sum Y - 2\sum a - 2\sum bX$

$\frac{d}{da} = 0 \quad 2n\bar{Y} - 2na - 2nb\bar{X} = 0$

$\bar{Y} - a - b\bar{X} = 0 \quad a = \bar{Y} - b\bar{X}$

Derive $\hat{\beta}$

$\frac{d}{db} = \sum 2(Y - a - bX)(-X) = -2\sum (X)(Y - a - bX) \quad \frac{d}{db} = 0 \quad \sum (XY - aX - bX^2) = 0$

$\sum XY - \sum aX - \sum bX^2 = 0 \quad \sum XY - a\sum X - b\sum X^2 = 0 \quad \sum XY - n a\bar{X} - b\sum X^2 = 0$

$\sum XY - (\bar{Y} - b\bar{X}) n\bar{X} - b\sum X^2 = 0 \quad \sum XY - n\bar{X}\bar{Y} + nb\bar{X}^2 - b\sum X^2 = 0 \quad \sum (XY - \bar{X}\bar{Y}) - b\sum (X^2 - \bar{X}^2) = 0$

$nS_{XY} - bnS_X^2 = 0 \quad b = S_{XY}/S_X^2$

- These inference tools include:
  1. standard errors and confidence intervals
  2. F-statistics for joint tests
- Details TBD in LN7, but the ideas behind 1 are not really new, so we'll start using these tools today.

$n s_{xy} = \sum(x-\bar{x})(y_i-\bar{y})$
$= \sum(xy + \bar{x}\bar{y} - x\bar{y} - y\bar{x})$
$= \sum xy + n\bar{x}\bar{y} - \bar{y}\sum x - \bar{x}\sum y$
$= \sum xy + n\bar{x}\bar{y} - \bar{y}n\bar{x} - \bar{x}n\bar{y}$
$= \sum xy - n\bar{x}\bar{y}$

$n s_x^2 = \sum(x-\bar{x})^2$
$= \sum(x^2 - 2\bar{x}x + \bar{x}^2)$
$= \sum x^2 - 2\bar{x}\sum x + n\bar{x}^2$
$= \sum x^2 - 2n\bar{x}^2 + n\bar{x}^2$
$= \sum x^2 - n\bar{x}^2 = \sum(x^2 - \bar{x}^2)$

Also by definition

$E[Y|D] = P(D=0)\,E[Y|D=0] + P(D=1)\,E[Y|D=1]$

4 Regression for Dummies

When the conditioning variable is a dummy, say D_i, then $E[Y_i \mid D_i]$ is linear:

$$E[Y_i \mid D_i] = \underbrace{E[Y_i \mid D_i = 0]}_{\alpha} + \underbrace{(E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0])}_{\beta} D_i$$

$D_i = 1 \rightarrow E[Y|D=1]$
$D_i = 0 \rightarrow E[Y|D=0]$

The regression slope and intercept must therefore be: * From derivation of α 2 pages earlier

$$\alpha = E[Y_i|D_i = 0] = E[Y_i] - E[D_i]\beta$$
$$\beta = E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = C(D_i, Y_i)/V(D_i)$$

* From derivation of β 2 pages earlier

$E[Y_i|D_i]$ can only take on two values as D_i is a dummy. Define
$E[Y_i|D_i=0] = \alpha + D_i\beta = \alpha$
$E[Y_i|D_i=1] = \alpha + D_i\beta = \alpha + \beta$
$\therefore \beta = E[Y_i|D_i=1] - E[Y_i|D_i=0]$

* $\alpha = E[Y_i] - \beta E[X_{1i}]$
$\beta = COV(Y_i, X_{1i})/V(X_{1i})$
(why? explain this in Pset 2)

- *Regression estimates differences in means.* We've seen regression used to estimate differences in means in comparisons of treatment and control groups using data from experiments like ALO (2009) and CGW (2017).
- When regressors are discrete and our regression model includes dummies for all of the possible values they might assume, the regression model is said to be *saturated*. The regression function and CEF coincide for saturated models.

Matchmaker Reprise

- We introduced regression by assuming the CEF is linear, in which case regression is it. Otherwise, regression approximates a nonlinear CEF linearly. That leaves us with my claim that regression is also an automatic matchmaker, that is, a simple strategy to make *ceteris paribus* comparisons involving a focal regressor, D_i, while holding control variables, W_i, fixed.
- Here's the formal result behind this claim: Consider coefficient δ in the regression of Y_i on dummy D_i and a vector of saturated dummy controls, W_i:

$$Y_i = W_i'\gamma + \delta D_i + \varepsilon_i, \tag{9}$$

where ε_i is the regression error (controls here include a variable thats always equal to 1; the coefficient on this is the constant)

- Note that (9) has a single treatment effect. In general, however, the difference in means with D_i switched on and off is a function of W_i. That is,

$$\delta(W_i) = E[Y_i|D_i = 1, W_i] - E[Y_i|D_i = 0, W_i],$$

is not a constant

- The regression coefficient δ in (9) can be shown to be a weighted average of $\delta(W_i)$ contrasts. In particular,

$$\delta = E[\delta(W_i)\sigma_D^2(W_i)], \tag{10}$$

where $\sigma_D^2(W_i)$ is the conditional variance function for D_i given W_i. MHE proves this. We'll demonstrate it here by computer.

$\beta = C(D_i,Y_i)/V(D_i) = \dfrac{E[D_iY_i] - E[D_i]E[Y_i]}{p(1-p)} = \dfrac{E[D_iY_i|D_i=1]\cdot P(D_i=1) + E[D_iY_i|D_i=0]\cdot P(D_i=0) - p\left[E(Y_i|D_i=1)\cdot P(D_i=1) + E(Y_i|D_i=0)\cdot P(D_i=0)\right]}{p(1-p)}$

$= \dfrac{E[Y|D_i=1]\cdot p - p^2 E[Y|D_i=1] + p\,E[Y_i|D_i=0](1-p)}{p(1-p)} = \dfrac{E[Y|D_i=1] - p\,E[Y_i|D_i=1] + E[Y_i|D_i=0](1-p)}{1-p}$

$= \dfrac{(1-p)E[Y_i|D_i=1] + (1-p)E[Y_i|D_i=0]}{(1-p)} = E[Y_i|D_i=1] + E[Y_i|D_i=0]$

$\alpha = E[Y_i|D_i=0] = (1-p)E[Y_i|D_i=0] + pE[Y_i|D_i=0] = (1-p)E[Y_i|D_i=0] + (p-p)E[Y_i|D_i=1] + pE[Y_i|D_i=0]$

$= (1-p)E[Y_i|D_i=0] + pE[Y_i|D_i=1] + p[E[Y_i|D_i=1] + E[Y_i|D_i=0]] = E[Y_i] + p\beta = E[Y_i] + E[D_i]\beta$

5 Asians and Whites Under Control

- In a sample of prime age male high school grads in the 2016 American Community Survey, Asians (75% foreign-born) earn more than whites

```
59 . summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agep	57,696	44.632	2.834972	40	49
wagp	57,696	85197.99	88589.83	0	714000
wkhp	57,696	45.16632	10.0141	1	99
racasn	57,696	.0987243	.2982941	0	1
racpi	57,696	.0009706	.0311397	0	1
racwht	57,696	.9083299	.2885622	0	1
uhe	56,924	34.81603	29.37749	0	201.5789
loguhe	53,750	3.361481	.7235695	-6.437752	5.306181
immig	57,696	.1748475	.3798399	0	1
yearsEd	57,696	14.53616	2.42775	12	21
hsgrad	57,696	1	0	1	1
somecol	57,696	.5348551	.498788	0	1
colgrad	57,696	.4422664	.4966599	0	1
asianpac	57,696	.0987243	.2982941	0	1
white	57,289	.9076786	.2894816	0	1

```
60 . bys asianpac: summarize loguhe yearsEd colgrad immig
```

```
-> asianpac = 0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
loguhe	48,411	3.345458	.7155705	-6.437752	5.306181
yearsEd	52,000	14.40617	2.377595	12	21
colgrad	52,000	.4188462	.4933748	0	1
immig	52,000	.11025	.3132041	0	1

```
-> asianpac = 1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
loguhe	5,339	3.506776	.7775642	-3.912023	5.30231
yearsEd	5,696	15.72279	2.555968	12	21
colgrad	5,696	.6560744	.4750583	0	1
immig	5,696	.7645716	.4243035	0	1

- Is the Asian wage effect causal? (Ponder potential outcomes).
- Either way, ethnicity gaps in college graduation rates might explain it

				F(1, 53748)	= 240.08
Model	125.139748	1	125.139748	Prob > F	= 0.0000
Residual	28015.2987	53,748	.521234253	R-squared	= 0.0044
				Adj R-squared	= 0.0044
Total	28140.4384	53,749	.523552781	Root MSE	= .72197

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	.1613188	.0104113	15.49	0.000	.1409126	.1817249
_cons	3.345458	.0032813	1019.56	0.000	3.339026	3.351889

```
63 . reg loguhe asianpac colgrad
```

Source	SS	df	MS		
				Number of obs	= 53,750
				F(2, 53747)	= 5623.46
Model	4869.57938	2	2434.78969	Prob > F	= 0.0000
Residual	23270.859	53,747	.43297038	R-squared	= 0.1730
				Adj R-squared	= 0.1730
Total	28140.4384	53,749	.523552781	Root MSE	= .658

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	.016507	.0095892	1.72	0.085	-.002288	.0353019
colgrad	.6043304	.0057731	104.68	0.000	.593015	.6156457
_cons	3.091722	.0038496	803.14	0.000	3.084176	3.099267

```
64 . reg loguhe asianpac yearsEd
```

Source	SS	df	MS		
				Number of obs	= 53,750
				F(2, 53747)	= 6283.13
Model	5332.56924	2	2666.28462	Prob > F	= 0.0000
Residual	22807.8692	53,747	.424356134	R-squared	= 0.1895
				Adj R-squared	= 0.1895
Total	28140.4384	53,749	.523552781	Root MSE	= .65143

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	-.0109312	.0095219	-1.15	0.251	-.0295942	.0077317
yearsEd	.1301684	.0011751	110.78	0.000	.1278653	.1324715
_cons	1.469512	.0171914	85.48	0.000	1.435816	1.503207

```
65 .
66 . bys colgrad: reg loguhe asianpac
```

-> colgrad = 0

Source	SS	df	MS		
				Number of obs	= 29,903
				F(1, 29901)	= 25.12
Model	9.74715785	1	9.74715785	Prob > F	= 0.0000
Residual	11600.8634	29,901	.387975766	R-squared	= 0.0008
				Adj R-squared	= 0.0008
Total	11610.6105	29,902	.388288761	Root MSE	= .62288

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	-.0755548	.0150739	-5.01	0.000	-.1051003	-.0460093
_cons	3.097319	.0037168	833.34	0.000	3.090034	3.104604

-> colgrad = 1

Source	SS	df	MS		
				Number of obs	= 23,847
				F(1, 23845)	= 29.15
Model	14.2407638	1	14.2407638	Prob > F	= 0.0000
Residual	11647.2907	23,845	.488458407	R-squared	= 0.0012
				Adj R-squared	= 0.0012
Total	11661.5315	23,846	.48903512	Root MSE	= .6989

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	.068885	.0127577	5.40	0.000	.0438791	.0938908
_cons	3.688318	.0049022	752.39	0.000	3.67871	3.697927

- Compare the college-controlled regression estimate of 0.0167 (above) to the average conditional-on-college Asian effect:

$$-.0756\frac{29903}{53750}(=.556)+.0689\frac{23847}{53750}(=.444)\simeq -.01$$

6 Regression Anatomy

System (3) doesn't immediately reveal exactly how multivariate regression works its matching magic. Here's a way to look inside the box.

- Start with a regression equation with two regressors:

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\varepsilon_i \tag{11}$$

- Consider the following two *auxiliary regressions*:

$$X_{1i}=\delta_{10}+\delta_{12}X_{2i}+\tilde{x}_{1i} \quad \text{Regression of } X_1 \text{ on } X_2$$
$$X_{2i}=\delta_{20}+\delta_{21}X_{1i}+\tilde{x}_{2i} \quad \text{Regression of } X_2 \text{ on } X_1$$

where the δ's are bivariate regression coefficients [e.g., $\delta_{12}=COV(X_{1i},\,X_{2i})/V(X_{2i})$], while $\tilde{x}_{1i}$ is the residual from a regression of X_{1i} on X_{2i} and $\tilde{x}_{2i}$ is the residual from a regression of X_{2i} on X_{1i}

The following theoretical result is key to our understanding of multivariate regression models:

The Regression-Anatomy Theorem.

$$\beta_1=COV(Y_i,\tilde{x}_{1i})/V(\tilde{x}_{1i})$$
$$\beta_2=COV(Y_i,\tilde{x}_{2i})/V(\tilde{x}_{2i})$$

Proof. Substitute for Y_i using $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\varepsilon_i$, where ε_i must be mean-zero and uncorrelated with regressors (why?).

- The multivariate β_1 captures the effect of $\tilde{x}_{1i}$, that is, the part of X_1 that is not explained (in a regression sense) by X_2
- The multivariate β_2 captures the effect of $\tilde{x}_{2i}$, that is, the part of X_2 that is not explained (in a regression sense) by X_1

$$COV(Y,\tilde{x}_1)/V(\tilde{x}_1)=\frac{COV(\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon,\tilde{x}_1)}{V(\tilde{x}_1)}=\frac{COV(\beta_1X_1,\tilde{x}_1)}{V(\tilde{x}_1)}=\frac{\beta_1COV(X_1,\tilde{x}_1)}{V(\tilde{x}_1)}=\frac{\beta_1COV(\delta_{10}+\delta_{12}X_2+\tilde{x}_1,\tilde{x}_1)}{V(\tilde{x}_1)}$$

$$=\frac{\beta_1COV(\tilde{x}_1,\tilde{x}_1)}{V(\tilde{x}_1)}=\beta_1$$

Likewise for β_2

```
70 . reg loguhe asianpac yearsEd age
```

Source	SS	df	MS	Number of obs	=	53,750
				F(3, 53746)	=	4223.15
Model	5368.08594	3	1789.36198	Prob > F	=	0.0000
Residual	22772.3525	53,746	.423703205	R-squared	=	0.1908
				Adj R-squared	=	0.1907
Total	28140.4384	53,749	.523552781	Root MSE	=	.65092

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
asianpac	-.008028	.0095198	-0.84	0.399	-.0266869	.0106309
yearsEd	.1302794	.0011742	110.95	0.000	.1279779	.1325808
agep	.0090692	.0009906	9.16	0.000	.0071276	.0110107
_cons	1.062983	.0476094	22.33	0.000	.9696681	1.156298

```
71 .
72 . **step 1
73 .
74 . reg asianpac age yearsEd if e(sample)==1
```

Source	SS	df	MS	Number of obs	=	53,750
				F(2, 53747)	=	766.95
Model	133.428423	2	66.7142115	Prob > F	=	0.0000
Residual	4675.24747	53,747	.086986203	R-squared	=	0.0277
				Adj R-squared	=	0.0277
Total	4808.67589	53,749	.089465402	Root MSE	=	.29493

asianpac	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
agep	-.003466	.0004486	-7.73	0.000	-.0043452	-.0025868
yearsEd	.0200877	.0005249	38.27	0.000	.0190588	.0211166
_cons	-.0381703	.0215712	-1.77	0.077	-.08045	.0041095

```
75 . predict ap_resid, residuals

76 .
77 . **step 2
78 .
79 . reg loguhe ap_resid
```

Source	SS	df	MS	Number of obs	=	53,750
				F(1, 53748)	=	0.58
Model	.30131172	1	.30131172	Prob > F	=	0.4481
Residual	28140.1371	53,748	.523556915	R-squared	=	0.0000
				Adj R-squared	=	-0.0000
Total	28140.4384	53,749	.523552781	Root MSE	=	.72357

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ap_resid	-.008028	.0105823	-0.76	0.448	-.0287693	.0127134
_cons	3.361481	.003121	1077.06	0.000	3.355364	3.367599

```
80 .
81 . log close
```

- It works! Phew!

Fear no more!

7 Regression, Causality, and Control

Does MIT matter? The Dale and Krueger (DK; 2002) study on our reading list looks at difference in earnings between graduates of more and less selective colleges, as measured by average SAT scores of those enrolled. We turn this into a Bernoulli treatment (here and in MM, Chpt 2) by considering the effect of graduation from a private institution (which are more selective than public, on average). Two of my former Ph.D. students were admitted to Harvard yet attended their local state (public) schools. Today, these students are professors in top econ departments - not bad! But perhaps they would have done better if they had attended (private) Harvard instead. Who knows, they might have found jobs on Wall Street!

These are just two data points, of course. But in larger and more representative samples, comparisons between private and state school graduates consistently show higher earnings for those who went private. No surprise! *Something* must justify the many thousands of tuition dollars these schools collect.

Yet, part of the difference in earnings between private and public college grads is surely attributable to differences in the characteristics ($Y_{0i}'s$) of people who did and didn't attend private schools. Variables likely to differ with school type include students' own SAT scores (which are correlated with their earnings), the selectivity of schools they applied to (which says something about students' own judgements of their ability) and family income (which is also correlated with earnings).

- We'd like to hold these things constant, that is, to control for them when comparing groups of students who went to different types of schools
- We hope this control brings us one giant step closer to the average causal effect that would be revealed by an experiment that randomly assigns private attendance

7.1 The Payoff to Private College

The DK (2002) research design, as implemented in Chapter 2 of MM, compares public and private graduates who applied to and were admitted to schools of similar selectivity.

- Consider a hypothetical set of applicants, all of whom applied to one or more schools among three public (All State, Tall State, and Altered State) and three private (Ivy, Leafy, and Smart). The matching matrix these students face appears below:

TABLE 2.1
The college matching matrix

		Private			Public			
Applicant group	Student	Ivy	Leafy	Smart	All State	Tall State	Altered State	1996 earnings
A	1		Reject	Admit		Admit		110,000
	2		Reject	Admit		Admit		100,000
	3		Reject	Admit		Admit		110,000
B	4	Admit			Admit		Admit	60,000
	5	Admit			Admit		Admit	30,000
C	6		Admit					115,000
	7		Admit					75,000
D	8	Reject			Admit	Admit		90,000
	9	Reject			Admit	Admit		60,000

Note: Enrollment decisions are highlighted in gray.

- Five of nine students (numbers 1,2,4,6,7) attended private schools. Average earnings in this group are $92,000. The other four, with average earnings of $72,500, went to a public school. The almost $20,000 gap between these two groups suggests a large private school advantage.

The hypothesis motivating a DK-style analysis is that, conditional on the identity (or selectivity) of schools to which an applicant applies, and the identity (or selectivity) of schools to which an applicant has been admitted, comparisons of students who went to different schools (say, one to public and one to private) are likely to be "apples to apples." In other words, we uncover the effects of private school attendance by ...

- Comparing students 1 and 2 with student 3 in group A and by comparing student 4 and student 5 in Group B
- Discarding students in groups C and D (why?)
- The average of the -5 thousand dollars gap for group A and the 30,000 gap dollars for group B is $12,500. This is a good estimate of the effect of private school attendance on average earnings because it controls (at least partially) for applicants' ambition and ability (a weighted average reflecting the fact that 3/5 of applicants are in Group A is $9,000)
- Notice that overall average earnings in Group A are much higher than overall average earnings in group B. Our within-group matching estimates of 12,500 or 9,000 eliminate this source of selection bias in public-private comparisons

Instead of averaging these group-specific contrasts by hand, regress!

- Limit the analysis to Groups A an B. With only one control variable needed, A_i (a dummy for those in Group A), the regression of interest can be written:

$$Y_i = \alpha + \beta P_i + \gamma A_i + \varepsilon_i \tag{12}$$

- The distinction between the causal variable, P_i, and the control variable, A_i, in equation (12) is conceptual, not formal: nothing in equation (12) indicates which is which.
- Using data for the five students in Groups A and B generates $\beta = 10,000$ and $\gamma = 60,000$. The private school coefficient in this case is $10,000$, close to the estimate obtained by averaging the public-private contrasts within groups A and B and well below the raw public-private difference of almost 20,000.

Public-Private Face-Off

The *College and Beyond* (C&B) data set used in the DK study includes over 14,000 college graduates who attended one of 30 schools. About 2,000 graduates have an exact match, that is, they can be put into groups of two or more who applied to and were accepted by the same schools.

We can increase the number of useful comparisons by deeming schools to be "matched" if they are equally selective instead of insisting on identical matches.

- To fatten up the selectivity categories, call schools comparable if they fall into the same Barron's selectivity categories

9,202 students have Barron's matches, that is, they can put put into groups of two or more who applied to and were accepted by sets of schools in the same Barron's categories. Because we're interested in public-private comparisons, however, our Barron's matched sample is also limited to matched applicant groups that contain both public and private school graduates. This leaves 5,583 matched applicants for analysis. These matched applicants fall into 151 different selectivity groups containing both public and private graduates.

The operational regression model for the Barron's selectivity-matched sample includes many control variables, while the stylized 6-student example controls only for the dummy variable A_i, indicating students in group A. The key controls in the operational model consist of a set of many dummy variables indicating all Barron's matches represented in the sample (with one group left out as a reference category). These controls capture applicant ambition and ability as measured by the selectivity of the schools to which they applied and were admitted in the real world, where many combinations of schools are possible.

The resulting regression model looks like this:

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=1}^{150} \gamma_j GROUP_{ji} + \delta_1 SAT_i + \delta_2 \ln PI_i + \varepsilon_i \tag{13}$$

- The parameter β in this model is the coefficient of interest, an estimate of the causal effect of attendance at a private school
- This model controls for 151 groups instead of the two groups in our stylized example. The parameters γ_j, for $j = 1$ to 150, are the coefficients on 150 selectivity-group dummies, denoted $GROUP_{ji}$
- The model includes two further control variables: individual SAT scores and the log of parental income, plus a few more we haven't bother to write out. The table below reports key findings:

TABLE 2.2
Private school effects: Barron's matches

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.135	.095	.086	.007	.003	.013
	(.055)	(.052)	(.034)	(.038)	(.039)	(.025)
Own SAT score ÷ 100		.048	.016		.033	.001
		(.009)	(.007)		(.007)	(.007)
Log parental income			.219			.190
			(.022)			(.023)
Female			−.403			−.395
			(.018)			(.021)
Black			.005			−.040
			(.041)			(.042)
Hispanic			.062			.032
			(.072)			(.070)
Asian			.170			.145
			(.074)			(.068)
Other/missing race			−.074			−.079
			(.157)			(.156)
High school top 10%			.095			.082
			(.027)			(.028)
High school rank missing			.019			.015
			(.033)			(.037)
Athlete			.123			.115
			(.025)			(.027)
Selectivity-group dummies	No	No	No	Yes	Yes	Yes

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column reports coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The results in columns (4)–(6) are from models that include applicant selectivity-group dummies. The sample size is 5,583. Standard errors are reported in parentheses.

- Perhaps it's enough to control linearly for the average SAT scores of the schools to which I'm admitted, as well as the number of schools to which I apply. Here's how that comes out:

TABLE 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

- This buys us a larger sample and doesn't much change the results

- What about school selectivity (as in DK02) instead of the public/private distinction?

TABLE 2.4
School selectivity effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
School average SAT score ÷ 100	.109	.071	.076	−.021	−.031	.000
	(.026)	(.025)	(.016)	(.026)	(.026)	(.018)
Own SAT score ÷ 100		.049	.018		.037	.009
		(.007)	(.006)		(.006)	(.006)
Log parental income			.187			.161
			(.024)			(.025)
Female			−.403			−.396
			(.015)			(.014)
Black			−.023			−.034
			(.035)			(.035)
Hispanic			.015			.006
			(.052)			(.053)
Asian			.173			.155
			(.036)			(.037)
Other/missing race			−.188			−.193
			(.119)			(.116)
High school top 10%			.061			.063
			(.018)			(.019)
High school rank missing			.001			−.009
			(.024)			(.022)
Athlete			.102			.094
			(.025)			(.024)
Average SAT score of schools applied to ÷ 100				.138	.116	.089
				(.017)	(.015)	(.013)
Sent two applications				.082	.075	.063
				(.015)	(.014)	(.011)
Sent three applications				.107	.096	.074
				(.026)	(.024)	(.022)
Sent four or more applications				.153	.143	.106
				(.031)	(.030)	(.025)

Notes: This table reports estimates of the effect of alma mater selectivity on earnings. Each column shows coefficients from a regression of log earnings on the average SAT score at the institution attended and controls. The sample size is 14,238. Standard errors are reported in parentheses.

- Pity my poor parents, whom I made even poorer by attending Oberlin, a pricey private college. It seems I could just as well have gone to Penn State!

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 6

Understanding Multivariate Regression: The OVB Formula

1 The OVB formula

1.1 One vs. Two

Suppose you'd like to regress log wages (Y_i) on years of schooling (S_i), controlling for ability (A_i):

$$Y_i = \alpha + \rho S_i + \gamma A_i + \varepsilon_i \tag{1}$$

You seek this regression in the hope that controlling for ability mitigates selection bias in estimates of the economic returns to schooling. Alas, you don't observe ability, so you make do with the short regression on schooling alone:

$$Y_i = \alpha^* + \rho^* S_i + v_i$$

$A_i = \delta_0 + \delta_{AS} S_i$

- Substituting (1) into the formula for a bivariate regression slope reveals that:

$$\underbrace{\rho^*}_{short} = \frac{C(Y_i, S_i)}{V(S_i)} = \underbrace{\rho}_{long} + \underbrace{\gamma \delta_{AS}}_{OVB},$$

where δ_{AS} is the regression of A_i on S_i

1 way to prove

$$\frac{C(\alpha + \rho S_i + \gamma A_i + \varepsilon_i, S_i)}{V(S_i)} = \frac{\rho V(S_i)}{V(S_i)} + \gamma \frac{C(A_i, S_i)}{V(S_i)} = \rho + \gamma \delta_{AS}$$

Other way to prove

$$Y_i = \alpha + \rho S_i + \gamma \delta_0 + \gamma \delta_{AS} S_i + \varepsilon_i$$
$$Y_i = (\alpha + \gamma \delta_0) + (\rho + \gamma \delta_{AS}) S_i + \varepsilon_i$$
$$\alpha^* = \alpha + \gamma \delta_0$$
$$\rho^* = \rho + \gamma \delta_{AS}$$

- Neat formula! We say: Basically, the effect we observe is of both the regressor and a covariate related to both the regressor and independent variable

Short equals long -plus- the effect of omitted -times- the regression of omitted on included

  - This *omitted variables bias* (OVB) *formula* is regression's golden rule (the OVB term is $\gamma \delta_{AS}$)

- In a wage equation like (1) where the omitted variable is ability, labor economists refer to OVB as *ability bias*

  - Ponder the sign of ability bias (Is the short reg ρ^* too big or too small relative to the long-regression ρ that you seek?) Ability positively correlated with wages and schooling so will overestimate (ρ^* too big)

1.2 Two vs. Four

Suppose now that your long regression includes four regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i \qquad E[\varepsilon_i X_{ji}] = 0;\ j = 1,\ 2,\ 3,\ 4 \tag{2}$$

- You'd like to estimate the parameters of equation (1), the long regression of your dreams. Alas, you're missing data on X_{3i} and X_{4i}. So, you settle for the short regression with only two:

$$Y_i = \beta_0^* + \beta_1^* X_{1i} + \beta_2^* X_{2i} + \nu_i \qquad E[\nu_i X_{ji}] = 0;\ j = 1,\ 2 \tag{3}$$

- What's the relationship between β_1^* and β_1? Between β_2^* and β_2? The regression anatomy formula formula for β_1^* gives

$$(short)\ \beta_1^* = \frac{C(Y_i, \tilde{x}_{1i})}{V(\tilde{x}_{1i})}, \tag{4}$$

where this *auxiliary regression:*

$$X_{1i} = \gamma_{10} + \gamma_{11} X_{2i} + \tilde{x}_{1i},$$

[remember that residual must be uncorrelated with regressor]

partials out (removes) the influence of X_{2i} on X_{1i}.

- Now, substitute the long reg for Y_i in (4):

$C(X_{1i}, \tilde{x}_{1i})$
$= C(\gamma_{10} + \gamma_{11} X_{2i} + \tilde{x}_{1i}, \tilde{x}_{1i})$
$= C(\tilde{x}_{1i}, \tilde{x}_{1i})$
$= V(\tilde{x}_{1i})$

$$\begin{aligned}
\beta_1^* &= \frac{C(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i,\ \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\
&= \frac{C(\beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i},\ \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\
&= \beta_1 \frac{C(X_{1i},\ \tilde{x}_{1i})}{V(\tilde{x}_{1i})} + \frac{C(\beta_3 X_{3i} + \beta_4 X_{4i},\ \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \\
&= \beta_1 + \beta_3 \frac{C(X_{3i},\ \tilde{x}_{1i})}{V(\tilde{x}_{1i})} + \beta_4 \frac{C(X_{4i},\ \tilde{x}_{1i})}{V(\tilde{x}_{1i})}
\end{aligned}$$

Write this as:

$$\beta_1^* = \beta_1 + \beta_3 \delta_{31.2} + \beta_4 \delta_{41.2},$$

where

Confirm actual equation

$$\delta_{31.2} = \frac{C(X_{3i}, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \quad \text{(Regression of } X_3 \text{ on } X_1 \text{ in a model that includes } X_2\text{)}$$

$X_{3|1.2} = a + bX_1 + cX_2 = a + b(\gamma_{10} + \gamma_{11} X_2 + \tilde{x}_{1i}) + cX_2 = a + b\gamma_{10} + b\tilde{x}_{1i} + (b\gamma_{11} + c)X_2$

$$\delta_{41.2} = \frac{C(X_{4i}, \tilde{x}_{1i})}{V(\tilde{x}_{1i})} \quad \text{(Regression of } X_4 \text{ on } X_1 \text{ in a model that includes } X_2\text{)}$$

$X_{4|1.2} = a + bX_1 + cX_2 = a + b(\gamma_{10} + \gamma_{11} X_2 + \tilde{x}_{1i}) + cX_2 = a + b\gamma_{10} + b\tilde{x}_{1i} + (b\gamma_{11} + c)X_2$

Likewise:

$$\beta_2^* = \beta_2 + \beta_3 \delta_{32.1} + \beta_4 \delta_{42.1},$$

where

Confirm actual equation

$$\delta_{32.1} = \frac{C(X_{3i}, \tilde{x}_{2i})}{V(\tilde{x}_{2i})} \quad \text{(Regression of } X_3 \text{ on } X_2 \text{ in a model that includes } X_1\text{)}$$

$$\delta_{42.1} = \frac{C(X_{4i}, \tilde{x}_{2i})}{V(\tilde{x}_{2i})} \quad \text{(Regression of } X_4 \text{ on } X_2 \text{ in a model that includes } X_1\text{)}$$

- OVB, same as it ever was:

  Short equals long plus {effect(s) of omitted times regression(s) of omitted on included}, with auxiliary regressions computed in a models maintaining the set of controls included in both short and long.

1.3 Sample Short and Long

- OVB formulas hold in any sample as well as in the population from which samples are drawn.
  - Let $\hat{\beta}_1^*$ be the OLS estimate of β_1^* in (3) and let $\hat{\beta}_2^*$ be the corresponding estimate of β_2^*. Then, we have:

$$\hat{\beta}_1^* = \frac{\sum Y_i \tilde{x}_{1i}}{\sum \tilde{x}_{1i}^2} = \hat{\beta}_1 + \hat{\beta}_3 \hat{\delta}_{31.2} + \hat{\beta}_4 \hat{\delta}_{41.2},$$

    where hats denote estimates and $\tilde{x}_{1i}$ is the residual from a regression of X_1 on X_2 in the sample. Likewise,

$$\hat{\beta}_2^* = \frac{\sum Y_i \tilde{x}_{2i}}{\sum \tilde{x}_{2i}^2} = \hat{\beta}_2 + \hat{\beta}_3 \hat{\delta}_{32.1} + \hat{\beta}_4 \hat{\delta}_{42.1},$$

    where $\tilde{x}_{2i}$ is the residual from a regression of X_2 on X_1 in the sample.
  - OVB holds in your data! (Show this at home)

1.4 When Short Equals Long

Two scenarios yield short=long:

1. Omitted variables have coefficients of zero in long (in which case they're not really "omitted") $\gamma = 0$
2. Omitted variables are uncorrelated with included variables, conditional on maintained covariates in both short and long $\delta_{AS} = 0$

2 Empirical OVB

Immigrant and native wages (working men aged 40-49 in the 2016 ACS)

Variable	Obs	Mean	Std. Dev.	Min	Max
agep	67,179	44.59836	2.843473	40	49
wagp	67,179	77017	70468.34	4	665000
wkhp	67,179	44.73532	9.786132	1	99
uhe	67,179	33.90219	27.61486	.0016	201.5789
loguhe	67,179	3.264571	.7410341	-6.437752	5.306181
immig	67,179	.2214829	.4152479	0	1
yearsEd	67,179	13.83362	3.240573	0	21
hsgrad	67,179	.9243365	.264461	0	1
somecol	67,179	.4721565	.4992279	0	1
colgrad	67,179	.3860581	.4868478	0	1
asianpac	67,179	.0833147	.2763594	0	1
white	66,790	.7721216	.4194669	0	1
married	67,179	.7207461	.4486359	0	1

```
58 .
59 . ***short vs long***
60 .
61 . reg loguhe immig
```

Source	SS	df	MS	Number of obs	=	67,179
				F(1, 67177)	=	570.52
Model	310.655619	1	310.655619	Prob > F	=	0.0000
Residual	36578.9048	67,177	.544515307	R-squared	=	0.0084
				Adj R-squared	=	0.0084
Total	36889.5604	67,178	.549131567	Root MSE	=	.73791

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.1637641	.0068562	-23.89	0.000	-.1772022	-.1503259
_cons	3.300842	.0032267	1022.99	0.000	3.294518	3.307166

```
62 .         gen beta_short=_b[immig]

63 . reg loguhe immig yearsEd
```

Source	SS	df	MS	Number of obs	=	67,179
				F(2, 67176)	=	8096.70
Model	7165.30841	2	3582.6542	Prob > F	=	0.0000
Residual	29724.252	67,176	.442483208	R-squared	=	0.1942
				Adj R-squared	=	0.1942
Total	36889.5604	67,178	.549131567	Root MSE	=	.66519

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.0346034	.0062671	-5.52	0.000	-.0468869	-.02232
yearsEd	.0999527	.0008031	124.46	0.000	.0983786	.1015267

```
64 .         gen beta_long=_b[immig]

65 .         gen gamma_long=_b[yearsEd]

66 .
67 . **Regression of omitted on included (aux reg)**
68 .
69 . reg yearsEd immig
```

Source	SS	df	MS	Number of obs	=	67,179
				F(1, 67177)	=	1893.82
Model	19342.5545	1	19342.5545	Prob > F	=	0.0000
Residual	686114.857	67,177	10.2135382	R-squared	=	0.0274
				Adj R-squared	=	0.0274
Total	705457.411	67,178	10.5013161	Root MSE	=	3.1959

yearsEd	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-1.292219	.0296939	-43.52	0.000	-1.350419	-1.234019
_cons	14.11983	.0139745	1010.40	0.000	14.09244	14.14722

```
70 .         gen delta=_b[immig]

71 .
72 . **check OVB formula**
73 .
74 .         gen short_chk = beta_long + delta*gamma_long

75 .
76 .         sum short_chk beta_short beta_long gamma delta
```

Variable	Obs	Mean	Std. Dev.	Min	Max
short_chk	67,179	-.1637641	0	-.1637641	-.1637641
beta_short	67,179	-.1637641	0	-.1637641	-.1637641
beta_long	67,179	-.0346034	0	-.0346034	-.0346034
gamma_long	67,179	.0999527	0	.0999527	.0999527
delta	67,179	-1.292219	0	-1.292219	-1.292219

```
77 .
78 . ***repeat with maintained controls***
79 .
80 . cap drop delta short_chk beta_short beta_long gamma_long delta

81 .
82 . reg loguhe immig married age
```

Source	SS	df	MS	Number of obs	=	67,179
				F(3, 67175)	=	1261.06
Model	1966.79504	3	655.598348	Prob > F	=	0.0000
Residual	34922.7653	67,175	.519877415	R-squared	=	0.0533
				Adj R-squared	=	0.0533
Total	36889.5604	67,178	.549131567	Root MSE	=	.72103

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.1844893	.0067131	-27.48	0.000	-.197647	-.1713317
married	.3467611	.0062125	55.82	0.000	.3345845	.3589376
agep	.0071519	.0009788	7.31	0.000	.0052334	.0090704
_cons	2.736543	.0439402	62.28	0.000	2.65042	2.822666

```
83 .         gen beta_short=_b[immig]

84 . reg loguhe immig yearsEd married age
```

Source	SS	df	MS	Number of obs	=	67,179
				F(4, 67174)	=	4753.57
Model	8138.3322	4	2034.58305	Prob > F	=	0.0000
Residual	28751.2282	67,174	.428011257	R-squared	=	0.2206
				Adj R-squared	=	0.2206
Total	36889.5604	67,178	.549131567	Root MSE	=	.65423

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.055513	.0061851	-8.98	0.000	-.0676358	-.0433901
yearsEd	.095556	.0007958	120.08	0.000	.0939963	.0971157
married	.2638873	.0056791	46.47	0.000	.2527563	.2750183
agep	.0086363	.0008882	9.72	0.000	.0068954	.0103772
_cons	1.379621	.0414398	33.29	0.000	1.298399	1.460843

```
-------------+------------------------------------------------------------

85 .         gen beta_long=_b[immig]

86 .         gen gamma_long=_b[yearsEd]

87 .
88 . **Regression of omitted on included (aux reg)**
89 .
90 . reg yearsEd immig married age

      Source |       SS           df       MS      Number of obs   =    67,179
-------------+----------------------------------   F(3, 67175)     =    979.45
       Model |  29564.8411         3  9854.94703   Prob > F        =    0.0000
    Residual |   675892.57    67,175  10.0616683   R-squared       =    0.0419
-------------+----------------------------------   Adj R-squared   =    0.0419
       Total |  705457.411    67,178  10.5013161   Root MSE        =     3.172

------------------------------------------------------------------------------
     yearsEd |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       immig |  -1.349747   .0295329   -45.70   0.000    -1.407631   -1.291862
     married |   .8672798   .0273309    31.73   0.000     .8137113    .9208482
        agep |  -.0155344   .0043061    -3.61   0.000    -.0239744   -.0070944
       _cons |   14.20029   .1933065    73.46   0.000     13.82141    14.57917
------------------------------------------------------------------------------

91 .         gen delta=_b[immig]

92 .
93 . **check OVB formula**
94 .
95 .         gen short_chk = beta_long + delta*gamma_long

96 .
97 .         sum short_chk beta_short beta_long gamma_long delta

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   short_chk |     67,179   -.1844894           0  -.1844894  -.1844894
  beta_short |     67,179   -.1844893           0  -.1844893  -.1844893
   beta_long |     67,179    -.055513           0   -.055513   -.055513
  gamma_long |     67,179     .095556           0    .095556    .095556
       delta |     67,179   -1.349747           0  -1.349747  -1.349747

98 .
99 .         log close
       name:  <unnamed>
        log:  /Users/joshangrist/Documents/teaching/14.32/2020/1432apps/LN8log.smcl
   log type:  smcl
  closed on:   2 Mar 2020, 14:32:28
-------------------------------------------------------------------------------
```

- works again - phew!

Private college redux

TABLE 2.3
Private school effects: Average SAT score controls

	No selection controls			Selection controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	.212	.152	.139	.034	.031	.037
	(.060)	(.057)	(.043)	(.062)	(.062)	(.039)
Own SAT score ÷ 100		.051	.024		.036	.009
		(.008)	(.006)		(.006)	(.006)
Log parental income			.181			.159
			(.026)			(.025)
Female			−.398			−.396
			(.012)			(.014)
Black			−.003			−.037
			(.031)			(.035)
Hispanic			.027			.001
			(.052)			(.054)
Asian			.189			.155
			(.035)			(.037)
Other/missing race			−.166			−.189
			(.118)			(.117)
High school top 10%			.067			.064
			(.020)			(.020)
High school rank missing			.003			−.008
			(.025)			(.023)
Athlete			.107			.092
			(.027)			(.024)
Average SAT score of schools applied to ÷ 100				.110	.082	.077
				(.024)	(.022)	(.012)
Sent two applications				.071	.062	.058
				(.013)	(.011)	(.010)
Sent three applications				.093	.079	.066
				(.021)	(.019)	(.017)
Sent four or more applications				.139	.127	.098
				(.024)	(.023)	(.020)

Notes: This table reports estimates of the effect of attending a private college or university on earnings. Each column shows coefficients from a regression of log earnings on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

- Why does the private wage premium fall as we move from column 1 to column 2?

Controlling for own SAT score which is positively correlated with both earnings & likelihood of attending private colleges.

TABLE 2.5
Private school effects: Omitted variables bias

	Dependent variable					
	Own SAT score ÷ 100			Log parental income		
	(1)	(2)	(3)	(4)	(5)	(6)
Private school	1.165	1.130	.066	.128	.138	.028
	(.196)	(.188)	(.112)	(.035)	(.037)	(.037)
Female		−.367			.016	
		(.076)			(.013)	
Black		−1.947			−.359	
		(.079)			(.019)	
Hispanic		−1.185			−.259	
		(.168)			(.050)	
Asian		−.014			−.060	
		(.116)			(.031)	
Other/missing race		−.521			−.082	
		(.293)			(.061)	
High school top 10%		.948			−.066	
		(.107)			(.011)	
High school rank missing		.556			−.030	
		(.102)			(.023)	
Athlete		−.318			.037	
		(.147)			(.016)	
Average SAT score of schools applied to ÷ 100			.777			.063
			(.058)			(.014)
Sent two applications			.252			.020
			(.077)			(.010)
Sent three applications			.375			.042
			(.106)			(.013)
Sent four or more applications			.330			.079
			(.093)			(.014)

Notes: This table describes the relationship between private school attendance and personal characteristics. Dependent variables are the respondent's SAT score (divided by 100) in columns (1)–(3) and log parental income in columns (4)–(6). Each column shows the coefficient from a regression of the dependent variable on a dummy for attending a private institution and controls. The sample size is 14,238. Standard errors are reported in parentheses.

- Test the OVB formula: Take *Short* to be the bivariate reg reported col 1 of MM T2.3, and *Long* to be the reg that adds individual SAT scores, reported in col 2. We see that:

$$Short - Long = OVB = .212 - .152 = .06.$$

As can also be seen in column 2 of T2.3, the effect of SAT in the long regression is .051, while MM T2.5 (above) shows the regression of SAT (omitted in short) on a private school dummy (included in short) produces a coefficient of 1.165. Putting these pieces together, we confirm *OVB= Reg of omitted on included × Effect of omitted in Long = 1.165×.051=.06.* Phew!

- Why, then, are SAT_i and other controls irrelevant for the private premium in cols 4-6 of MM T2.3?

3 OVB and Selection Bias

- The OVB formula is algebra, true for any short-vs-long regression comparison
- Even so, we often use the OVB formula to understand selection bias -- the math behind this is simple but the idea is subtle – the key is a conditional independence assumption much like that used in LN4 to understand RCTs. Our thinking is thus:
  1. Why do we *care* to go long? As we've seen, private Y_{0i}'s are better (on average)!
  2. Regression with the right controls reduces, maybe even eliminates, the selection bias that arises from imbalanced Y_{0i}
  3. Verily, a regression coefficient so blessed should have no OVB in the sense that, once key controls are included, it matters not whether we add more
- A constant-effects causal effects model helps formalize this argument:
  - Let $Y_{0i} = \alpha + \eta_i$, where $E[Y_{0i}] = \alpha$; assume $Y_{1i} - Y_{0i} = \rho$
  - This means

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})P_i = \alpha + \rho P_i + \eta_i \tag{5}$$

- Private college isn't randomly assigned, so

$$E[\eta_i|P_i] \neq 0$$

  - Because P_i is Bernoulli, the regression of Y_i on P_i produces

$$E[Y_i|P_i = 1] - E[Y_i|P_i = 0] = \rho + \{E[\eta_i|P_i = 1] - E[\eta_i|P_i = 0]\},$$

that is, the sum of a causal effect plus selection bias

$E[Y_i|P_i=1] = E[\alpha + \rho P_i + \eta_i | P_i = 1]$
$= E[\alpha|P_i=1] + E[\rho P_i|P_i=1] + E[\eta_i|P_i=1]$
$= \alpha + \rho + E[\eta_i|P_i=1]$

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

$E[Y_i|P_i=1] - E[Y_i|P_i=0] =$
$\rho + E[\eta_i|P_i=1] - E[\eta_i|P_i=0]$

Regression in Pursuit of Causal Effects

- The causal interpretation of regression estimates is based on a key claim, also called an *identifying assumption* (we'll learn more about econometric identification later in the course)
  - The claim supporting causal inference in this case is that you observe a vector of control variables, X_i, that satisfy the following *conditional independence assumption* (CIA):

$$E[Y_{0i}|\underbrace{P_i}_{poof!}; X_i] = E[Y_{0i}|X_i] = \alpha + \gamma' X_i \tag{6}$$

    Equivalently,

$$Y_{0i} = \alpha + \gamma' X_i + u_i,$$

    where $E[u_i X_i] = 0$ *by construction* and $E[u_i P_i] = 0$ by virtue of the CIA
  - In other words, conditional on X_i, mean Y_{0i} does not depend on P_i → This means selection bias is eliminated
- This leads to a causal regression model: $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})P_i = \alpha + \gamma' X_i + u_i + \rho P_i$

$$Y_i = \alpha + \gamma' X_i + \rho P_i + u_i, \tag{7}$$

where X_i and P_i are both uncorrelated with u_i and coefficient ρ is the causal effect defined in (5)

$E[Y_{0i}|P_i,X_i] - E[Y_{0i}|X_i]$
$E[E[Y_{0i}|P_i,X_i,W_i]|P_i,X_i] = E[E[Y_{0i}|X_i,P_i,W_i]|X_i]$
$E[\alpha_1+\gamma_1'X_i+\delta'W_i|P_i,X_i] = E[\alpha_1+\gamma_1'X_i+\delta'W_i|X_i]$
$E[\delta'W_i|P_i,X_i] = E[\delta'W_i|X_i]$
$E[W_i|P_i,X_i] = E[W_i|X_i]$

– X_i in DK02 and MM Chpt. 2 contains a vector of dummies for Barrons selectivity groups; DK02 argues that the CIA holds given these key controls

- Consider now a longer regression that conditions on pre-treatment covariates W_i as well as X_i. If these additional controls are worth considering they should predict Y_{0i}. This suggests:

$$E[Y_{0i}|P_i; X_i, W_i] = \alpha_1 + \gamma_1' X_i + \delta' W_i, \tag{8}$$

where α_1 is a new intercept, γ_1 is a new slope, and $\delta \neq 0$.

– Given equation (6), however,

$$E[Y_{0i}|P_i, X_i] = E\{E[Y_{0i}|P_i; X_i, W_i]|P_i, X_i\} \overset{(8)}{=} E\{\alpha_1 + \gamma_1' X_i + \delta' W_i|P_i, X_i\} \overset{(6)}{=} E[Y_{0i}|X_i]. \quad = \alpha + \gamma' X_i$$

using LIE

– Hence, conditional on X_i, the CIA says conditional mean W_i does not depend on P_i

- Whence OVB? For purposes of this discussion, short is equation (7), while long adds W_i to this

– Given equation (6), we expect these short and long models to yield similar estimates of ρ

– In other words, no OVB!

This means the regression of W_i on P_i controlling for X_i should be 0

- Can you tell when your long regression is long enough? Not conclusively. Still, given a set of key controls that you believe is sufficient to eliminate selection bias, other controls should change estimates of the causal effect of interest little

– This happy scenario appears in cols 4-6 of MM T2.2-2.4

4 An OVB Classic: Ability Bias

- MHE Table 3.2.1 compares schooling coefficients estimated with and without controls for family background, AFQT scores (a measure of ability), and occupation

Table 3.2.1: Estimates of the returns to education, males

	(1)	(2)	(3)	(4)	(5)
Controls:	None	Age dummies	Col. (2) and additional controls*	Col. (3) and AFQT score	Col. (4), with occupational dummies
	0.132	0.131	0.114	0.087	0.066
	(0.007)	(0.007)	(0.007)	(0.009)	(0.010)

Notes: Data are from the National Longitudinal Survey of Youth (1979 cohort, 2002 survey)
The number in the first row is the coefficient on years of education in a weighted least squares regression of education on wages with the indicated controls. The number in parentheses is the associated standard error. The sample is restricted to males, weighted by NLSY sampling weights, and the sample size is 2434.
* Additional controls are mother's/father's years of education, and dummy variables for race and Census region.

(Beware bad control!)

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 7

Regression Inference

Start with the bivariate population slope and intercept, $\beta = \frac{C(Y_i, X_i)}{V(X_i)}$ and $\alpha = E[Y_i] - \beta E[X_i]$. The corresponding OLS estimates of these parameters can be written:

$$\hat{\beta} = s_{XY}/s_X^2$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X},$$

where s_{XY} and s_X^2 are sample covariance and variance, respectively. Like all sample statistics, OLS estimates vary from sample to sample; the corresponding standard errors quantify the resulting sampling variance.

These notes summarize the theory of statistical inference for regression. We focus on $\hat{\beta}$, first deriving the relevant sampling distribution assuming fixed regressors and Normally distributed errors. These are *classical regression assumptions* (like the Normal data assumption in our first pass at a sampling distribution for sample means in LN2). We then follow up with *asymptotic theory*, which uses the CLT to derive approximate sampling distributions without the need for restrictive, unrealistic assumptions.

1 The Wisdom of the Ancients

Classical Regression Assumptions

1. The CEF of Y_i given X_i is linear, in which case we can write:

$$E[Y_i \mid X_i] = \alpha + \beta X_i,$$

   where α and β are as defined above

2. Define the regression residual $\varepsilon_i = Y_i - E[Y_i \mid X_i] = Y_i - [\alpha + \beta X_i]$. Classicists assume:

   (a) $E[\varepsilon_i^2 \mid X_i] = E[\varepsilon_i^2] = \sigma_\varepsilon^2$ Variance of residual is constant — homoskedasticity

   (b) $E[\varepsilon_i \varepsilon_j] = 0;\ i \neq j$ Residuals uncorrelated with each other [No serial correlation] — random sampling

   (c) ε_i is normally distributed — Normality

3. X_i is fixed in repeated samples

Ancient econometricians derived the sampling distribution of $\hat{\beta}$ under a sampling scheme such that, for the first observation someone with $X_i = x_1$ is always picked, but among these Y_i is drawn randomly; for the second observation someone with $X_i = x_2$ is always picked, but among these Y_i is drawn randomly, etc. In this sampling scheme, we get a fixed *distribution* of X_i in every sample. Imagine, for example, you plan to study the difference in wages between MIT and Harvard grads. Let X_i be a dummy indicating MIT grads. A sample design that always selects 100 MIT grads and 100 Harvard grads fixes the distribution of X_i in your sample.

Why should we assume the distribution of regressors is fixed while the dependent variable is treated as random? Expedience: this assumption simplifies sampling theory and doesn't matter much in practice.

- Most econometrics texts also list the following "assumptions":
  (d) $E[\varepsilon_i] = 0$
  (e) $E[X_i\varepsilon_i] = 0$
  - Given our definition of ε_i, however, these are not assumptions. Rather, the regression slope and intercept *are defined* so as to make these things true (a point noted in LN5, and discussed in the appendix to MM Chapter 2 and MHE Section 3.1.3).
- We're much concerned with whether and when regression parameters have a causal interpretation. As discussed in LN6 and MHE 3.2, this turns on a conditional independence assumption that links these parameters with potential outcomes. The question of whether regression *parameters* are causal is divorced from the technical question of how to quantify the sampling variance of regression *estimates*.

We derive the sampling distribution of OLS estimators using this rewrite:

$$\begin{aligned}\hat{\beta} = s_{XY}/s_X^2 &= \frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2} \\ &= \frac{\sum(X_i-\bar{X})(\alpha+\beta X_i+\varepsilon_i)}{\sum(X_i-\bar{X})^2} = \beta + \frac{\sum(X_i-\bar{X})\varepsilon_i}{\sum(X_i-\bar{X})^2} \qquad (1)\end{aligned}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = \alpha + \beta\bar{X} + \sum\frac{\varepsilon_i}{n} - \hat{\beta}\bar{X} = \alpha + (\beta-\hat{\beta})\bar{X} + \sum\frac{\varepsilon_i}{n} \qquad (2)$$

This isolates the randomness in OLS estimates: under classical assumptions, this is due to randomness in residuals alone.

Classical Regression Statistical Properties

Result 1 The sample slope and intercept are unbiased:

Let $a_i = \frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}$ $\quad E[\sum a_i\varepsilon_i] = \sum a_i E[\varepsilon_i] = 0$

$$E(\hat{\beta}) = \beta + E\left[\frac{\sum(X_i-\bar{X})\varepsilon_i}{\sum(X_i-\bar{X})^2}\right] = \beta$$

$$E(\hat{\alpha}) = \alpha + E[(\beta-\hat{\beta})\bar{X}] + E\left[\sum\tfrac{\varepsilon_i}{n}\right] = \alpha$$

Result 2 The sampling variance of $\hat{\beta}$ is $\frac{\sigma_\varepsilon^2}{ns_X^2}$.

$s_X^2 = \frac{\sum(X_i-\bar{X})^2}{n} \Rightarrow ns_X^2 = \sum(X_i-\bar{X})^2$

Note that

$$V(\hat{\beta}) = V\left[\frac{\sum(X_i-\bar{X})\varepsilon_i}{\sum(X_i-\bar{X})^2}\right] = V[\sum a_i\varepsilon_i] = \sum V(a_i\varepsilon_i) = \sum a_i^2 V(\varepsilon_i)$$
$$= \sigma_\varepsilon^2\sum a_i^2 = \sigma_\varepsilon^2/\sum(X_i-\bar{X})^2 = \sigma_\varepsilon^2/ns_X^2$$

where $a_i = \frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}$. Now use rules for the variance of l.c.'s of r.v.s (review these in LN1) to simplify $V[\sum a_i\varepsilon_i] = \frac{\sigma_\varepsilon^2}{ns_X^2}$

Take square root of sampling variance

- Larger samples and increased variability in X_i produce a more precisely estimated slope. Variance in regressors is good!
- $\frac{\sigma_\varepsilon}{\sqrt{n}s_X}$ is the *standard error* of the sample slope

$a_i = \frac{(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}$ $\quad \sum a_i = \frac{\sum(X_i-\bar{X})}{\sum(X_i-\bar{X})^2}$

$\sum a_i^2 = \frac{\sum(X_i-\bar{X})^2}{[\sum(X_i-\bar{X})^2]^2} = \frac{\sum(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2\cdot\sum(X_i-\bar{X})^2}$

$= \frac{1}{\sum(X_i-\bar{X})^2}$

- In practice, we work with *estimated standard errors*: $\frac{s_e}{\sqrt{n}s_X}$, where *estimated residuals* are

$$e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i,$$

and s_e^2 is their sample variance (this is an estimate of σ_ε^2)

Result 3 $\hat{\beta}$ is Normally distributed; in particular, $\hat{\beta} \sim N\left[\beta, \frac{\sigma_\varepsilon^2}{ns_X^2}\right]$

Proof. We've already derived the mean and variance of $\hat{\beta}$. Note that $\hat{\beta} = \beta + \sum a_i\varepsilon_i$, and we've assumed ε_i is Normally distributed. Since a linear combination of Normal r.v.s is Normal, we're done.

Result 4 Under classical assumptions, the OLS estimator $\hat{\beta}$ is a best linear unbiased estimator (BLUE) of β. This famous result is called the *Gauss-Markov Theorem*.

- $\hat{\beta}$ is said to be a 'linear estimator" because it's a linear combination of the Y_i. In particular, $\hat{\beta} = \sum a_i Y_i$
- $\hat{\beta}$ is "best" in the linear class because any other linear unbiased estimator, say $\hat{b} = \sum b_i Y_i$, for some other constants b_i not equal to a_i such that $E[\hat{b}] = \beta$, must have sampling variance no smaller than that of $\hat{\beta} = \sum a_i Y_i$.

Proof that OLS is BLUE (TBD in recitation).

OLS estimator has lowest sampling variance within class of linear unbiased estimators.

(Wait - you mean there are *other* unbiased estimators of β besides OLS? See Pset 4!)

2 Using the Theory

Hypothesis testing

Test $H_0 : \beta = \beta_0$ using the regression t-statistic:

$$T_n = \frac{\hat{\beta} - \beta_0}{s_e/(\sqrt{n}s_X)} \sim t(n-2)$$

where s_e^2 is an estimate of σ_ε^2 as before (why $n-2$ df?). Recall that:

$$T_n \sim_{approx} N(0,1)$$

We therefore reject H_0 when T_n is a sufficiently surprising draw from a $t(n-2)$ or from a standard Normal distribution.

Confidence intervals

Let c_α be the critical value for a two-sided α-level hypothesis test. Then:

$$P[-c_\alpha \leq T_n \leq c_\alpha] = 1 - \alpha$$

Substituting, we have:

$$P\left[\hat{\beta} - c_\alpha\left(\frac{s_e}{\sqrt{n}s_X}\right) \leq \beta \leq \hat{\beta} + c_\alpha\left(\frac{s_e}{\sqrt{n}s_X}\right)\right] = 1 - \alpha$$

For large n, $T_n \approx N(0,1)$, so for $\alpha = .05$, we have $c_\alpha \approx 2$.

Midterm Stuff Ends Here

3 Multivariate Regression Standard Errors

Suppose you've got four regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i \qquad E[\varepsilon_i X_{ji}] = 0;\ j = 1,\ 2,\ 3,\ 4 \tag{3}$$

The estimated standard error of $\hat{\beta}_1$ is

$$\frac{s_e}{\sqrt{n} s_{\tilde{x}_1}}$$

where s_e is the sample standard deviation of estimated residuals in (3) and $s^2_{\tilde{x}_1}$ is the variance of the residual from a regression of X_{1i} on other regressors in the model (with fixed regressors, we needn't distinguish between population and estimated residuals in auxiliary regressions).

4 Check the Theory: A Small Sampling Experiment

- The pop parameter here is an immigration coefficient estimated using data from the March 2012 Current Population Survey, with controls for age, age^2, and female
- Don't give me p-vals, yo. It's all about that standard error!

```
 8 .
 9 .         // 14.32 Lecture Note 7: c.i. coverage for regression estimates of immigrant-native wage gap
10 .
11 .         // Insheet data from 2016 ACS PUMS
12 .         import delimited "ss16pusa.csv", clear
   (284 vars, 1,623,216 obs)

13 .
14 .         // Sample of interest is working men and women aged 40-49
15 .         keep if age>=40 & age <=49
   (1,427,225 observations deleted)

16 .         keep if wkw==1
   (62,348 observations deleted)

17 .
18 .         // Generate usual hourly earnings
19 .         lab var wagp "Raw annual earnings"

20 .         lab var wkhp "Usual hours per week"

21 .         gen uhe = wagp/(50*wkhp)

22 .         quietly sum uhe, d

23 .         replace uhe=. if uhe>`r(p99)'
   (1,331 real changes made, 1,331 to missing)

24 .         lab var uhe "Usual hourly earnings (truncated at 99pct)"

25 .         gen loguhe=log(uhe)
   (8,324 missing values generated)

26 .         lab var loguhe "Log hourly earnings"

27 .
28 .         // Code regressors
29 .         gen immig=(nativity==2)

30 .         lab var immig "Foreign born"

31 .         gen female=(sex==2)

32 .         gen age2=agep^2

33 .
34 . reg loguhe immig agep age2 female
```

Source	SS	df	MS	Number of obs	=	125,319
				F(4, 125314)	=	1013.10
Model	2067.20528	4	516.80132	Prob > F	=	0.0000
Residual	63925.2846	125,314	.510120853	R-squared	=	0.0313
				Adj R-squared	=	0.0313
Total	65992.4899	125,318	.526600248	Root MSE	=	.71423

loguhe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
immig	-.1498594	.0049374	-30.35	0.000	-.1595366	-.1401823
agep	.0829035	.0246283	3.37	0.001	.0346324	.1311746
age2	-.0008836	.0002765	-3.20	0.001	-.0014254	-.0003417
female	-.2275821	.0040461	-56.25	0.000	-.2355124	-.2196519
_cons	1.358982	.5467118	2.49	0.013	.2874358	2.430527

```
35 . gen popbeta=_b[immig]

36 .
37 .     // Keep only key variables in a newdata set called "workingextract"
```

```
39 .         drop if missing(loguhe)
   (8,324 observations deleted)

40 .     keep wagp loguhe uhe wkhp immig agep age2 female popbeta

41 .         save workingextract, replace
   file workingextract.dta saved

42 .
43 . summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agep	125,319	44.61104	2.845942	40	49
wagp	125,319	65587.55	60054.73	4	665000
wkhp	125,319	42.56813	9.995274	1	99
uhe	125,319	30.06187	23.8018	.0013333	183.2727
loguhe	125,319	3.15407	.7256723	-6.620073	5.210975
immig	125,319	.2121466	.4088297	0	1
female	125,319	.4655479	.4988136	0	1
age2	125,319	1998.244	253.5185	1600	2401
popbeta	125,319	-.1498594	0	-.1498594	-.1498594

```
44 . keep if _n==1
   (125,318 observations deleted)

45 . keep popbeta

46 . save expresults, replace
   file expresults.dta saved

47 .
48 . /* draw samples of N=100 500 times compute regression se */
49 .
50 .    forvalues s=1/500 {
  2.         quietly use workingextract, clear
  3.         quietly bsample 100
  4.         quietly reg loguhe immig agep age2 female
  5.         quietly gen betahat=_b[immig]
  6.         quietly gen sebeta=_se[immig]
  7.         quietly gen hi95 = betahat + (sebeta*1.96)
  8.         quietly gen lo95 = betahat - (sebeta*1.96)
  9.         quietly gen cover95=0
 10.         quietly replace cover95 = 1 if lo95<=popbeta &  popbeta<=hi95
 11.         quietly keep if _n==1
 12.         quietly append using expresults
 13.         quietly save expresults, replace
 14.    }

51 .
52 .     keep popbeta betahat sebeta hi95 lo95 cover95

53 .     drop if missing(betahat)
   (1 observation deleted)

54 .
55 .         /* sampling experiment results */
56 .
57 .         list if _n<=5
```

	popbeta	betahat	sebeta	hi95	lo95	cover95
1.	-.1498594	-.3002854	.1793579	.0512562	-.6518269	1
2.	-.1498594	.1582616	.195466	.5413749	-.2248517	1
3.	-.1498594	-.0601462	.1675861	.2683227	-.388615	1
4.	-.1498594	-.2011411	.2040096	.1987176	-.6009999	1
5.	-.1498594	.0736945	.2200027	.5048999	-.3575108	1

```
58 .         summarize betahat lo95 hi95 cover95
```

Variable	Obs	Mean	Std. Dev.	Min	Max
betahat	500	-.1592314	.19725	-.7738841	.5400975
lo95	500	-.511066	.1978349	-1.083843	.1105932
hi95	500	.1926032	.2073171	-.4639261	.9696018
cover95	500	.918	.2746395	0	1

- Coverage not perfect, but not bad. And this in spite of the fact that the dependent variable isn't Normally distributed, the regressor is just as random as the dependent variable, the CEF isn't quite linear, and the residuals aren't homoskedastic. What's up with that?! As we'll soon see, it's the CLT!
- As we'll also shortly see, *robust* standard errors, which allow for heteroskedasticity, improve on this.

5 Modern Times: Regression in Asymptopia

Life in asymptopia requires large samples but few assumptions. In particular, we no longer need assume linear CEFs, fixed regressors, homoskedastic or Normal residuals. Random sampling of independent observations is enough.

- Asymptopians work with *probability limits*, that is, the LLN and its corollaries, instead of relying on unbiasedness to justify choice of estimator
- Asymptopians use the CLT to argue that, in large samples, sampling distributions are approximately Normal, instead of deriving an exact sampling distribution valid for any sample size

The relevant asymptotic theory is detailed in MHE 3.1.3, and sketched below.

5.1 Consistency of the Sample Slope

1. *Laws of large numbers* (LLN; review definition of *plim* in LN2)
   (a) Let $\bar{X}_n$ denote the sample mean of X_i in a sample of size n. Then: $plim_{n\to\infty}\bar{X}_n = \mu_X$.
   - We say: "The sample mean is a *consistent estimator* of the pop mean."

   (b) LLN for other sample moments. Let $m_n^{r+s} = \sum \frac{X_i^r Y_i^s}{n}$. Then: $plim_{n\to\infty} m_n^{r+s} = E[X_i^r Y_i^s]$.
   - *Any* sample moment is a consistent estimator of the corresponding population moment.

   (c) Plimming functions. Suppose that $plim\, A_n = a$ and $plim\, B_n = b$. Let $h(A_n, B_n)$ be any function continuous at a, b. Then
   $$plim\ h(A_n, B_n) = h(a, b).$$
   - This is called the *Continuous Mapping Theorem*.

Result 1 $\hat{\beta}_n$, the OLS slope estimator computed in a sample of size n, is consistent for β.

Proof. Recall that $\beta = \frac{C(X_i,Y_i)}{V(X_i)}$ and that OLS is the sample analog of this, $\hat{\beta}_n = s_{XY}/s_X^2$. By the Continuous Mapping Theorem,
$$plim\ \hat{\beta}_n = plim(s_{XY}/s_X^2) = plim(s_{XY})/plim(s_X^2).$$
Also,
$$s_{XY} = \frac{1}{n}\sum X_i Y_i - \bar{X}_n \bar{Y}_n,$$
so, plimming inside the covariance formula shows that
$$plim\, s_{XY} = plim\left[\frac{1}{n}\sum X_i Y_i\right] - plim\left[\bar{X}_n\right] plim\left[\bar{Y}_n\right] = E[X_i Y_i] - \mu_X \mu_Y = C(X, Y).$$
Likewise, $plim\, s_X^2 = V(X)$. So, we see that
$$plim(s_{XY}/s_X^2) = C(X, Y)/V(X) = \beta.$$

5.2 Asymptotic Distribution of the Sample Slope

1. *Convergence in Distribution*

The sequence of statistics S_n has asymptotic or limiting distribution F if:

$$Lim_{n\to\infty} P(S_n \leq c) = F(c) \qquad \text{for any constant } c \text{ for which } F \text{ is defined.}$$

When treating S_n as if it has distribution F, we are said to be "using an asymptotic approximation."

(a) *Central Limit Theorem*

The standardized sample mean from a random sample has an asymptotic Normal distribution. This means that:

$$Lim_{n\to\infty} P\left[\frac{\bar{X} - \mu_X}{s_X/\sqrt{n}} < c\right] = Lim_{n\to\infty} P\left[\frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} < c\right] = \Phi(c), \tag{4}$$

where Φ is the standard Normal CDF.

$\bar{X}_n \to N\left(\mu_X, \frac{\sigma^2}{n}\right)$

- Note that $\frac{\bar{X}-\mu_X}{s_X/\sqrt{n}}$ is our usual regression t-statistic
- The first equals sign in (4) means that replacing σ_X with the estimated std dev, s_X, doesn't (asymptotically) matter. Hence, what we've been calling a "$t-statistic$" is asymptotically standard Normal.
- The variance of a sample statistic's limiting distribution is referred to as its *asymptotic variance*. The asymptotic standard error is the square root of this, divided by $\sqrt{n}$. For a sample mean, the asymptotic SE is $\sigma_X/\sqrt{n}$, estimated by $s_X/\sqrt{n}$, the same as the finite-sample standard error.

(b) CLT for sample moments

In general, any SE-standardized sample moment has a limiting Normal distribution. In particular:

$$(Sample\ moment - Population\ moment)/(asymptotic\ SE\ of\ sample\ moment) \sim_a N(0,1)$$

(c) CLT for functions of moments

Suppose again that A_n and B_n are sample moments and that C_n is a continuous function of these moments:

$$C_n = h(A_n, B_n).$$

Then C_n has a limiting Normal distribution:

$$(C_n - plim\, C_n)/(asymptotic\ SE\ of\ C_n) \sim_a N(0,1)$$

This is a remarkably powerful result; it encapsulates most of the statistical theory you really need to know.

- Formulas for the asymptotic standard error of a function of sample moments are typically more complicated than formulas for the asymptotic standard errors of the underlying sample moments themselves. But we can look these up when needed (and Stata knows them).

Result 2

Consider regression model

$$Y_i = \alpha + \beta X_i + \varepsilon_i.$$

The OLS estimator $\hat{\beta}_n$ computed in a random sample of size n is approximately Normally distributed with probability limit β and asymptotic variance equal to:

$$AV(\hat{\beta}_n) = \frac{E[(X_i - \mu_X)^2 \varepsilon_i^2]}{(\sigma_X^2)^2}.$$

This in turn means that $\frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\sqrt{AV(\hat{\beta}_n)}} \sim_a N(0,1)$ $= \frac{\hat{\beta}_n - \beta}{\sqrt{AV(\hat{\beta}_n)}/\sqrt{n}}$

Proof. *plim* $\hat{\beta}_n = \beta$ is Result 1, above. Asymptotic normality is a consequence of the CLT for functions of sample moments. The formula for the asymptotic variance is derived in MHE 3.1.3.

$AV(\hat{\beta})$ is not a beautiful formula, but we can easily use it by substituting sample values for population values. It becomes an *asymptotic standard error* when we take the square root and divide by $\sqrt{n}$. Stata *estimates* the asymptotic standard error of a regression coefficient using the *robust standard error* formula:

$$\widehat{RSE}(\hat{\beta}_n) = \frac{1}{\sqrt{n}} \left(\frac{\frac{1}{n}\sum_i [(X_i - \bar{X})^2 e_i^2]}{(s_X^2)^2} \right)^{\frac{1}{2}} \tag{5}$$

This is $\frac{1}{\sqrt{n}}\sqrt{AV(\hat{\beta}_n)}$ after replacing expectations with sums, ε_i with the estimated resids, e_i, and replacing σ_X^2 with s_X^2. Stata reports this standard error when you use option "robust." MM notation differs slightly, using $RSE(\hat{\beta}_n)$ to denote population robust standard errors, without worrying about whether resids and σ_X^2 are estimated or known.

RSE for multivariate regression

When the slope coefficient of interest is $\hat{\beta}_j$, that is, the jth coefficient in a multivariate model, we get the relevant robust standard error by modifying (5) to be ... (see MM p. 97 for the answer)

Homoskedasticity in large samples

Result 3 When residuals are homoskedastic, i.e., $E[\varepsilon_i^2 \mid X_i] = \sigma_\varepsilon^2$, then $AV(\hat{\beta}_n) = \frac{\sigma_\varepsilon^2}{\sigma_X^2}$.

Proof: The general AV formula is $\frac{E[(X_i - \mu_X)^2 \varepsilon_i^2]}{(\sigma_X^2)^2}$

Using Law of Iterated Expectations $= \frac{E[E((X_i - \mu_X)^2 \varepsilon_i^2 \mid X_i)]}{(\sigma_X^2)^2} = \frac{E[(X_i - \mu_X)^2 E(\varepsilon_i^2 \mid X_i)]}{(\sigma_X^2)^2}$

$$= \frac{\sigma_\varepsilon^2 E[(X_i - \mu_X)^2]}{(\sigma_X^2)^2} \quad (\text{since } E[\varepsilon_i^2 \mid X] = \sigma_\varepsilon^2)$$

$$= \frac{\sigma_\varepsilon^2}{\sigma_X^2}.$$ (From Page 63)

$RSE = \frac{1}{\sqrt{n}}\sqrt{\frac{\sigma_\varepsilon^2}{\sigma_X^2}} = \frac{\sigma_\varepsilon}{\sigma_X \sqrt{n}}$

- Assuming homoskedastic resids, the RSE std error formula simplifies to the old-fashioned formula
- Ancient econometricians fretted much about heteroskedasticity. Modern masters have only one thing to say in the face of this: "robust"

5.3 Empirical SE Formulas Compared

```
    opened on:   5 Apr 2018, 10:00:06

 1 .
 2 .         // Merge housing and person record from 2016 ACS PUMS
 3 .         use pfile, clear

 4 .         merge m:1 serialno using hfile

       Result                           # of obs.
       -----------------------------------------
       not matched                       124,853
           from master                         0  (_merge==1)
           from using                    124,853  (_merge==2)

       matched                         3,156,487  (_merge==3)
       -----------------------------------------

 5 .         keep if _m==3
   (124,853 observations deleted)

 6 .         drop _m

 7 .
 8 .         // Sample of interest is all women aged 25-49
 9 .         keep if sex==2 & age>=25 & age <=49
   (2,684,757 observations deleted)

10 .
11 .         // Children
12 .         lab var noc "Number of own children"

13 .
14 .         // Employed variable
15 .         gen employed=(cow>=1 & cow<=7)

16 .         lab var employed  "Individual is employed"

17 .
18 .         // Weekly hours
19 .         replace wkhp=0 if employed==0
   (65,870 real changes made)

20 .         lab var wkhp "Usual hours per week"
```

```
. sum employed wkhp noc yearsEd age white
```

Variable	Obs	Mean	Std. Dev.	Min	Max
employed	471,730	.860365	.3466083	0	1
wkhp	435,760	32.05255	17.07636	0	99
noc	464,543	1.082944	1.219274	0	16
yearsEd	471,730	13.98169	3.078403	0	21
agep	471,730	37.15064	7.276517	25	49
white	471,730	.7336273	.4420619	0	1

```
46 .         // Hours worked on number of kids, old-fashioned vs. robust SEs
47 .         reg wkhp noc yearsEd age white
```

Source	SS	df	MS		
Model	11395392.4	4	2848848.1	Number of obs	= 429,843
Residual	111868309	429,838	260.256907	F(4, 429838)	= 10946.29
				Prob > F	= 0.0000
				R-squared	= 0.0924
				Adj R-squared	= 0.0924
Total	123263701	429,842	286.765139	Root MSE	= 16.132

wkhp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
noc	-2.274926	.0203607	-111.73	0.000	-2.314832	-2.23502
yearsEd	1.353838	.0080881	167.39	0.000	1.337986	1.369691
agep	-.0496699	.0033886	-14.66	0.000	-.0563114	-.0430284
white	.4837675	.0560483	8.63	0.000	.3739146	.5936205
_cons	17.14356	.1782207	96.19	0.000	16.79425	17.49286

```
48 .         reg wkhp noc yearsEd age white, r
```

Linear regression

Number of obs	= 429,843
F(4, 429838)	= 9620.16
Prob > F	= 0.0000
R-squared	= 0.0924
Root MSE	= 16.132

wkhp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
noc	-2.274926	.0218837	-103.96	0.000	-2.317817	-2.232035
yearsEd	1.353838	.0090033	150.37	0.000	1.336192	1.371484
agep	-.0496699	.0033541	-14.81	0.000	-.0562439	-.0430959
white	.4837675	.0578343	8.36	0.000	.370414	.5971211
_cons	17.14356	.1879141	91.23	0.000	16.77525	17.51186

- Robustness matters little in this case, but the fact that robust SEs are a little bigger than old-fashioned SEs improves confidence interval coverage

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 8

Residuals, Fitted Values, and Goodness of Fit

1 Regression Recap

Conceptual

Regression is a many-splendored thing: If $E[Y_i|X_i] = \alpha + \beta X_i$, then $\alpha = E[Y_i] - \beta E[X_i]$ and $\beta = C(X_i, Y_i)/V(X_i)$. If the CEF is nonlinear, the regression slope and intercept provide the best linear approximation to it (and the best linear predictor for Y_i). Regression estimates of treatment effects approximate what we'd get by matching on the values of regressors included as controls and then averaging these conditional effects. This note discusses further regression features – properties that hold for all regressions, regardless of your reason for running 'em.

Estimation Recap

We estimate the population slope and intercept using their sample analogs. With a single regressor, these OLS estimators are

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$
$$\hat{\beta} = s_{XY}/s_X^2$$

Multivariate regression models replace each regressor in the bivariate formula with the residuals remaining after partialing out all others. For example, the first regressor, X_{1i}, in a model with $k-1$ additional controls has sample slope

$$\hat{\beta}_1 = s_{\tilde{x}_1 Y}/s_{\tilde{x}_1}^2$$

where $\tilde{x}_{1i}$ is the residual from a regression of X_{1i} on $X_{2i}, ..., X_{ki}$. This *OLS estimator* minimizes the *residual sum of squares* (RSS),

$$RSS = \sum (Y_i - a - b_1 X_{1i} - ... - b_k X_{ki})^2$$

in your data.

Regression Inference Recap

As we've seen, under classical assumptions, OLS estimates are unbiased, Normally distributed, and BLUE. More generally, assuming only random sampling, OLS estimates are consistent and asymptotically Normally distributed. We use this to test hypotheses and construct confidence intervals for regression parameters under very weak assumptions.

- We've focused so far on interpreting regression coefficients and on the sampling distributions of corresponding OLS estimates.
- Regression models generate additional information as well. The ideas here are the same whether there's one regressor or many, so we'll do the math for bivariate regression only.

2 Regression Fission: Two Pieces of Y

2.1 Residuals

Population residuals are defined as

$$\varepsilon_i \equiv Y_i - \alpha - \beta X_i$$

By definition, these satisfy:

$$E(\varepsilon_i) = 0 \tag{1}$$
$$E(X_i \varepsilon_i) = 0 \tag{2}$$

- *Estimated residuals are*

  $$e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i,$$

  with properties analogous to (1) and (2) in the sample:

  $$\sum_{i=1}^{n} e_i = 0 \tag{3}$$
  $$\sum_{i=1}^{n} X_i e_i = 0 \tag{4}$$

  Proof : Substitute sample resids in sums,

  $$\sum e_i = \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = \sum (Y_i - \bar{Y}) - \hat{\beta} \sum (X_i - \bar{X}) = 0$$

  $$\sum X_i e_i = \sum X_i (Y_i - \bar{Y}) - \hat{\beta} \sum X_i (X_i - \bar{X}) = 0$$

  We've been using this already: these are the first-order conditions for OLS.

- Note again that we can't use (3) and (4) to "check" whether $E(\varepsilon_i) = 0$ and $E(X_i \varepsilon_i) = 0$: these properties hold in both population and sample

2.2 Fitted values

- Population *fitted values* are defined as:

  $$\hat{Y}_i^* = \alpha + \beta X_i$$

  So

  $$Y_i = \hat{Y}_i^* + \varepsilon_i$$

  This decomposes Y_i into the fitted value, said to be "explained by X_i" and the piece left over, ε_i.

Population fits and resids are uncorrelated:

$$E\left[\hat{Y}_i^* \varepsilon_i\right] = 0$$

Proof. Substitute for $\hat{Y}_i$:

$$E\left[\hat{Y}_i^* \varepsilon_i\right] = E\left[(\alpha + \beta X_i)\varepsilon_i\right] = \alpha \underbrace{E[\varepsilon_i]}_{0} + \beta \underbrace{E[X_i \varepsilon_i]}_{0}$$

Now use (1) and (2).

$\sum \hat{Y}_i e_i = \sum (\hat{\alpha} + \hat{\beta} X_i)(e_i) = \sum (\hat{\alpha} e_i + \hat{\beta} X_i e_i) = \sum \hat{\alpha} e_i + \sum \hat{\beta} X_i e_i = \hat{\alpha} \sum e_i + \hat{\beta} \sum X_i e_i = 0 + 0 = 0$

- *Estimated* fitted values are defined as
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i,$$
so
$$Y_i = \hat{Y}_i + e_i$$
- By virtue of (3) and (4), estimated resids and fits are uncorrelated, that is,
$$\sum \hat{Y}_i e_i = 0$$
in your data
  - MM Chpt 2 garbles the distinction between population fitted values and estimated fitted values - sorry!

3 R-squared

How much of the variance in Y_i can be attributed to variation in X_i? We have seen that resids and fits are uncorrelated. So
$$s_Y^2 = s_{\hat{Y}}^2 + s_e^2 = \hat{\beta}^2 s_X^2 + s_e^2 \tag{5}$$
in your sample. Likewise, in the population from which you're sampling
$$V(Y_i) = V(\hat{Y}_i^*) + V(\varepsilon_i).$$

Variance of sum = sum of variances when independent
$Y_i = \hat{Y}_i + e_i \Rightarrow s_Y^2 = s_{\hat{Y}}^2 + s_e^2$

Be sure you can show this.

- In both sample and pop, regression ANOVA holds:
Total variance = explained variance + residual variance
- It's customary to report the following as a measure of regression "goodness of fit":
$$R^2 \equiv \frac{s_{\hat{Y}}^2}{s_Y^2} = \frac{\hat{\beta}^2 s_X^2}{s_Y^2} = 1 - \frac{s_e^2}{s_Y^2}$$
uses (5)

$R^2 = \frac{\text{explained variance}}{\text{Total variance}}$

- This value, called R^2, is necessarily between 0 and 1
  - R-squared gives the fraction of variation in Y_i that is accounted for in a correlational sense by the regressor, X_i
  - R^2 is also equal to $\left[CORR(\hat{Y}_i, Y_i)\right]^2$, sometimes said to be the square of the coefficient of multiple correlation (show this)
- Down the road, we'll use R^2 to construct hypothesis tests that *compare* alternative multivariate regression models, such as short and long
- The R^2 from any *single* model is hard to interpret without a standard of comparison.
  - Life is random; stuff happens. How random, you ask? Hard to say.
- Our discussion of resids, fits, and R^2 carries over to multivariate models (e.g., R^2 is the proportion of dependent variable variance explained by all regressors in a multivariate model)

$CORR(\hat{Y}_i, Y_i) = \frac{Cov(\hat{Y}_i, Y_i)}{\sqrt{V(\hat{Y}_i)\,V(Y_i)}} = \frac{Cov(\hat{Y}_i + e_i, \hat{Y}_i)}{\sqrt{V(\hat{Y}_i + e_i)\,V(\hat{Y}_i)}} = \frac{V(\hat{Y}_i)}{\sqrt{(V(\hat{Y}_i) + V(e_i))\,V(\hat{Y}_i)}} = \frac{[V(\hat{Y}_i)]^{1/2}}{\sqrt{V(\hat{Y}_i) + V(e_i)}}$

$[CORR(\hat{Y}_i, Y_i)]^2 = \frac{V(\hat{Y}_i)}{V(\hat{Y}_i) + V(e_i)} = \frac{s_{\hat{Y}}^2}{s_{\hat{Y}}^2 + s_e^2} = \frac{s_{\hat{Y}}^2}{s_Y^2}$

Math from here same for population & sample

- See attached regression output, which shows that schooling is better than sex ... in an R^2 sense

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |     10054    31.99881    1.419198         30         34
     incwage |     10054    34664.13    39664.05          0     496759
         uwe |      8468    783.3483    530.3319   .1041667       4000
      loguwe |      8468     6.42724    .7603143  -2.261763   8.294049
     yearsEd |     10054    13.43973    2.521863          0         16
-------------+--------------------------------------------------------
       white |     10054           1           0          1          1
     working |     10054    .8490153    .3580518          0          1
      female |     10054    .5277501    .4992542          0          1

. // regress log weekly wage on education
. reg loguwe yearsEd, robust

Linear regression                                      Number of obs =    8468
                                                       F(  1,  8466) = 1022.00
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1135
                                                       Root MSE      =  .71591

------------------------------------------------------------------------------
             |               Robust
      loguwe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1075207   .0033633    31.97   0.000     .1009278    .1141136
       _cons |   4.965565   .0466539   106.43   0.000     4.874112    5.057018
------------------------------------------------------------------------------

. // regress log weekly wage on female
. reg loguwe female, robust

Linear regression                                      Number of obs =    8468
                                                       F(  1,  8466) =  494.69
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0561
                                                       Root MSE      =  .73873

------------------------------------------------------------------------------
             |               Robust
      loguwe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.3605673   .0162113   -22.24   0.000    -.3923455   -.3287891
       _cons |   6.597901   .0102161   645.83   0.000     6.577875    6.617927
------------------------------------------------------------------------------
```

3.1 Regression F-Statistics

In the classical bivariate regression model (Normal errors, etc), the statistic:

$$\frac{s_X^2 \hat{\beta}^2}{s_e^2/(n-2)} = (n-2)\frac{R^2}{1-R^2}$$

is distributed $F_{1,n-2}$ under H_0: $\beta = 0$

- An F with one d.f. in the numerator is the square of the t-statistic for the single regressor in a bivariate model, so there's no new information here (check this above).
- F tests shine when it comes time to compare multivariate regression models

$$\frac{S_x^2 \hat{\beta}^2}{S_e^2/(n-2)} = (n-2)\frac{S_x^2 \hat{\beta}^2}{S_e^2} = (n-2)\frac{R^2 S_y^2}{(1-R^2)S_y^2} = (n-2)\frac{R^2}{1-R^2}$$

$$+\sqrt{\frac{S_x^2 \hat{\beta}^2}{S_e^2/n-2}} = \frac{\hat{\beta} S_x}{S_e/\sqrt{n-2}} = \hat{\beta}\left(\frac{S_x}{S_e}\right)\sqrt{n-2}$$

$$T_n = \frac{\hat{\beta}-\beta_0}{S_e/(\sqrt{n}\,S_x)} = (\hat{\beta}-\beta_0)\left(\frac{S_x}{S_e}\right)\sqrt{n}$$

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 9

Modeling with Multivariate Regression Models

1 Saturated Regression Models and Interaction Terms

- LN5 introduces the idea of *saturated regression models* – let's expand on this
  - A saturated regression model has as many parameters as the corresponding CEF
  - When counting parameters, we include the intercept
- Consider a single multinomial regressor $X_i \in \{10, 11, 12, ...19\}$. The CEF $E[Y_i|X_i]$ assumes up to 10 distinct values. 10 distinct values → 10 distinct parameters
  - What's the corresponding saturated regression model look like? An intercept plus 9 dummies, one for each value of $X_i > 10$ the intercept would just be the value for $X=10$
  - Other 10-parameter schemes also work
- What about multiple regressors? Suppose X_{1i} and X_{2i} are dummies for college graduation status and sex (not that these ever go together):

$$X_{1i} = 1(colgrad_i = yes)$$
$$X_{2i} = 1(sex_i = female)$$

- The CEF of Y_i conditional on X_{1i}, X_{2i} can be written:

$$E[Y_i \mid X_{1i}, X_{2i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i}$$

How many values can this CEF assume?

- The corresponding saturated regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{1i} X_{2i} + \epsilon_i \tag{1}$$

- Think of saturated models as maximum flexibility or a perfect fit
  - Equation (1) allows the college wage effect to differ for men and women:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \epsilon_i && \text{when } X_{2i} = 0 \quad \text{(men)} \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 + \beta_{12} X_{1i} + \epsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) X_{1i} + \epsilon_i && \text{when } X_{2i} = 1 \quad \text{(women)} \end{aligned}$$

while also allowing the female effect to differ between those with and without a college degree::

$$\begin{aligned} Y_i &= \beta_0 + \beta_2 X_{2i} + \epsilon && \text{when } X_{1i} = 0 \\ Y_i &= \beta_0 + \beta_1 + \beta_2 X_{2i} + \beta_{12} X_{2i} + \epsilon_i \\ &= (\beta_0 + \beta_1) + (\beta_2 + \beta_{12}) X_{2i} + \epsilon && \text{when } X_{1i} = 1 \end{aligned}$$

β_0 is baseline for men
$\beta_0 + \beta_2$ is baseline for women
β_1 is effect of college for men
$\beta_1 + \beta_{12}$ is effect of college for women

- Regression talk
  - The coefficients on X_{1i} and X_{2i} are said to be *main effects*, while β_{12} is an *interaction term* or *2nd-order term.* Models with more regressors may have higher-order terms
  - Models without interactions are said to be *additive*. The additive version of (1) is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{2}$$

    Additive models can be viewed as either a restriction on, or an approximation to, the CEF
  - When choosing the details of a regression model (including or omitting interaction terms) we are said to be *specifying* or *parameterizing* it.
  - Critics of your model may object to your *specification* of it

2 Not Only for Dummies: Ordinal and Continuous Interactions

Interesting interactions arise in models with ordinal and continuous regressors.

- Let

$$S_i = years\ of\ schooling$$
$$X_{2i} = 1(sex_i = female)$$

- Repeat equation (1):

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 X_{2i} + \beta_{12} S_i X_{2i} + \epsilon_i \tag{3}$$

  - This model isn't saturated even though it includes an interaction term (why not?) Multiple values for years of schooling (could have used dummy for each value)
  - Regression model (3) therefore restricts or approximates the CEF (how?)

- With ordinal S_i and dummy X_{2i}, models like (3) are usually interpreted (asymmetrically) as allowing the regression of Y_i on S_i to differ by X_{2i}:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 S_i + \epsilon_i && \text{when } X_{2i} = 0 \\ Y_i &= \beta_0 + \beta_1 S_i + \beta_2 + \beta_{12} S_i + \epsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_{12}) S_i + \epsilon_i && \text{when } X_{i2} = 1 \end{aligned}$$

  - The (female-male) difference in intercepts is β_2
  - The (female-male) difference in the economic returns to schooling is β_{12}
  - Note that this interpretation works equally well for continuous S_i

- This parameterization produces results identical to those from separate regressions on S_i by sex
  - Why? Both the S_i slope and the model intercept are free to vary with X_{2i}, which takes on two values
  - We say that model (3) is *fully interacted with* X_{2i} (though it's not saturated)
- A test of whether the regression on schooling is the same for men and women is a joint test of:

$$H_0 : \beta_2 = \beta_{12} = 0 \tag{4}$$

Multiple restrictions (how many?) of this sort are tested with an F statistic

3 Testing Linear Restrictions

- We're often interested in testing sets of linear restrictions. A set of q linearly independent restrictions applied to a regression model with k coefficients can be written:

$$c_{11}\beta_1 + c_{12}\beta_2 + c_{13}\beta_3 + ... + c_{1k}\beta_k = c_{10}$$
$$\vdots$$
$$c_{q1}\beta_1 + c_{q2}\beta_2 + c_{q3}\beta_3 + ... + c_{qk}\beta_k = c_{qo},$$

  where the c's are constants.

  - Hypothesis (4) contains two restrictions:

$$\beta_2 = 0$$
$$\beta_{12} = 0$$

    simple! $(c_{11} = c_{21} = 1; c_{10} = c_{20} = 0)$

[r means restricted, ur means unrestricted]

$q = df_r - df_{ur}$

$$F = \frac{(RSS_r - RSS_{ur})/(df_r - df_{ur})}{RSS_{ur}/df_{ur}}$$

$df_{ur} = n-k-1$

- Let e_{Ri} be the residuals from the model that imposes these restrictions and let e_{Ui} be the residuals from the model that doesn't impose them. Estimated residual variances from these models are $s_R^2 = \frac{1}{n-k-1}\sum e_{Ri}^2$ and $s_U^2 = \frac{1}{n-k-1}\sum e_{Ui}^2$. Given classical regression modeling assumptions (linear CEF, fixed X's, homo Normal independent resids), we have that

$$F = \frac{(\sum e_{Ri}^2 - \sum e_{Ui}^2)/q}{\sum e_{Ui}^2/(n-k-1)} = (\frac{n-k-1}{q}) \times \frac{(s_R^2 - s_U^2)}{s_U^2} \sim F_{q,\, n-k-1},$$

  under the null hypothesis that the restrictions are satisfied

  - This *F statistic* is used to do an *F test*
  - The F distribution has two parameters: *numerator degrees of freedom (df)* equal to the number of restrictions being tested and *denominator df* equal to the sample size minus the number of parameters in the unrestricted model
  - As usual, we test a null by looking for surprising draws from the null distribution, fixed the probability of Type I error
  - The F-stat's two df determine critical values (numerator *df* matter most; denominator *df* can often be taken to be infinite)

- We often test *zero restrictions* where $c_{10=...} = c_{q0} = 0$ and all other c's are equal to either 1 or 0.

R^2 version of the F-test

- Write $R_R^2 = 1 - (s_R^2/s_Y^2)$ and $R_U^2 = 1 - (s_U^2/s_Y^2)$ for restricted and unrestricted R^2, respectively. Provided restrictions don't transform the dependent variable (thereby changing its variance), we can write:

$$F = (\frac{n-k-1}{q}) \times \frac{(s_R^2 - s_U^2)}{s_U^2} = (\frac{n-k-1}{q}) \times \frac{(s_R^2/s_Y^2 - s_U^2/s_Y^2)}{s_U^2/s_Y^2}$$
$$= \left(\frac{n-k-1}{q}\right) \times \frac{(R_U^2 - R_R^2)}{(1 - R_U^2)}$$

- An F-test gauges the extent to which restrictions reduce R^2 (as they must) by asking whether the change is small enough to be put down to sampling variance under the null

3.1 Restriction Examples

- Testing additivity

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 X_{2i} + \beta_3 X_{3i} + \underbrace{\beta_{12} S_i X_{2i} + \beta_{13} S_i X_{3i} + \beta_{23} X_{2i} X_{3i}}_{2nd\ order\ terms} + \underbrace{\beta_{123} S_i X_{2i} X_{3i}}_{3rd\ order\ term} + \varepsilon_i \quad (UR)$$

$$Y_i = \beta_0 + \beta_1 S_i + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i \quad (R)$$

In this case, a no-interactions null imposes four restrictions:

$$H_0 : \beta_{12} = \beta_{13} = \beta_{23} = \beta_{123} = 0$$

- Testing constant returns to scale (CRTS) in Cobb-Douglas production. Start with:

If sum of β's is 1, we have CRTS

$$Y_i = A X_{1i}^{\beta_1} X_{2i}^{\beta_2} X_{3i}^{\beta_3} \eta_i$$

Logging:

$$\ln Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \varepsilon_i \quad (UR)$$

CRTS says

$$H_0 : \beta_1 + \beta_2 + \beta_3 = 1$$

This is a single restriction on an unrestricted model with 4 parameters, so the relevant null distribution is $F_{1,\,N-4}$

  - The restricted model can be rewritten:

$\beta_1 = 1 - (\beta_2 + \beta_3)$ can substitute this in to get 1 fewer parameter

$$Y_i = \beta_0 + (1 - \beta_2 - \beta_3) X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$$Y_i - X_{1i} = \beta_0 + \beta_2 (X_{2i} - X_{1i}) + \beta_3 (X_{3i} - X_{1i}) + \varepsilon_{ii}$$

$$Y_i^* = \beta_0 + \beta_2 X_{2i}^* + \beta_3 X_{3i}^* + \epsilon_i \quad (R)$$

  - F can't be constructed from R_U^2 and R_R^2 since the restricted model modifies the dependent variable. But Stata's test command get it right.

4 Testing Gender and Race Interactions in the Returns to Military Service (attached)

5 Krueger (1993): Computer Interactions in Action (attached)

```
clear
set more off
```

Data from the 2008 IPUMS-CPS http://cps.ipums.org/cps/

```
use cps_00013.dta
```

Generate usual weekly and hourly earnings

```
gen uwe =  incwage / WKSWORK1
label var uwe "Usual weekly earnings"
gen loguwe = log(uwe)
label var loguwe "log(usual weekly earnings)"
gen logahe = log(incwage/(WKSWORK1*uhrswork))
label var logahe "Log usual hourly earnings"
gen logearn = log(incwage)
label var logearn "log(annual wage and salary earnings)"
```

```
(99,260 missing values generated)
(105,868 missing values generated)
(105,868 missing values generated)
(56,097 missing values generated)
```

Generate approximate years of education

```
gen yearsEd = .
replace yearsEd=0 if educ==2
// additional labeling and output omitted from logs
label var yearsEd "Years of education (approximate)"
gen colgrad = yearsEd>=16

gen age2 = age*age
gen potex = age-yearsEd-6
label var potex "Potential experience"
gen potex2 = potex*potex
```

Code vet, mil, race, sex

```
gen military = (popstat==2)
gen veteran = (vetstat==2) | (military==1)
```

black or part black

```
gen black = (race==200 | race==801 | race==805 | race==806 | race==807)
gen blackEd = black*yearsEd
gen blackvet = black*veteran

gen female=(sex==2)
gen femEd = female*yearsEd
gen femvet = female*veteran

gen blackfem=female*black
```

```
(49,771 missing values generated)
```

keep men and women aged 30–49, include active duty, sample=inLF

```
keep if age>=30 & age<50
```

```
(146,432 observations deleted)
```

add LFvars

```
gen working = WKSWORK1>0
```

Summary Stats

```
sum age yearsEd female black veteran femvet blackvet military ///
        colgrad incwage WKSWORK1 uhrswork logahe loguwe working
```

```
sum age yearsEd female black veteran femvet blackvet military colgrad incwage
 WKSWORK1 uhrswork logahe loguwe working
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	59,972	39.88245	5.681624	30	49
yearsEd	59,972	13.7431	2.900316	0	21
female	59,972	.5254786	.4993546	0	1
black	59,972	.1104015	.3133922	0	1
veteran	59,972	.0612619	.2398121	0	1
femvet	59,972	.0090042	.0944631	0	1
blackvet	59,972	.0094711	.0968584	0	1
military	59,972	.0067698	.0820007	0	1
colgrad	59,972	.3235843	.4678474	0	1
incwage	59,972	38108.07	48959.97	0	688117
WKSWORK1	59,972	41.47669	19.27549	0	52
uhrswork	59,972	34.92955	17.70783	0	99
logahe	48,009	2.879673	.730158	-5.897154	9.069273
loguwe	48,009	6.549787	.8281364	-2.261763	11.15625
working	59,972	.852131	.354973	0	1

Additive model

```
reg logahe yearsEd veteran black female
```

Source	SS	df	MS	Number of obs	=	48,009
				F(4, 48004)	=	3026.50
Model	5154.67766	4	1288.66941	Prob > F	=	0.0000
Residual	20439.8621	48,004	.425794977	R-squared	=	0.2014
				Adj R-squared	=	0.2013
Total	25594.5397	48,008	.533130723	Root MSE	=	.65253

logahe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yearsEd	.104886	.0010596	98.98	0.000	.1028091	.1069629
veteran	.0367164	.0121419	3.02	0.002	.012918	.0605147
black	-.0988435	.009582	-10.32	0.000	-.1176244	-.0800627
female	-.2972256	.0060911	-48.80	0.000	-.3091643	-.2852869
_cons	1.570199	.0152843	102.73	0.000	1.540242	1.600157

Regression of wages on eduation and veteran status, by sex, control for race

```
bys female: reg logahe yearsEd veteran black
```

```
-------------------------------------------------------------------------------
-> female = 0

      Source |       SS           df       MS      Number of obs   =    24,552
-------------+----------------------------------   F(3, 24548)     =   1844.86
       Model |  2348.76452         3  782.921508   Prob > F        =    0.0000
    Residual |  10417.6685    24,548  .424379524   R-squared       =    0.1840
-------------+----------------------------------   Adj R-squared   =    0.1839
       Total |  12766.4331    24,551  .519996459   Root MSE        =    .65144

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1024476   .0014073    72.80   0.000     .0996893    .1052059
     veteran |   .0348462   .0131176     2.66   0.008     .0091349    .0605575
       black |  -.1908072   .0143528   -13.29   0.000    -.2189395   -.1626749
       _cons |    1.61261   .0199472    80.84   0.000     1.573512    1.651707
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> female = 1

      Source |       SS           df       MS      Number of obs   =    23,457
-------------+----------------------------------   F(3, 23453)     =   1523.13
       Model |  1945.86697         3  648.622324   Prob > F        =    0.0000
    Residual |  9987.43719    23,453  .425849025   R-squared       =    0.1631
-------------+----------------------------------   Adj R-squared   =    0.1630
       Total |  11933.3042    23,456  .508752735   Root MSE        =    .65257

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1081747   .0016071    67.31   0.000     .1050246    .1113248
     veteran |   .0681413   .0318518     2.14   0.032     .0057097    .1305729
       black |  -.0251327   .0128612    -1.95   0.051    -.0503416    .0000761
       _cons |   1.216732   .0231711    52.51   0.000     1.171315    1.262149
------------------------------------------------------------------------------
```

Regression of wages on eduation and veteran status interacted with sex

```
reg logahe yearsEd femEd veteran femvet black female blackfem
```

```
      Source |       SS           df       MS      Number of obs   =    48,009
-------------+----------------------------------   F(7, 48001)     =   1743.95
       Model |  5189.43401         7  741.347715   Prob > F        =    0.0000
    Residual |  20405.1057    48,001  .425097513   R-squared       =    0.2028
-------------+----------------------------------   Adj R-squared   =    0.2026
       Total |  25594.5397    48,008  .533130723   Root MSE        =      .652

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1024476   .0014084    72.74   0.000      .099687    .1052082
       femEd |   .0057271   .0021359     2.68   0.007     .0015407    .0099135
     veteran |   .0348462   .0131287     2.65   0.008     .0091138    .0605786
      femvet |   .0332951   .0344254     0.97   0.333    -.0341792    .1007694
       black |  -.1908072   .0143649   -13.28   0.000    -.2189626   -.1626518
      female |  -.3958777   .0305699   -12.95   0.000    -.4557951   -.3359604
    blackfem |   .1656745   .0192735     8.60   0.000     .1278981    .2034508
       _cons |    1.61261   .0199641    80.78   0.000      1.57348     1.65174
------------------------------------------------------------------------------
```

Test interactions jointly

```
test femEd femvet
test femEd femvet blackfem
lincom black+blackfem
```

```
 ( 1)  femEd = 0
 ( 2)  femvet = 0

       F(  2, 48001) =    4.14
            Prob > F =    0.0160

 ( 1)  femEd = 0
 ( 2)  femvet = 0
 ( 3)  blackfem = 0

       F(  3, 48001) =   27.25
            Prob > F =    0.0000

 ( 1)  black + blackfem = 0

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0251327   .0128499    -1.96   0.050    -.0503186    .0000532
------------------------------------------------------------------------------
```

Regression of wages on education and veteran status, by race, control for sex

```
bys black: reg logahe yearsEd veteran female
```

```
-------------------------------------------------------------------------------
-> black = 0

      Source |       SS           df       MS      Number of obs   =    42,771
-------------+----------------------------------   F(3, 42767)     =   3644.61
       Model |  4732.42079         3   1577.4736   Prob > F        =    0.0000
    Residual |  18510.5762    42,767  .432823818   R-squared       =    0.2036
-------------+----------------------------------   Adj R-squared   =    0.2036
       Total |   23242.997    42,770  .543441595   Root MSE        =    .65789

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1048038     .00111    94.41   0.000     .1026281    .1069795
     veteran |   .0245867   .0132534     1.86   0.064    -.0013903    .0505637
      female |  -.3160608   .0064956   -48.66   0.000    -.3287923   -.3033293
       _cons |   1.581149   .0159949    98.85   0.000     1.549799      1.6125
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> black = 1

      Source |       SS           df       MS      Number of obs   =     5,238
-------------+----------------------------------   F(3, 5234)      =    325.53
       Model |  353.444082         3  117.814694   Prob > F        =    0.0000
    Residual |  1894.24545     5,234  .361911627   R-squared       =    0.1572
-------------+----------------------------------   Adj R-squared   =    0.1568
       Total |  2247.68954     5,237  .429194107   Root MSE        =    .60159

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1066224   .0035997    29.62   0.000     .0995655    .1136792
     veteran |   .1253284   .0292934     4.28   0.000     .0679011    .1827557
      female |  -.1386085   .0172364    -8.04   0.000    -.1723989    -.104818
       _cons |   1.349591   .0506443    26.65   0.000     1.250307    1.448875
------------------------------------------------------------------------------
```

Regression of wages on education and veteran status interacted with race

```
reg logahe yearsEd blackEd veteran blackvet black female blackfem
```

```
      Source |       SS           df       MS      Number of obs   =    48,009
-------------+----------------------------------   F(7, 48001)     =   1744.07
       Model |  5189.71807         7  741.388295   Prob > F        =    0.0000
    Residual |  20404.8217    48,001  .425091595   R-squared       =    0.2028
-------------+----------------------------------   Adj R-squared   =    0.2027
       Total |  25594.5397    48,008  .533130723   Root MSE        =    .65199

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     yearsEd |   .1048038   .0011001    95.27   0.000     .1026476      .10696
     blackEd |   .0018186   .0040534     0.45   0.654    -.0061261     .0097633
     veteran |   .0245867   .0131345     1.87   0.061    -.0011571     .0503306
    blackvet |   .1007417   .0343572     2.93   0.003      .033401     .1680823
       black |  -.2315583   .0571302    -4.05   0.000    -.3435343    -.1195824
      female |  -.3160608   .0064373   -49.10   0.000    -.3286781    -.3034436
    blackfem |   .1774524   .0197584     8.98   0.000     .1387256     .2161792
       _cons |   1.581149   .0158514    99.75   0.000     1.550081     1.612218
------------------------------------------------------------------------------
```

Test interactions jointly

```
test blackEd blackvet
test blackEd blackvet blackfem
lincom black+blackfem
```

```
 ( 1)  blackEd = 0
 ( 2)  blackvet = 0

       F(  2, 48001) =    4.47
            Prob > F =    0.0114

 ( 1)  blackEd = 0
 ( 2)  blackvet = 0
 ( 3)  blackfem = 0

       F(  3, 48001) =   27.48
            Prob > F =    0.0000

 ( 1)  black + blackfem = 0

------------------------------------------------------------------------------
      logahe |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   -.054106   .0574678    -0.94   0.346    -.1667436     .0585316
------------------------------------------------------------------------------
```

TABLE I
PERCENT OF WORKERS IN VARIOUS CATEGORIES WHO DIRECTLY USE A COMPUTER AT WORK

Group	1984	1989
All workers	24.6	37.4
Gender		
Men	21.2	32.3
Women	29.0	43.4
Education		
Less than high school	5.0	7.8
High school	19.3	29.3
Some college	30.6	45.3
College	41.6	58.2
Postcollege	42.8	59.7
Race		
White	25.3	38.5
Black	19.4	27.7
Age		
Age 18–24	19.7	29.4
Age 25–39	29.2	41.5
Age 40–54	23.6	39.1
Age 55–65	16.9	26.3
Occupation		
Blue-collar	7.1	11.6
White-collar	33.0	48.4
Union status		
Union member	20.2	32.5
Nonunion	28.0	41.1
Hours		
Part-time	23.7	36.3
Full-time	28.9	42.7
Region		
Northeast	25.5	38.0
Midwest	23.4	36.0
South	23.2	36.5
West	27.0	39.9

Source. Author's tabulations of the 1984 and 1989 October Current Population Surveys. The sample size is 61,712 for 1984 and 62,748 for 1989.

TABLE II

OLS REGRESSION ESTIMATES OF THE EFFECT OF COMPUTER USE ON PAY

(DEPENDENT VARIABLE: ln (HOURLY WAGE))

Independent variable	October 1984			October 1989		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	1.937	0.750	0.928	2.086	0.905	1.094
	(0.005)	(0.023)	(0.026)	(0.006)	(0.024)	(0.026)
Uses computer at work (1 = yes)	0.276	0.170	0.140	0.325	0.188	0.162
	(0.010)	(0.008)	(0.008)	(0.009)	(0.008)	(0.008)
Years of education	—	0.069	0.048	—	0.075	0.055
		(0.001)	(0.002)		(0.002)	(0.002)
Experience	—	0.027	0.025	—	0.027	0.025
		(0.001)	(0.001)		(0.001)	(0.001)
Experience-squared ÷ 100	—	−0.041	−0.040	—	−0.041	−0.040
		(0.002)	(0.002)		(0.002)	(0.002)
Black (1 = yes)	—	−0.098	−0.066	—	−0.121	−0.092
		(0.013)	(0.012)		(0.013)	(0.012)
Other race (1 = yes)	—	−0.105	−0.079	—	−0.029	−0.015
		(0.020)	(0.019)		(0.020)	(0.020)
Part-time (1 = yes)	—	−0.256	−0.216	—	−0.221	−0.183
		(0.010)	(0.010)		(0.010)	(0.010)
Lives in SMSA (1 = yes)	—	0.111	0.105	—	0.138	0.130
		(0.007)	(0.007)		(0.007)	(0.007)
Veteran (1 = yes)	—	0.038	0.041	—	0.025	0.031
		(0.011)	(0.011)		(0.012)	(0.011)
Female (1 = yes)	—	−0.162	−0.135	—	−0.172	−0.151
		(0.012)	(0.012)		(0.012)	(0.012)
Married (1 = yes)	—	0.156	0.129	—	0.159	0.143
		(0.011)	(0.011)		(0.011)	(0.011)
Married*Female	—	−0.168	−0.151	—	−0.141	−0.131
		(0.015)	(0.015)		(0.015)	(0.015)
Union member (1 = yes)	—	0.181	0.194	—	0.182	0.189
		(0.009)	(0.009)		(0.010)	(0.010)
8 Occupation dummies	No	No	Yes	No	No	Yes
R^2	0.051	0.446	0.491	0.082	0.451	0.486

Notes. Standard errors are shown in parentheses. Sample size is 13,335 for 1984 and 13,379 for 1989. Columns (2), (3), (5), and (6) also include three region dummy variables.

TABLE III

THE RETURN TO VARIOUS USES OF COMPUTERS, OCTOBER 1989[a]

(DEPENDENT VARIABLE: ln (HOURLY WAGE))

Use of computer at work	Proportion	Coefficient (std. error)
Uses computer at work for any task[b]	0.398	0.145
		(0.010)
Specific Task[c]		
Word processing	0.165	0.017
		(0.012)
Bookkeeping	0.100	−0.058
		(0.013)
Computer-assisted design	0.039	0.026
		(0.020)
Electronic mail	0.063	0.149
		(0.016)
Inventory control	0.102	−0.056
		(0.013)
Programming	0.077	0.052
		(0.031)
Desktop publishing or newsletters	0.036	−0.047
		(0.021)
Spread sheets	0.094	0.079
		(0.015)
Sales	0.060	−0.002
		(0.016)
Computer games	0.019	−0.109
		(0.026)
R^2		0.495

a. The sample and other explanatory variables are the same as in column (6) of Table II.

b. The computer use dummy variable equals one if the worker uses computers for any of the ten enumerated tasks or for any other task.

c. The dummy variables for any specific computer task, and the dummy variable for any computer use, are not mutually exclusive.

TABLE IV

THE RETURN TO COMPUTER USE AT WORK, HOME, AND WORK AND HOME

(STANDARD ERRORS ARE SHOWN IN PARENTHESES.)

Type of computer use	October 1984 (1)	October 1989 (2)	Percent of sample, 1989 (3)
Uses computer at work	0.165	0.177	39.8
	(0.009)	(0.009)	
Uses computer at home	0.056	0.070	12.5
	(0.021)	(0.019)	
Uses computer at home and work	0.006	0.017	8.6
	(0.029)	(0.023)	
Sample size	13,335	13,379	

Notes. The table reports coefficients for three dummy variables estimated from log hourly wage regressions. The other explanatory variables in the regressions are education, experience and its square, two race dummies, three region dummies, dummy variables indicating part-time status, residence in an SMSA, veteran status, gender, marital status, union membership, and an interaction between marital status and gender. Covariates are the same as in columns (2) and (5) of Table II.

TABLE VII

OLS REGRESSION ESTIMATES OF THE EFFECT OF COMPUTER USE ON PAY

(DEPENDENT VARIABLE: ln (HOURLY WAGE))

Independent variable	October 1984			October 1989		
	(1)	(2)	(3)	(4)	(5)	(6)
Uses computer at work (1 = yes)	—	0.170	0.073	—	0.188	0.005
		(0.008)	(0.048)		(0.008)	(0.043)
Computer use*Education	—	—	0.007	—	—	0.013
			(0.003)			(0.003)
Years of education	0.076	0.069	0.067	0.086	0.075	0.071
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
Experience	0.027	0.027	0.027	0.027	0.027	0.027
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Experience-squared ÷ 100	−0.042	−0.041	−0.042	−0.044	−0.041	−0.042
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Black (1 = yes)	−0.106	−0.098	−0.099	−0.141	−0.121	−0.122
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Other race (1 = yes)	−0.120	−0.105	−0.106	−0.037	−0.029	−0.032
	(0.020)	(0.020)	(0.020)	(0.021)	(0.020)	(0.020)
Part-time (1 = yes)	−0.287	−0.256	−0.256	−0.261	−0.221	−0.221
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
Lives in SMSA (1 = yes)	0.123	0.111	0.111	0.148	0.138	0.138
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
Veteran (1 = yes)	0.043	0.038	0.039	0.027	0.025	0.029
	(0.011)	(0.011)	(0.011)	(0.012)	(0.012)	(0.012)
Female (1 = yes)	−0.140	−0.162	−0.160	−0.142	−0.172	−0.168
	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Married (1 = yes)	0.162	0.156	0.156	0.169	0.159	0.158
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Married*Female	−0.171	−0.168	−0.168	−0.146	−0.141	−0.139
	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)
Union member (1 = yes)	0.167	0.181	0.181	0.164	0.182	0.182
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)	(0.010)
R^2	0.429	0.446	0.446	0.428	0.451	0.452
Mean-squared error	0.168	0.163	0.163	0.176	0.169	0.169

Notes. Standard errors are shown in parentheses. Sample size is 13,335 for 1984 and 13,379 for 1989. Regressions also include three region dummy variables and an intercept.

TABLE I
PERCENT OF WORKERS IN VARIOUS CATEGORIES WHO USE DIFFERENT TOOLS ON THEIR JOB

Group	U. S. 1984	U. S. 1989	U. S. 1993	Germany 1979	Germany 1985–1986	Germany 1991–1992
	Percentage that are computer users					
All workers	25.1	37.4	46.6	8.5	18.5	35.3
Men	21.6	32.2	41.1	7.9	18.5	36.4
Women	29.6	43.8	53.2	9.7	18.5	33.5
Less than high school	5.1	7.7	10.4	3.2	4.3	9.9
High school	19.2	28.4	34.6	8.5	18.3	32.7
Some college	30.6	45.0	53.1	8.5	24.8	48.4
College	42.4	58.8	70.2	13.4	30.5	61.6
Age 18–24	20.5	29.6	34.3	10.1	13.8	27.8
Age 25–39	29.6	41.4	49.8	9.6	21.6	39.9
Age 40–54	23.9	38.9	50.0	6.6	17.2	35.9
Age 55–64	17.7	27.0	37.3	5.9	13.5	23.7
Blue-collar	7.1	11.2	56.6	1.2	3.5	10.7
White-collar	39.7	56.6	67.6	12.8	28.9	50.2
Part-time	14.8	24.4	29.3	6.4	14.7	26.5
Full-time	29.3	42.3	51.0	8.7	19.1	37.0
	Percentage of all workers who use a specific tool					
Computer	25.1	37.4	46.6	8.5	18.5	35.3
Calculator				19.6	35.7	44.2
Telephone				41.8	43.7	58.4
Pen/pencil				54.9	53.4	65.6
Work while sitting[a]				30.8	19.3	—
Hand tool (e.g., hammer)				29.4	32.9	30.5
Number of obs.	61,704	62,748	59,852	19,427	22,353	20,042

a. Variable definition differs in 1979 and 1985–1986. In 1979 it refers to "Never or rarely standing," and in 1985–1986 it refers to "Often or almost always sitting."

Columns 1 to 3 are from Table 3 in Autor, Katz, and Krueger [1996] and come from the October *Current Population Survey*. German data are from the *Qualification and Career Survey*.

resulting approximation should be rather good because of the large number of brackets. Adopting a similar specification, we find that this earnings variable yields the same return to schooling in 1985–1986 as reported by Krueger and Pischke [1995] with a continuous earnings variable for 1988. Years of education are imputed from information on schools attended and degrees obtained following Krueger and Pischke [1995]. When we bracket the earnings variable in the October 1984 CPS to be comparable with our 1979 data, in a regression similar to Krueger's we find a computer coefficient of 0.1697 using the original wage variable, and 0.1701 using the bracketed variable. Standard errors are about 3 percent larger with the bracketed variable.

TABLE II

OLS REGRESSIONS FOR THE EFFECT OF COMPUTER USE ON PAY

DEPENDENT VARIABLE: LOG HOURLY WAGE

(STANDARD ERRORS IN PARENTHESES)

Independent variable	U. S. 1984	U. S. 1989	U. S. 1993	Germany 1979	Germany 1985–1986	Germany 1991–1992
Computer	0.171	0.188	0.204	0.112	0.157	0.171
	(0.008)	(0.008)	(0.008)	(0.010)	(0.007)	(0.006)
Years of schooling	0.068	0.075	0.081	0.073	0.063	0.072
	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Experience	0.028	0.028	0.026	0.030	0.035	0.030
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Experience^2/100	−0.043	−0.043	−0.041	−0.052	−0.058	−0.046
	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
R^2	0.444	0.448	0.424	0.267	0.280	0.336
Number of obs.	13,335	13,379	13,305	19,427	22,353	20,042

Columns 1 to 3 are from Table 4 in Autor, Katz, and Krueger [1996]. Data for columns 1 to 3 are from the October *Current Population Survey;* data for columns 4 to 6 are from the *Qualification and Career Survey.* All models also include an intercept, a dummy for part-time, large city/SMSA status, female, married, female*married. Regressions for the United States in columns 1 to 3 also include dummies for black, other race, veteran status, union membership, and three regions. Regressions for Germany in columns 4 to 6 also include a dummy for civil servants (*Beamter*).

TABLE III

OLS REGRESSION FOR THE EFFECT OF DIFFERENT TOOLS ON PAY

DEPENDENT VARIABLE: LOG HOURLY WAGE

(STANDARD ERRORS IN PARENTHESES)

Independent variable	Germany 1979	Germany 1985–86	Germany 1991–92	Germany 1979	Germany 1979	Germany 1985–1986	Germany 1991–1992
Occupation indicators	No	No	No	501	501	742	1071
Grades and father's Occupation[a]	No	No	No	No	Yes	No	No
				Tools entered separately			
Computer	0.112	0.157	0.171	0.025	0.022	0.076	0.083
	(0.010)	(0.007)	(0.006)	(0.011)	(0.011)	(0.008)	(0.007)
Calculator	0.087	0.128	0.129	0.027	0.025	0.061	0.054
	(0.007)	(0.006)	(0.006)	(0.008)	(0.008)	(0.007)	(0.006)
Telephone	0.131	0.114	0.136	0.060	0.057	0.059	0.072
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Pen/pencil	0.123	0.112	0.127	0.055	0.052	0.055	0.050
	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)	(0.007)
Work while sitting	0.106	0.101	—	0.042	0.041	0.036	—
	(0.006)	(0.007)		(0.008)	(0.008)	(0.008)	
Hand tool	−0.117	−0.086	−0.091	−0.048	−0.045	−0.020	−0.020
(e.g., hammer)	(0.007)	(0.006)	(0.006)	(0.009)	(0.009)	(0.008)	(0.008)

	Tools entered together						
Computer	0.066	0.105	0.126	0.027	0.024	0.067	0.069
	(0.010)	(0.008)	(0.007)	(0.011)	(0.011)	(0.008)	(0.007)
Calculator	0.017	0.053	0.044	0.015	0.014	0.032	0.022
	(0.008)	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)	(0.007)
Telephone	0.072	0.043	0.045	0.043	0.041	0.035	0.048
	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Pen/pencil	0.062	0.031	0.035	0.040	0.038	0.024	0.007
	(0.007)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Work while sitting	0.058	0.050	—	0.036	0.035	0.032	—
	(0.007)	(0.007)		(0.008)	(0.008)	(0.008)	

a. Two variables for self-reported grades in math and German and eleven dummy variables for father's education.

Data are from the *Qualification and Career Survey*. All regressions also include an intercept, years of schooling, experience and experience squared, dummies for part-time, city, female, married, married*female, and for civil servants (*Beamter*).

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 10

Standard Standard Error Issues

LN7 outlines the tools of regression inference for the carefree world of random samples. We elaborate here on two complications, both arising in samples in which observations are dependent. The first complication is *serial correlation;* the second is *clustering*. Serial correlation arises most often in time series data, when one observation predicts the next. For example, the unemployment rate this quarter is likely to be similar to the unemployment rate last quarter. Clustering arises when when groups of observations are similar in identifiable groups, but are otherwise independent. In a randomized trial involving school-age children, for example, those in the same school or classroom are likely to have similar outcomes.

1 Serial Correlation in Time Series

Time series econometrics is a world unto itself. We visit this world only briefly (Stock and Watson pay a much longer visit). A few big-picture lessons:

- Dependent observations carry less information than independent observations. We should allow for this information downgrade when undertaking statistical inference
- The practical fix-ups for serial correlation are simple, but can be extraordinarily consequential
- Our fix-ups work only in large samples: we must therefore embrace an asymptopian vision of econometric inference

1.1 Defining Serial Correlation

- Consider a model relating US quarterly GDP growth to the federal funds rate, an important tool of monetary policy. For this time series regression, we write:

$$Y_t = \alpha + \beta X_t + \varepsilon_t \,;\, t = 1, \ldots, T \tag{1}$$

- In this case, it's likely that

$$C(\varepsilon_t, \varepsilon_s) = E[\varepsilon_t \varepsilon_s] \neq 0 \text{ for } s \neq t.$$

  In particular, we expect $E[\varepsilon_t \varepsilon_s] > 0$, that is, positive *serial correlation*
- Most economic time series are positively serially correlated. This reflects the fact that macro variables like unemployment rates, GDP levels and growth, financial variables, aggregate consumption, interest rates, public policy variables like government spending and interest rates are highly *persistent.*

Consequences of serial correlation

- When residuals are positively serially correlated, conventional regression SEs are probably too small. Robust SEs do not fix this.
- Even if classical assumptions (like homoskedasticity) are otherwise satisfied, OLS isn't BLUE
- OLS estimates are consistent and may still be unbiased if regressors are non-stochastic or the CEF is linear

1.2 Serial Correlation Fix-ups

Generalized Least Squares

- Serial correlation is often described using *autoregressive models*. The simplest is an $AR(1)$:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t;\ -1 < \rho < 1, \tag{2}$$

The error term is a linear function of the previous time period's error term

where the error term in this equation satisfies:

$\nu_t = \varepsilon_t - \rho\varepsilon_{t-1}$

  - $E[\nu_t] = 0$ for all t (not a restriction since $E[\varepsilon_t] = 0$)
  - $E[\nu_t^2] = \sigma_\nu^2$ for all t (homoskedasticity)
  - $E[\nu_t\nu_s] = 0$ for any $s \neq t$ (serially uncorrelated leftovers)

- We require $|\rho| < 1$ so that the time series error process is *stationary* (the variance of a non-stationary process is infinite)
  - Because economic data are persistent, we expect $\rho \geq 0$

- Write the model of interest and ρ-times-the-lagged-model:

$$Y_t = \alpha + \beta X_t + \varepsilon_t \tag{3}$$
$$\rho Y_{t-1} = \rho\alpha + \rho\beta X_{t-1} + \rho\varepsilon_{t-1} \tag{4}$$

  - Subtract (4) from (3):

$$\begin{aligned}(Y_t - \rho Y_{t-1}) &= (1-\rho)\alpha + \beta(X_t - \rho X_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1} \\ &= (1-\rho)\alpha + \beta(X_t - \rho X_{t-1}) + \nu_t\end{aligned} \tag{5}$$

- The *quasi-differenced* equation, (5), has a serially uncorrelated error term
  - OLS on the transformed equation is a version of *generalized least squares* (GLS), a strategy that transforms the original model into one with residuals that have classical properties

- Provided υ_t is homoskedastic, conventional standard errors for estimates of equation (5) should not be misleading. We therefore estimate:

$$Y_t^* = \alpha^* + \beta X_t^* + \nu_t \tag{6}$$

β is the same in original and quasi-differenced equations

where

$$\begin{aligned}Y_t^* &= (Y_t - \hat\rho Y_{t-1}) \\ X_t^* &= (X_t - \hat\rho X_{t-1}) \\ \alpha^* &= (1-\hat\rho)\alpha,\end{aligned}$$

From equation 5 except use $\hat\rho$ instead of ρ

and $\hat\rho$ is a consistent estimate of ρ, computed from OLS residuals

  - Does it matter that we quasi-difference using $\hat\rho$ rather than ρ? Not in asymptopia!
  - Stata automates quasi-differencing using a command called `Prais` (after Prais-Winsten, who invented a version of the procedure). You'll use a variant known as Cochrane-Orcutt (CORC) on Pset 5

The Durbin-Watson Test

- How to decide whether serial correlation is a problem? Simplest is to look at the regression of $\hat{\epsilon}_t$ on $\hat{\epsilon}_{t-1}$, a direct estimate of ρ
- In some cultures (e.g., among the Lost Tribes of Macroeconomia, isolated from modern applied microeconomics), it's customary to look at the Durbin-Watson (DW) statistic:

$$\begin{aligned} DW &= \sum_{t=2}^{T}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2 / \sum_{t=1}^{T}\hat{\varepsilon}_t^2 \\ &= \sum_{t=2}^{T}(\hat{\varepsilon}_t^2 - 2\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} + \hat{\varepsilon}_{t-1}^2) / \sum_{t=1}^{T}\hat{\varepsilon}_t^2 \\ &= \left\{\sum_{t=2}^{T}\hat{\varepsilon}_t^2 / \sum_{t=1}^{T}\hat{\varepsilon}_t^2\right\} - 2\left\{\sum_{t=2}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} / \sum_{t=1}^{T}\hat{\varepsilon}_t^2\right\} + \left\{\sum_{t=2}^{T}\hat{\varepsilon}_{t-1}^2 / \sum_{t=1}^{T}\hat{\varepsilon}_t^2\right\} \\ &\approx 1 - 2\hat{\rho} + 1 = 2(1-\hat{\rho}) \end{aligned}$$

- Moral: Look for DW that is "close to 2" when testing $H_0 : \rho = 0$. How close? Stata computes DW p-values

Modern Times: Newey-West Standard Errors

- Hetero *and* serial ... what's a young master to do?
- The modern analog of "robust" standard errors for time series data is Newey-West, named after our colleague, Whitney K. Newey, who invented these in his youth (with collaborator Ken West)
- These are also called HAC standard errors (*Heteroskedasticity and Autocorrelation Consistent*)
- HAC SEs allow for unrestricted serial correlation and heteroskedasticity, with the former not limited to AR(1) as in the quasi-differencing GLS procedure
  - Stata can HAC it (but you must pick the lag length)

1.3 Something Fishy in the Data (from Graddy 1995)

Our next 'metrics app comes from from Katy Graddy's analysis of the Fulton Fish Market. The Fulton Fish Market, which moved from lower Manhattan to the South Bronx in 2005, is the second largest wholesale fish market in the world (Tokyo's Tsukiji is the largest).

- Graddy looks for evidence of non-competitive behavior at Fulton by comparing the prices of fish paid by buyers of different ethnicities (she observes a single white seller).
- Does the law of one price hold for fish - or is there something fishy about fish prices? The economic rationale for Asian-white differences in price paid (for the same fish) is *price discrimination.*
- As luck would have it, the fish of interest in this study is called *whiting.*

```
      name:  <unnamed>
       log:  /Users/joshangrist/Documents/teaching/14.32/FA2020/notes/LN11/newfishformat/newfish_oct2020.smcl
  log type:  smcl
 opened on:  30 Oct 2020, 17:52:41

 1 . /*******************************************************************************
   > *Title:         Serial Correlation Fix-ups
   > *Author:        Robert Upton 04-17-20; JA edit 10-27-20
   > *******************************************************************************/
 2 .
 3 . cd "/Users/joshangrist/Documents/teaching/14.32/FA2020/notes/LN11/newfishformat/"
   /Users/joshangrist/Documents/teaching/14.32/FA2020/notes/LN11/newfishformat

 4 .
 5 . use fish.dta, clear

 6 . *** data start wide, with asians and whites as columns, rows are days
 7 . *** generate a unique id identifier (i)
 8 . gen t = _n

 9 .
10 . *** reshape to panel format
11 .
12 . reshape long price_ qty_, i(t) j(race) string
   (note: j = a w)

   Data                               wide   ->   long
   -----------------------------------------------------------------------------
   Number of obs.                       97   ->     194
   Number of variables                  15   ->      14
   j variable (2 values)                     ->   race
   xij variables:
                          price_a price_w   ->   price_
                              qty_a qty_w   ->   qty_
   -----------------------------------------------------------------------------

13 .
14 . list race t price* day* in 1/10

      +-----------------------------------------------------+
      | race   t     price_   day1   day2   day3   day4     |
      |-----------------------------------------------------|
   1. |    a   1   .6222222      1      0      0      0     |
   2. |    w   1   .7666667      1      0      0      0     |
   3. |    a   2   .9722222      0      0      1      0     |
   4. |    w   2      1.175      0      0      1      0     |
   5. |    a   3   1.233333      0      0      0      1     |
      |-----------------------------------------------------|
   6. |    w   3      1.475      0      0      0      1     |
   7. |    a   4   1.928571      0      0      0      0     |
   8. |    w   4      1.625      0      0      0      0     |
   9. |    a   5    .803125      1      0      0      0     |
  10. |    w   5   .8642857      1      0      0      0     |
      +-----------------------------------------------------+

15 .
16 . /**********                    Regression Analysis (Time Series)         ********************/
17 .
18 . gen ln_price = log(price)

19 . *** create asian dummy in long file format
20 . gen asian = race == "a"

21 . replace t = t + 100 if asian == 1
   (97 real changes made)

22 . *** regress on asian dummy, no special time series issues, but drop obs 101 to mimic tsset t processing for DW
23 . reg ln_price asian day* wave* if t!=101

      Source |       SS           df       MS      Number of obs   =       193
-------------+----------------------------------   F(7, 185)       =     12.58
       Model |  10.3491393         7  1.47844847   Prob > F        =    0.0000
    Residual |  21.7484389       185  .117559129   R-squared       =    0.3224
-------------+----------------------------------   Adj R-squared   =    0.2968
       Total |  32.0975782       192  .167174886   Root MSE        =    .34287

------------------------------------------------------------------------------
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asian |  -.0944506   .0493681    -1.91   0.057    -.1918474    .0029462
        day1 |  -.0126533    .079684    -0.16   0.874    -.1698593    .1445528
        day2 |  -.0232779   .0778549    -0.30   0.765    -.1768754    .1303197
        day3 |   .0531658   .0775971     0.69   0.494    -.0999232    .2062547
        day4 |   .1091356   .0770476     1.42   0.158    -.0428692    .2611404
       wave2 |   .0968629   .0148107     6.54   0.000     .0676433    .1260825
       wave3 |   .0542783   .0139477     3.89   0.000     .0267614    .0817953
```

```
       _cons |  -.9783857   .1041993    -9.39   0.000    -1.183957    -.772814
------------------------------------------------------------------------------

24 .
25 . *** tell stata "what time is" so we can get the DW, a time series command
26 .
27 . tsset t
        time variable:  t, 1 to 197, but with a gap
                delta:  1 unit

28 . estat dwatson

   Number of gaps in sample:  1

   Durbin-Watson d-statistic(  8,   193) =  .8134966

29 .
30 . predict e, residual

31 . gen Le = e[_n-1]
   (1 missing value generated)

32 . regress e Le if t!=101

      Source |       SS           df       MS      Number of obs   =       192
-------------+----------------------------------   F(1, 190)       =     96.25
       Model |    7.231917         1    7.231917   Prob > F        =    0.0000
    Residual |   14.2755252       190  .075134343   R-squared       =    0.3363
-------------+----------------------------------   Adj R-squared   =    0.3328
       Total |   21.5074422       191  .112604409   Root MSE        =    .27411

------------------------------------------------------------------------------
           e |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          Le |   .5768094   .0587929     9.81   0.000     .4608388    .6927801
       _cons |   .0069138   .0197869     0.35   0.727    -.0321165    .0459441
------------------------------------------------------------------------------

33 .
34 . *** to install newey2: ssc install newey2
35 . *** tell stata the file is a panel to avoid a bad lag over the asian-white seam
36 . *** (can't then estat DW, but this traps an error for Newey2; otherwise use option force)
37 .
38 . tsset asian t
        panel variable:  asian (weakly balanced)
         time variable:  t, 1 to 197
                 delta:  1 unit

39 .
40 . newey2 ln_price asian day* wave2 wave3 if t!=101 , lag(1)

   Regression with Newey-West standard errors          Number of obs  =       193
   maximum lag : 1                                     F(  7,   185)  =     15.37
                                                       Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asian |  -.0944506   .0621183    -1.52   0.130     -.217002    .0281008
        day1 |  -.0126533   .0692561    -0.18   0.855    -.1492865    .1239799
        day2 |  -.0232779   .0841152    -0.28   0.782    -.1892263    .1426705
        day3 |   .0531658   .0775967     0.69   0.494    -.0999225    .2062541
        day4 |   .1091356   .0595184     1.83   0.068    -.0082864    .2265577
       wave2 |   .0968629   .0139071     6.97   0.000     .0694261    .1242997
       wave3 |   .0542783   .0119934     4.53   0.000      .030617    .0779397
       _cons |  -.9783857   .1070929    -9.14   0.000    -1.189666   -.7671053
------------------------------------------------------------------------------

41 . newey2 ln_price asian day* wave2 wave3 if t!=101 , lag(2)

   Regression with Newey-West standard errors          Number of obs  =       193
   maximum lag : 2                                     F(  7,   185)  =     15.51
                                                       Prob > F       =    0.0000

------------------------------------------------------------------------------
             |             Newey-West
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asian |  -.0944506   .0688601    -1.37   0.172    -.2303026    .0414014
        day1 |  -.0126533   .0657705    -0.19   0.848    -.1424098    .1171033
        day2 |  -.0232779    .078556    -0.30   0.767    -.1782586    .1317028
        day3 |   .0531658   .0745459     0.71   0.477    -.0939035    .2002351
        day4 |   .1091356    .055023     1.98   0.049     .0005824    .2176888
       wave2 |   .0968629   .0144706     6.69   0.000     .0683142    .1254116
```

```
       wave3 |   .0542783    .011804     4.60   0.000     .0309906    .0775661
       _cons |  -.9783857   .1130564    -8.65   0.000    -1.201431   -.7553402
------------------------------------------------------------------------------

. 
. *** oldschool: quasi-diff
. *** stata prais command does corchrane-orcutt quasi-differencing, use option "corc twostep" to prevent iteration
. *** (corc will drop the first obs in each time series sequence)
. 
. prais ln_price asian day* wave2 wave3, corc twostep

Number of gaps in sample:  1   (gap count includes panel changes)
(note: computations for rho restarted at each gap)

Iteration 0:  rho = 0.0000
Iteration 1:  rho = 0.5792

Cochrane-Orcutt AR(1) regression -- twostep estimates

      Source |       SS           df       MS      Number of obs   =       192
-------------+----------------------------------   F(7, 184)       =      5.09
       Model |  2.64091725         7  .377273893   Prob > F        =    0.0000
    Residual |  13.6427548       184  .074145407   R-squared       =    0.1622
-------------+----------------------------------   Adj R-squared   =    0.1303
       Total |  16.283672        191  .085254827   Root MSE        =     .2723

------------------------------------------------------------------------------
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       asian |  -.0963775   .0933951    -1.03   0.303    -.2806405    .0878854
        day1 |   .0051914   .0499291     0.10   0.917    -.0933156    .1036985
        day2 |   -.014794   .0560661    -0.26   0.792    -.1254091    .0958211
        day3 |   .0604477   .0562757     1.07   0.284    -.0505809    .1714764
        day4 |   .1013305   .0473123     2.14   0.034     .0079862    .1946749
       wave2 |   .0605689   .0126723     4.78   0.000     .0355673    .0855705
       wave3 |   .0435863   .0128231     3.40   0.001     .0182871    .0688855
       _cons |  -.7280926   .1216098    -5.99   0.000    -.9680215   -.4881636
-------------+----------------------------------------------------------------
         rho |   .5791789
------------------------------------------------------------------------------

Durbin-Watson statistic (original)    0.820225
Durbin-Watson statistic (transformed) 1.639846

. 
. log close
      name:  <unnamed>
       log:  /Users/joshangrist/Documents/teaching/14.32/FA2020/notes/LN11/newfishformat/newfish_oct2020.smcl
  log type:  smcl
 closed on:  30 Oct 2020, 17:52:41
-------------------------------------------------------------------------------
```

2 Clustering in Data with a Group Structure

Many samples of interest to us have a group structure. For example, data on K-12 test scores come from samples of children who are naturally grouped into classes and schools. The principal econometric challenge in such settings arises from the fact that observations within groups are correlated.

- Regressions for data with a group structure can be written like this:

$$Y_{ig} = \beta_0 + \beta_1 x_g + \varepsilon_{ig} \tag{7}$$

  - Y_{ig} is the dependent variable for individual i in group g
  - Regressor x_g varies only at the group level

- Do small classes enhance learning? Krueger (1999) and Angrist and Lavy (1999) study the effects of class size on test scores. These studies analyze samples of children grouped in schools and classes.
  - Children in the same classroom have much in common; they have, for example, the same teacher. This makes their test scores dependent or *clustered.*
  - Clustering is often a big deal: clustered standard errors take the shine off many a bright idea!

- Data from the STAR experiment analyzed by Krueger (1999) consist of Y_{ig}, the test score of student i in class g, and class size, x_g. Scores of students in the same class are probably correlated. We might therefore assume that, for students i and j in the same class, g:

$$E[\varepsilon_{ig}\varepsilon_{jg}] = \rho\sigma_e^2 > 0 \qquad \rho = Corr(\varepsilon_{ig}, \varepsilon_{jg}) = \frac{Cov(\varepsilon_{ig}, \varepsilon_{jg})}{\sigma_\varepsilon^2} = \frac{E[\varepsilon_{ig}\varepsilon_{jg}]}{\sigma_\varepsilon^2} \tag{8}$$

  while

$$E[\varepsilon_{ig}\varepsilon_{jh}] = 0;\ i \neq j, g \neq h \quad \text{[for students in different classes]}$$

  In other words, resids are assumed to be correlated within classes though not between them.
  - This intra-class correlation means that observations on children in the same classroom are less informative about class size effects than data drawn from other classrooms. To see why, imagine that children in the same class are so similar that they have the same values of Y_{ig}. We then learn about the entire class by observing a single student.

- Notation notes: ρ is the (positive) intra-class correlation coefficient and σ_ε^2 is the residual variance (ρ is called an *intra-class correlation coefficient* even when the groups of interest are not classrooms)
  - Why is ρ a correlation coefficient? Like any regression residual, $E[\varepsilon_{ig}] = 0$, so $E[\varepsilon_{ig}\varepsilon_{jg}] = C(\varepsilon_{ig}, \varepsilon_{jg})$. Note also that ε_{ig} and ε_{jg} are presumed to have the same variance, σ_ε^2, a homoskedasticity assumption
  - Correlation is covariance divided by the product of standard deviations. So, for $i \neq j$:

$$CORR(\varepsilon_{ig}, \varepsilon_{jg}) = \frac{\rho\sigma_\varepsilon^2}{\sigma_\varepsilon\sigma_\varepsilon} = \rho$$

  - The random-effects model is also said to *equi-correlated* because, within groups, any two observations are equally correlated (contrast this with autoregressive time series models, in which correlation diminishes over time)

2.1 Random-Effects Fix-ups for Clustering

Random groups effect (common across everyone in the group)

- A *random-effects model* postulates

$$\varepsilon_{ig} = v_g + \eta_{ig} \quad \text{(9)}$$

(η_{ig}: individual level error term)

  - v_g is an error component ("random effect") specific to class g, assumed to be uncorrelated across classes:

$$E[v_g v_h] = 0; \; g \neq h$$
$$E[v_g^2] = \sigma_v^2$$

  - η_{ig} is a student-level error component assumed to satisfy:

$E[v_g, \eta_{ig}] = 0$ Also $E[\eta_{ig}, \eta_{jg}] = 0$

$$E[\eta_{ig}\eta_{jh}] = 0; \; i \neq j, g \neq h$$
$$E[\eta_{ig}^2] = \sigma_\eta^2$$

  - These error components are assumed to be homoskedastic (we're focusing on dependence within clusters) and defined so that they're uncorrelated with each other (η_{ig} is the resid from a regression of ε_{ig} on group dummies).

- In this random-effects model, the intra-class correlation coefficient becomes:

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} \quad \text{(10)}$$

(show this)

$$\rho = \frac{Cov(\varepsilon_{ig}, \varepsilon_{jg})}{\sqrt{V(\varepsilon_{ig})V(\varepsilon_{ig})}} = \frac{Cov(v_g+\eta_{ig}, v_g+\eta_{jg})}{(V(v_g+\eta_{ig})V(v_g+\eta_{jg}))^{1/2}} = \frac{\sigma_v^2}{\sigma_v^2+\sigma_\eta^2}$$

Clustering is consequential

- Let $V_c(\widehat{\beta}_1)$ be the classical OLS sampling variance and let $V(\widehat{\beta}_1)$ be the variance that accounts for random-effects clustering. In a random-effects model with regressors constant within groups and groups of equal size, n, we have:

$SE_c = \sqrt{V_c(\hat{\beta}_1)} = \frac{\sqrt{V(\hat{\beta}_1)}}{\sqrt{1+(n-1)\rho}} = \frac{SE(\hat{\beta}_1)}{\text{Moulton Factor}}$

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + (n-1)\rho \quad \text{(11)}$$

- The square root of this is called the *Moulton factor*, after Moulton (1986); a simple cluster fix-up is to multiply conventional SEs by the Moulton factor. $\rightarrow SE_c \cdot \sqrt{\text{Moulton factor}} = SE(\hat{\beta}_1)$

- In Angrist and Lavy (2009), a randomized evaluation of high school achievement awards, 4000 students are grouped in 40 schools, so average $n = 100$. The regressor of interest is a treatment dummy indicating schools that offered large cash awards to students who pass a matriculation exam. The intra-class residual correlation in this study is around 0.10. Applying formula ((11)), the Moulton factor is over 3!

- The general Moulton formula is

Clustering has a bigger impact on standard errors with variable groupsizes and when ρ_x is large.

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + \left[\frac{V(n_g)}{\bar{n}} + (\bar{n}-1)\right]\rho_x\rho, \quad \text{(12)}$$

where n_g is the size of group g, $\bar{n}$ is average group size, and ρ_x is the intra-class correlation of the regressor, x_{ig} (this is less than 1 when regressors are not constant within groups)

- The Moulton worst-case scenario is when the regressor of interest is fixed within groups ($\rho_x = 1$), as it is for a regressor like class size

In this case, the Moulton factor is n

Better Head Back to Tennessee, Jed

- In the Krueger (1999) data, a regression of Tennessee STAR Kindergartners' percentile score on class size yields an estimate of -0.62 with a robust (HC_1) standard error of 0.09
- In this case, $\rho_x = 1$ because class size is fixed within classes while $V(n_g)$ is positive because classes vary in size ($V(n_g) = 17.1$)
- The intra-class correlation coefficient for residuals is .31 and the average class size is 19.4
- These numbers give a value of about 7 for $\frac{V(\widehat{\beta_1})}{V_c(\widehat{\beta_1})}$, so that conventional standard errors should be multiplied by a factor of $2.65 = \sqrt{7}$ to adjust for clustering

2.2 Other Cluster Fix-Ups

- GLS for random-effects models is a kind of quasi-differencing procedure similar to that used for serial correlation
- The Stata `cluster` option generalizes robust SEs to the clustered setting
  - Stata cluster works by treating entire clusters as the sampling unit instead of individual data
  - Clustered standard errors may be unreliable with few clusters

MHE Table 8.2.1 compares standard-error fix-ups in Tennessee.

TABLE 8.2.1
Standard errors for class size effects in the STAR data (318 clusters)

Variance Estimator	Std. Err.
Robust (HC_1)	.090
Parametric Moulton correction (using Moulton intraclass correlation)	.222
Parametric Moulton correction (using Stata intraclass correlation)	.230
Clustered	.232
Block bootstrap	.231
Estimation using group means (weighted by class size)	.226

Notes: The table reports standard errors for the estimates from a regression of kindergartners' average percentile scores on class size using the public use data set from Project STAR. The coefficient on class size is −.62. The group level for clustering is the classroom. The number of observations is 5,743. The bootstrap estimate uses 1,000 replications.

- The SE generated by running regression (7) on 318 group means instead if 5,743 students is very close to the clustered standard error
- In clustered samples, the effective sample size is closer to the number of clusters than the number of people

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 11

Instrumental Variables and 2SLS for OVB

1 Recap: Regression and the CIA

- Recall the potential outcomes model for effects of private university attendance (P_i) on wages:
  - $Y_{0i} = \alpha + \eta_i$, where $E[Y_{0i}] = \alpha$; assume $Y_{1i} - Y_{0i} = \delta$. This means:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})P_i = \alpha + \delta P_i + \eta_i \tag{1}$$

    where δ is a new Greek name for the causal effect of private college attendance

- The CEF of Y_i given P_i is linear, so the regression of Y_i on P_i generates a difference in means:

$$\frac{C(Y_i, P_i)}{V(P_i)} = E[Y_i|P_i = 1] - E[Y_i|P_i = 0] = \delta + \{E[\eta_i|P_i = 1] - E[\eta_i|P_i = 0]\}$$

  - Uncontrolled comparisons equal the causal effect of interest plus selection bias

- Regression captures causal effects by invoking a *conditional independence assumption* (CIA):

$$E[\eta_i|P_i, X_i] = E[\eta_i|X_i] = \gamma' X_i, \tag{2}$$

  for a set of observed control variables, X_i. Equivalently,

$$\eta_i = \gamma' X_i + u_i$$

  Basically saying that once you control for X_i, selection bias should be 0 as η_i would not depend on P_i

  where u_i and X_i are uncorrelated by construction and u_i and P_i are uncorrelated by the CIA
  - In MM Chpt 2, the variables in X_i are dummies for Barrons selectivity groups or simply the average SAT scores of colleges applied to and dummies for number of apps submitted

- The CIA yields a causal regression model:

$$Y_i = \alpha + \delta P_i + \gamma' X_i + u_i, \tag{3}$$

  that's free of OVB

- Nice work if you can get it! Alas, you won't always be so lucky in the controls department. What's a young master to do?

2 Instrumental Variables Eliminate Selection Bias

2.1 Waiting for Superman

- Many children in large urban school districts leave school with poor reading and math skills. Economists believe that lack of basic skills perpetuates poverty and increases inequality.

- Charter schools—privately managed public schools—offer a possible solution
  - Charters are funded by the host district, but free to deviate from local district requirements (such as restrictions on who can be hired to teach) and to opt out of teachers' union contracts that determine pay and job security at traditional public schools
  - The Knowledge is Power Program (KIPP) is iconic in the charter universe, serving mostly urban minority students. KIPP's "No Excuses" charter recipe includes a long school day and year, data-driven instruction, heavy use of TFA and tutoring, and emphasizes discipline and comportment
  - KIPP students tend to do better than other students in the district they hail from. But perhaps this is merely selection bias. The KIPP parent, after all, knows enough to find their child a coveted seat at KIPP.

    Here's what the critics say:

    > KIPP students, as a group, enter KIPP with substantially higher achievement than the typical achievement of schools from which they came. . . . [T]eachers told us either that they referred students who were more able than their peers, or that the most motivated and educationally sophisticated parents were those likely to take the initiative . . . and enroll in KIPP.

- The charter selection bias story
  - Let D_i denote attendance at KIPP and Y_i be an achievement test outcome
  - Under constant causal effects,

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = \alpha + \lambda D_i + \eta_i, \tag{4}$$

    where λ ("lambda") is a Greek name for the causal effect of interest. Once again, simple comparisons are confounded by selection bias:

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \lambda + \{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]\}$$

2.2 Defining Instruments

- An instrumental variable (Z_i) for KIPP attendance is correlated with D_i but uncorrelated with Y_{0i}. **In the context of equation (4), instrument Z_i is assumed to satisfy two conditions:**

Basically trying to isolate effect of regressor without affecting error term

$$C(Z_i,\, D_i) \neq 0 \quad \text{Correlated with } D_i \text{: [Relevance Assumption]} \tag{5}$$

$$C(Z_i,\, \eta_i) = 0 \quad \text{Uncorrelated with } \eta_i \text{: [Exclusion Restriction Assumption]} \tag{6}$$

These conditions imply:

$$C(Z_i,\, Y_i) = C(Z_i,\, D_i)\lambda_{IV} \tag{7}$$

Causal effect of interest

$$\lambda_{IV} = \frac{C(Z_i,\, Y_i)}{C(Z_i,\, D_i)} = \frac{C(Z_i,\, Y_i)/V(Z_i)}{C(Z_i,\, D_i)/V(Z_i)} \tag{8}$$

Because we can solve for λ given information on the joint distribution of $\{Y_i, D_i, Z_i\}$, parameter λ is said to be *identified* by instrument Z_i. In this case, the solution for λ is a ratio of regression coefficients

- *Identification* problems are distinct from *estimation* problems
- The subscript "IV" appears on λ_{IV} because, as we'll soon see, IV estimators identify a particular sort of causal effect

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

$C(Z_i, Y_i) = C(Z_i, \alpha + \lambda D_i + \eta_i)$

$= C(Z_i, \alpha) + C(Z_i, \lambda D_i) + C(Z_i, \eta_i)$

$= 0 + \lambda(C(Z_i, D_i)) + 0$

$C(Z_i, Y_i) = \lambda(C(Z_i, D_i))$ $\quad \lambda_{IV} = \dfrac{C(Z_i, Y_i)}{C(Z_i, D_i)} = \dfrac{C(Z_i, Y_i)/V(Z_i)}{C(Z_i, D_i)/V(Z_i)}$

- The *IV estimator* can be written as a ratio of estimated regression coefficients:

$$\hat{\lambda}_{IV} = \frac{s_{ZY}/s_Z^2}{s_{ZD}/s_Z^2} \tag{9}$$

where s_{ZY} and s_{ZD} are sample covariances and s_Z^2 is the sample variance of the instrument, Z_i

  - Given assumptions (5) and (6), $\hat{\lambda}_{IV}$ is a consistent (though not unbiased) estimator of λ, with an asymptotically Normal sampling distribution and standard error formulas that we will derive later

- The top and bottom of the IV ratio, (8), are central to the IV story, so we christen them:

Regress Outcome on Instrument / Outcome on Regressor of interest

$$\frac{\textit{The Reduced Form}}{\textit{The First Stage}} = \frac{C(Z_i, Y_i)/V(Z_i)}{C(Z_i, D_i)/V(Z_i)} = \frac{\rho}{\phi} = \lambda_{IV}$$

reduced form estimate / first stage estimate

Sample analogs, denoted $\hat{\rho}$ and $\hat{\phi}$, are called "reduced form estimates" and "first stage estimates"

An alternate path to IV: Long Regression w/o Controls

- When estimating KIPP effects, we'd like to control for factors like ability and family background
  - Suppose this is the "long regression" we'd like to run:

$$Y_i = \alpha_l + \lambda_l D_i + \gamma' A_i + u_i, \tag{10}$$

where A_i is a vector of ability and family background controls, and λ_l is the KIPP effect of interest

- Alas, important control variables are unobserved. For example, ability is hard to measure.
  - Instrumental Variables (IV) methods allow us to recover the coefficient of interest in a long regression when long-regression controls are unavailable. In addition to the first stage requirement, condition (5), this justification for IV requires that Z_i be uncorrelated with omitted variables and the residual thats left over. That is, we replace (6) with $C(Z_i, A_i) = C(Z_i, u_i) = 0$ in (10).

2.3 Playing the KIPP Lottery

- Like all Massachusetts charter schools, KIPP Lynn assigns seats by lottery when over-subscribed
  - A research jackpot!
- In this case, instrument Z_i is a dummy variable indicating the set of KIPP applicants randomly offered a KIPP seat
  - Because the lottery is how most KIPP applicants get seated there, (5) is satisfied → $C(Z_i, D_i) \neq 0$
  - Because lottery offers are randomly assigned, they're likely to be independent of potential outcomes, satisfying (6) $C(Z_i, \eta_i) = 0$

Correlated with attendance but not with outcomes

- Bernoulli (dummy) instruments generate a useful simplification of (8):

$$\lambda_{IV} = \frac{C(Z_i, Y_i)/V(Z_i)}{C(Z_i, D_i)/V(Z_i)} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \tag{11}$$

For Bernoulli Instruments:

- We can therefore construct IV estimates using a ratio of differences in means:

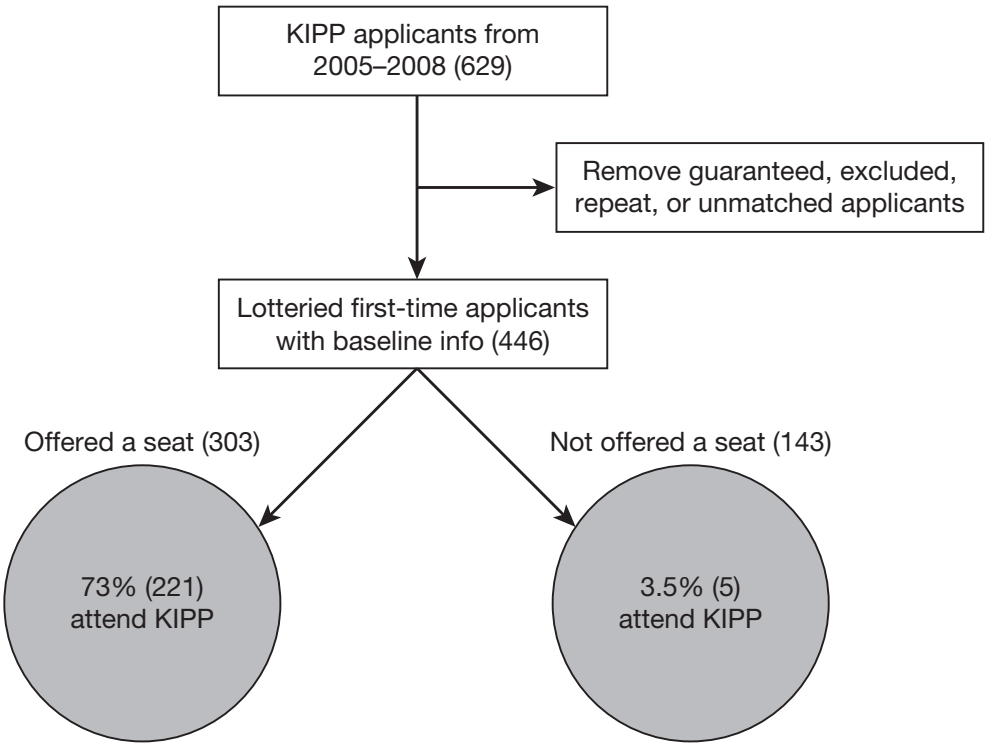
$$\hat{\lambda}_{IV} = (\bar{Y}_1 - \bar{Y}_0)/(\bar{D}_1 - \bar{D}_0), \tag{12}$$

  where $\bar{Y}_j$ and $\bar{D}_j$ are sample means of Y_i and D_i conditional on $Z_i = j$
- The formula in (12) is called a *Wald estimator* after Wald (1940)
- Since D_i is also Bernoulli, $\bar{D}_j$ is equal to the KIPP enrollment rate conditional on $Z_i = j$

The KIPP First Stage

- This figure diagrams the lottery first stage for applicants to KIPP Lynn, applying for 5th and 6th grade seats in the years 2005-2008

FIGURE 3.1
Application and enrollment data from KIPP Lynn lotteries



Note: Numbers of Knowledge Is Power Program (KIPP) applicants are shown in parentheses.

- Here, $\bar{D}_1 = .73$ and $\bar{D}_0 = .035$
- Lotteries at KIPP Lynn ensure *ceteris paribus* in comparisons between applicants with $Z_i = 0$ and $Z_i = 1$

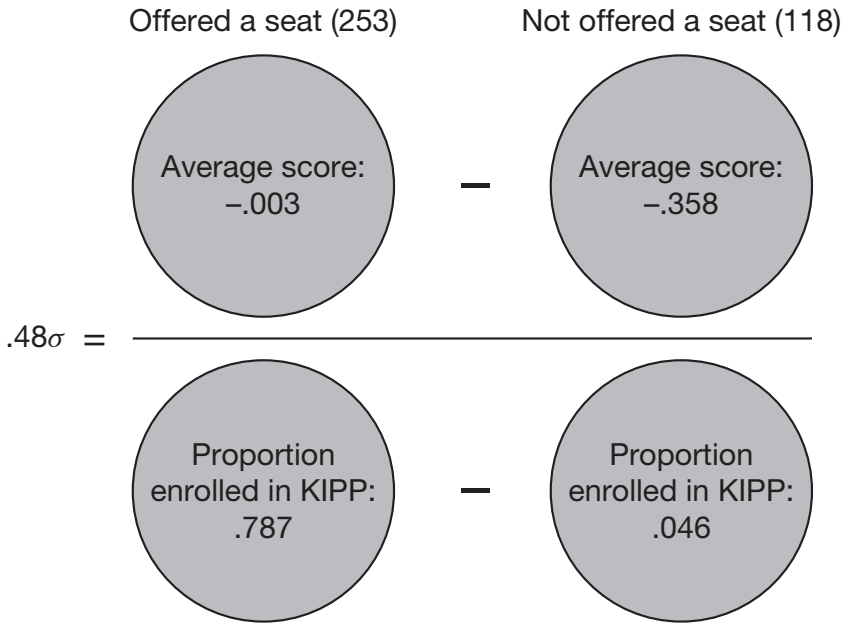
- The table below describes KIPP's 2005-8 applicants for 5th and 6th grade seats; outcomes are from the end of these grades (for the 371 tested lottery applicants; baseline scores are from 4th grade). Test scores are standardized to the state mean and variance.

TABLE 3.1
Analysis of KIPP lotteries

		KIPP applicants			
	Lynn public fifth graders (1)	KIPP Lynn lottery winners (2)	Winners vs. losers (3)	Attended KIPP (4)	Attended KIPP vs. others (5)
Panel A. Baseline characteristics					
Hispanic	.418	.510	−.058	.539	.012
			(.058)		(.054)
Black	.173	.257	.026	.240	−.001
			(.047)		(.043)
Female	.480	.494	−.008	.495	−.009
			(.059)		(.055)
Free/Reduced price lunch	.770	.814	−.032	.828	.011
			(.046)		(.042)
Baseline math score	−.307	−.290	.102	−.289	.069
			(.120)		(.109)
Baseline verbal score	−.356	−.386	.063	−.368	.088
			(.125)		(.114)
Panel B. Outcomes					
Attended KIPP	.000	.787	.741	1.000	1.000
			(.037)		—
Math score	−.363	−.003	.355	.095	.467
			(.115)		(.103)
Verbal score	−.417	−.262	.113	−.211	.211
			(.122)		(.109)
Sample size	3,964	253	371	204	371

Notes: This table describes baseline characteristics of Lynn fifth graders and reports estimated offer effects for Knowledge Is Power Program (KIPP) Lynn applicants. Means appear in columns (1), (2), and (4). Column (3) shows differences between lottery winners and losers. These are coefficients from regressions that control for risk sets, namely, dummies for year and grade of application and the presence of a sibling applicant. Column (5) shows differences between KIPP students and applicants who did not attend KIPP. Standard errors are reported in parentheses.

FIGURE 3.2
IV in school: the effect of KIPP attendance on math scores



Note: The effect of Knowledge Is Power Program (KIPP) enrollment described by this figure is $.48\sigma = .355\sigma/.741$.

3 IV is always LATE

3.1 The Four Types of Children

- KIPP lottery offers affect KIPP enrollment for many applicants . . . but not all
  - Some offered a seat at KIPP nevertheless go elsewhere
  - A few not offered a seat in the lottery manage to sneak in anyway
- How should we interpret IV estimates in light of this fact?

TABLE 3.2
The four types of children

		Lottery losers $Z_i = 0$	
		Doesn't attend KIPP $D_i = 0$	Attends KIPP $D_i = 1$
Lottery winners $Z_i = 1$	Doesn't attend KIPP $D_i = 0$	Never-takers (*Normando*)	Defiers
	Attends KIPP $D_i = 1$	Compliers (*Camila*)	Always-takers (*Alvaro*)

Note: KIPP = Knowledge Is Power Program.

(*Actually there are are only three: no defiers allowed!*)

- In a world of heterogeneous potential outcomes,

$$\lambda_{IV} = \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[Y_{1i} - Y_{0i}|C_i = 1],$$

  where C_i indicates compliers, like Camila
- Parameter $E[Y_{1i} - Y_{0i}|C_i = 1]$ is called a *local average treatment effect* (LATE)
- In general, LATE differs from the effect of treatment on the treated, $E[Y_{1i} - Y_{0i}|D_i = 1]$, because some treated, like Alvaro, are *always-takers*
  - As detailed in MHE, the proportion of always-takers is given by $E[D_i|Z_i = 0]$
  - With few always-takers (as in the KIPP lottery), we expect:

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

3.2 LATE Again: Effects of Vietnam-Era Military Service (Angrist 1990)

- From 1970-73, Uncle Sam selected soldiers in a *draft lottery*: Men born 1950-53 were called up by random sequence numbers (RSN), assigned to their DOB
- Men born in 1950 with RSN<195 were draft-eligible
- This table is from MHE:

Table 4.1.3

IV Estimates of the Effects of Military Service on the Earnings of White Men born in 1950

Earnings year	Earnings		Veteran Status		Wald Estimate of Veteran Effect
	Mean	Eligibility Effect	Inelig. Mean	Eligibility Effect	
	(1)	(2)	(3)	(4)	(5)
1981	16,461	-435.8 (210.5)	.182	.159 (.040)	-2,741 (1,324)
1971	3,338	-325.9 (46.6)			-2050 (293)
1969	2,299	-2.0 (34.5)			

Note: Adapted from Table 5 in Angrist and Krueger (1999) and author tabulations. Standard errors are shown in parentheses. Earnings data are from Social Security administrative records. Figures are in nominal dollars. Vete status data are from the Survey of Program Participation. There are about 13,500 individuals in the sample.

- What's the LATE interpretation here?
- A Vietnam-era update:

Angrist, Chen, and Song (2011)

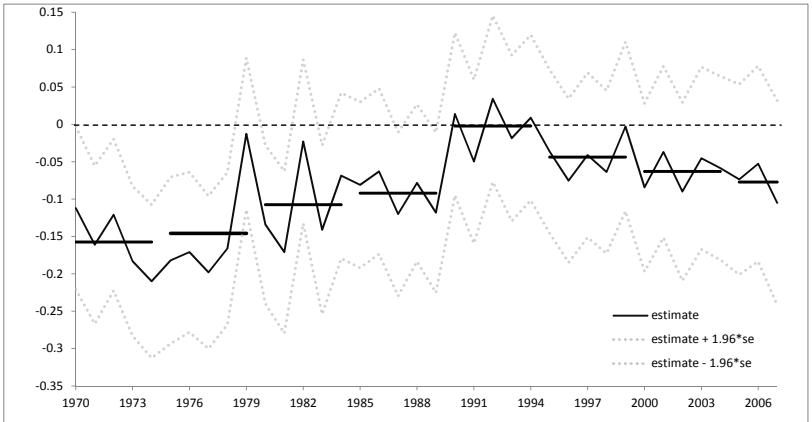


Figure 1. Draft-lottery Estimates of Vietnam-era Service Effects on ln(Earnings) for White Men Born 1950-52

- IV is everywhere! Reconsider, for example, the Carter, Greenberg and Walker (2017) class computer RCT and the OHP Medicaid effects in Taubman, et al. (2014). In fact, the OHP is done as IV in Finkelstein, et al. (2012). Ponder the first stage in these imperfectly-randomized randomized trials.
- Here's MM OHP Table 1.5 again:

TABLE 1.5
OHP effects on insurance coverage and health-care use

Outcome	Oregon Control mean (1)	Oregon Treatment effect (2)	Portland area Control mean (3)	Portland area Treatment effect (4)
A. Administrative data				
Ever on Medicaid	.141	.256 (.004)	.151	.247 (.006)
Any hospital admissions	.067	.005 (.002)		
Any emergency department visit			.345	.017 (.006)
Number of emergency department visits			1.02	.101 (.029)
Sample size	74,922		24,646	
B. Survey data				
Outpatient visits (in the past 6 months)	1.91	.314 (.054)		
Any prescriptions?	.637	.025 (.008)		
Sample size	23,741			

Notes: This table reports estimates of the effect of winning the Oregon Health Plan (OHP) lottery on insurance coverage and use of health care. Odd-numbered columns show control group averages. Even-numbered columns report the regression coefficient on a dummy for lottery winners. Standard errors are reported in parentheses.

4 Two-Stage Least Squares

In practice, we do IV by doing two-stage least squares (2SLS). This allows us to add covariates (controls) and to use multiple instruments to generate more precise IV estimates.

This is essentially the same as normal IV

4.1 2SLS Derived

- Here's a nifty way to compute IV estimates: First, regress D_i on Z_i

$$D_i = \alpha_1 + \phi Z_i + e_{1i}$$

and save the first-stage fitted values:

$$\hat{D}_i = \alpha_1 + \phi Z_i$$

Then regress Y_i on these

Regress Outcome on Fitted Values of Regressor of Interest

$$Y_i = \alpha_2 + \lambda_{2SLS}\hat{D}_i + e_{2i}, \tag{13}$$

It's easy to show (be sure you can) that λ_{2SLS} in (13) equals λ_{IV} in (8) in both population and sample

$\lambda_{2SLS} = \frac{Cov(\hat{D}_i, Y_i)}{V(\hat{D}_i)} = \frac{Cov(\alpha_1 + \phi Z_i, Y_i)}{V(\alpha_1 + \phi Z_i)} = \frac{\phi\, Cov(Z_i, Y_i)}{\phi^2 V(Z_i)} = \frac{Cov(Z_i, Y_i)/V(Z_i)}{\phi} = \frac{Cov(Z_i, Y_i)/V(Z_i)}{Cov(Z_i, D_i)/V(Z_i)}$

Covs in the mix

- Suppose the causal model of interest includes covariates, X_i:

$$Y_i = \alpha_2' X_i + \lambda_{2SLS} D_i + \eta_i \tag{14}$$

  In the Superman story, for example, X_i includes dummies for application year (KIPP offers are randomized conditional on this).

- Write the first stage with covariates as the sum of first-stage fitted values plus first-stage residuals:

$$D_i = X_i' \alpha_1 + \phi Z_i + e_{1i} = \hat{D}_i + e_{1i}$$

  The 2SLS second stage is OLS on:

$$Y_i = \alpha_2' X_i + \lambda_{2SLS} \hat{D}_i + e_{2i} \tag{15}$$

- Why does this work? The key is that the second-stage residual is

$$e_{2i} = \lambda_{2SLS} e_{1i} + \eta_i$$

  and both pieces of this resid are orthogonal to $\hat{D}_i$, so $E[\hat{D}_i e_{2i}] = 0$.

  - Unlike in LN8, we're not bothering here to distinguish between estimated and population first-fitted values. The fact that in practice we must estimate first-stage fitted values turns out not to matter for the IV idea to work (in asymptopia)

- The first stage and reduced form regressions for the model with covariates are:

$$D_i = X_i' \alpha_1 + \phi Z_i + e_{1i} \tag{16}$$

$$Y_i = X_i' \alpha_0 + \rho Z_i + e_{0i} \tag{17}$$

  Equation (17) is obtained by substituting (16) into (14).

  - λ_{2SLS} is still the ratio of *RF* to *1st Stage* coefficients:

$$\lambda_{2SLS} = \frac{\rho}{\phi}$$

  (show this)

Mightier with more instruments

- Blessed with more than one instrument?

  - In the Superman story, we might use dummies for lottery offers made immediately (on lottery night) and later (to applicants on a waiting list)

- Add 'em to the first stage when baking the fits:

$$D_i = X_i' \alpha_1 + \phi_1 Z_{1i} + \phi_2 Z_{2i} + e_{1i}$$

  The second stage, equation (15), stays the same

- Models with more instruments than necessary are said to be *over-identified*

5 So Where *Do* Babies Come From?

5.1 Family Size Effects on Female Labor Supply (Angrist and Evans, 1998)

- Lotteries are awesome! Other instruments come from deep institutional knowledge, revealing, for example, the effect of children on their parents' labor supply
- AE-98 come up with two instruments to identify effects of having more than two kids on parents' work and earnings
  - The *twins instrument*, Z_{1i} indicates multiple second births (buy one, get one free)
  - The *samesex instrument*, Z_{2i} indicates mothers of two boys and two girls at parities 1 and 2 (diversify your sibling-sex portfolio)

(Stata for AE98 follows)

```
 1 .
 2 . *All women sample:
 3 . keep if ((agem1>=21 & agem1<=35) & (kidcount>=2) & (ageq2nd1>4) & (agefstm>=15) & (asex==0) & (aage==0) & (aqtrbrth==0) & (asex2nd==0) & (aage2nd==0))
   (532,427 observations deleted)

 4 .   /*& (agefstd>=15 | agefstd==.)*/
 5 .
 6 . *keep if (msample==1)
 7 .
 8 . sum agem1 kidcount ageq2nd1 agefstm weeksm1 workedm morekids agem1 boy1st boy2nd blackm hispm othracem multi2nd samesex msample
```

Variable	Obs	Mean	Std. Dev.	Min	Max
agem1	394,840	30.1248	3.509685	21	35
kidcount	394,840	2.552069	.8083876	2	12
ageq2nd1	394,840	26.36489	14.61527	5	70
agefstm	394,840	20.13956	2.949069	15	33
weeksm1	394,840	20.83419	22.28601	0	52
workedm	394,840	.5654873	.4956935	0	1
morekids	394,840	.4020641	.4903154	0	1
agem1	394,840	30.1248	3.509685	21	35
boy1st	394,840	.511088	.4998777	0	1
boy2nd	394,840	.5109614	.4998805	0	1
blackm	394,840	.1189343	.3237115	0	1
hispm	394,840	.03004	.1706976	0	1
othracem	394,840	.028685	.16692	0	1
multi2nd	394,840	.0085604	.0921258	0	1
samesex	394,840	.5053895	.4999716	0	1
msample	394,840	.6449499	.4785291	0	1

```
 9 .
10 . *OLS:
11 . reg weeksm1 morekids agem1 agefstm boy1st boy2nd blackm hispm othracem, r
```

Linear regression

Number of obs	394,840
F(8, 394831)	4589.07
Prob > F	0.0000
R-squared	0.0778
Root MSE	21.402

weeksm1	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
morekids	-8.978191	.0705666	-127.23	0.000	-9.1165	-8.839883
agem1	1.466036	.0105266	139.27	0.000	1.445404	1.486668
agefstm	-1.423913	.0131709	-108.11	0.000	-1.449728	-1.398099
boy1st	-.1153498	.0681462	-1.69	0.091	-.2489143	.0182147
boy2nd	-.1773649	.0681483	-2.60	0.009	-.3109335	-.0437963
blackm	6.451669	.1103587	58.46	0.000	6.235369	6.667968
hispm	-.7810209	.1956389	-3.99	0.000	-1.164467	-.3975744
othracem	2.860371	.2109436	13.56	0.000	2.446928	3.273814
_cons	8.280615	.3199806	25.88	0.000	7.653463	8.907767

```
12 . reg workedm morekids agem1 agefstm boy1st boy2nd blackm hispm othracem, r
```

Linear regression

Number of obs	394,840
F(8, 394831)	3032.83
Prob > F	0.0000
R-squared	0.0537
Root MSE	.48222

workedm	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
morekids	-.1764489	.0016171	-109.11	0.000	-.1796184	-.1732793
agem1	.0241995	.0002424	99.84	0.000	.0237244	.0246745
agefstm	-.0291002	.0002967	-98.07	0.000	-.0296818	-.0285187
boy1st	-.0005312	.0015353	-0.35	0.729	-.0035404	.002478
boy2nd	-.0040863	.0015353	-2.66	0.008	-.0070955	-.0010771
blackm	.1060263	.0023474	45.17	0.000	.1014255	.110627
hispm	-.0309759	.0046057	-6.73	0.000	-.0400029	-.0219488
othracem	.0420805	.0046453	9.06	0.000	.0329759	.0511852
_cons	.4829654	.0075603	63.88	0.000	.4681474	.4977834

```
13 .
14 . *first stage and weeks reduced form: twins
15 . reg morekids multi2nd, r
```

```
Linear regression                               Number of obs     =    394,840
                                                F(0, 394838)      =          .
                                                Prob > F          =          .
                                                R-squared         =     0.0128
                                                Root MSE          =     .48716

------------------------------------------------------------------------------
             |               Robust
    morekids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    multi2nd |   .6030987    .000782   771.25   0.000      .601566    .6046313
       _cons |   .3969013    .000782   507.56   0.000     .3953687     .398434
------------------------------------------------------------------------------

16 . reg weeksm1 multi2nd, r

Linear regression                               Number of obs     =    394,840
                                                F(1, 394838)      =      27.19
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0001
                                                Root MSE          =     22.285

------------------------------------------------------------------------------
             |               Robust
    weeksm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    multi2nd |  -1.975956   .3789719    -5.21   0.000    -2.718729   -1.233182
       _cons |   20.8511   .0356232   585.32   0.000     20.78128    20.92092
------------------------------------------------------------------------------

17 . *Wald for twins
18 . ivregress 2sls weeksm1 (morekids = multi2nd)

Instrumental variables (2SLS) regression          Number of obs   =    394,840
                                                  Wald chi2(1)    =      26.71
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0138
                                                  Root MSE        =     22.132

------------------------------------------------------------------------------
    weeksm1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -3.276339   .6339241    -5.17   0.000    -4.518807   -2.033871
       _cons |   22.15149   .2573002    86.09   0.000     21.64719    22.65579
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   multi2nd

19 .
20 . *first stage and weeks reduced form: samesex
21 . reg morekids samesex, r

Linear regression                               Number of obs     =    394,840
                                                F(1, 394838)      =    1461.73
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0037
                                                Root MSE          =     .48941

------------------------------------------------------------------------------
             |               Robust
    morekids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |    .059544   .0015574    38.23   0.000     .0564915    .0625965
       _cons |   .3719712   .0010937   340.10   0.000     .3698276    .3741148
------------------------------------------------------------------------------

22 . reg weeksm1 samesex, r

Linear regression                               Number of obs     =    394,840
                                                F(1, 394838)      =      28.50
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0001
                                                Root MSE          =     22.285

------------------------------------------------------------------------------
             |               Robust
    weeksm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |  -.3786749   .0709378    -5.34   0.000    -.5177109   -.2396389
       _cons |   21.02557   .0505151   416.22   0.000     20.92656    21.12457
------------------------------------------------------------------------------

23 . *Wald for samesex
```

```
24 . ivregress 2sls weeksm1 (morekids = samesex)

Instrumental variables (2SLS) regression          Number of obs   =    394,840
                                                  Wald chi2(1)    =      29.00
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.0173
                                                  Root MSE        =     22.093

------------------------------------------------------------------------------
     weeksm1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -6.359578   1.181014    -5.38   0.000    -8.674324   -4.044833
       _cons |   23.39115   .4761434    49.13   0.000     22.45792    24.32437
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   samesex

25 .
26 . *check for balance
27 . reg agefstm multi2nd agem1 boy1st boy2nd blackm hispm othracem, r

Linear regression                                 Number of obs     =    394,840
                                                  F(7, 394832)      =   16568.12
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.1941
                                                  Root MSE          =     2.6475

------------------------------------------------------------------------------
             |               Robust
     agefstm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    multi2nd |   .1752039   .0457188     3.83   0.000     .0855964    .2648115
       agem1 |   .3290559   .0010836   303.66   0.000      .326932    .3311798
      boy1st |   .0094222   .0084286     1.12   0.264    -.0070976     .025942
      boy2nd |   .0193927   .0084301     2.30   0.021      .00287    .0359155
      blackm |  -1.427554   .0120888  -118.09   0.000    -1.451248   -1.40386
       hispm |  -.5792559   .0234584   -24.69   0.000    -.6252337   -.5332781
    othracem |   .6226536   .0280525    22.20   0.000     .5676715    .6776357
       _cons |   10.37992   .0320333   324.04   0.000     10.31713    10.4427
------------------------------------------------------------------------------

28 . reg educm multi2nd agem1 agefstm boy1st boy2nd blackm hispm othracem, r

Linear regression                                 Number of obs     =    394,840
                                                  F(8, 394831)      =    9264.86
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.2109
                                                  Root MSE          =     2.1325

------------------------------------------------------------------------------
             |               Robust
       educm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    multi2nd |  -.0074761   .0373593    -0.20   0.841    -.0806992    .0657471
       agem1 |   .0221166   .0010089    21.92   0.000     .0201392     .024094
     agefstm |   .3330336   .0014574   228.51   0.000     .3301771    .3358902
      boy1st |   .0036823   .0067884     0.54   0.588    -.0096229    .0169874
      boy2nd |   .0075476   .0067899     1.11   0.266    -.0057603    .0208555
      blackm |   .2191673   .0101611    21.57   0.000     .1992517    .2390828
       hispm |  -2.374502   .0326049   -72.83   0.000    -2.438406   -2.310597
    othracem |  -.531052   .0321654   -16.51   0.000    -.5940952   -.4680088
       _cons |   4.80712   .0341787   140.65   0.000     4.740131    4.874109
------------------------------------------------------------------------------

29 . reg agefstm samesex agem1 boy1st boy2nd blackm hispm othracem, r

Linear regression                                 Number of obs     =    394,840
                                                  F(7, 394832)      =   16567.41
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.1941
                                                  Root MSE          =     2.6475

------------------------------------------------------------------------------
             |               Robust
     agefstm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |   .0217124   .0084313     2.58   0.010     .0051874    .0382375
       agem1 |   .3290979   .0010836   303.71   0.000     .3269741    .3312217
      boy1st |   .0089868   .0084313     1.07   0.286    -.0075383     .025512
      boy2nd |   .0188476   .0084318     2.24   0.025     .0023216    .0353737
      blackm |  -1.427105    .012089  -118.05   0.000    -1.450799   -1.403411
       hispm |  -.5793764   .0234604   -24.70   0.000    -.625358   -.5333947
    othracem |   .6225523   .0280511    22.19   0.000      .567573    .6775316
       _cons |   10.36963   .0323018   321.02   0.000     10.30632    10.43294
```

```
30 . reg educm samesex agem1 agefstm boy1st boy2nd blackm hispm othracem, r

Linear regression                               Number of obs     =    394,840
                                                F(8, 394831)      =    9264.95
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2109
                                                Root MSE          =     2.1325

------------------------------------------------------------------------------
             |               Robust
      educm  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    samesex  |  -.0087736   .0067908    -1.29   0.196    -.0220833    .0045362
      agem1  |   .0221105   .0010089    21.92   0.000     .0201332    .0240878
    agefstm  |   .3330388   .0014574   228.51   0.000     .3301823    .3358953
     boy1st  |   .0038694   .0067906     0.57   0.569    -.0094399    .0171787
     boy2nd  |   .0077413   .0067907     1.14   0.254    -.0055683    .0210509
     blackm  |   .2191597   .0101598    21.57   0.000     .1992469    .2390726
      hispm  |  -2.374498   .0326047   -72.83   0.000    -2.438402   -2.310594
   othracem  |  -.5310898   .0321659   -16.51   0.000    -.594134    -.4680457
       _cons |   4.811376   .0343194   140.19   0.000     4.744112    4.878641
------------------------------------------------------------------------------

31 .
32 . *2sls: weeks (twins, w/covs)
33 . ivregress 2sls weeksm1 (morekids = multi2nd) agem1 agefstm boy1st boy2nd blackm hispm othracem, r

Instrumental variables (2SLS) regression        Number of obs     =    394,840
                                                Wald chi2(8)      =   18168.92
                                                Prob > chi2       =     0.0000
                                                R-squared         =     0.0654
                                                Root MSE          =     21.545

------------------------------------------------------------------------------
             |               Robust
   weeksm1   |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   morekids  |  -3.712292   .6036268    -6.15   0.000    -4.895379   -2.529205
      agem1  |   1.307164   .0209401    62.42   0.000     1.266122    1.348205
    agefstm  |  -1.186511   .0300822   -39.44   0.000    -1.245471   -1.127551
     boy1st  |  -.0804947   .0687157    -1.17   0.241    -.2151751    .0541857
     boy2nd  |  -.1385996   .0687455    -2.02   0.044    -.2733383   -.0038609
     blackm  |   6.075055   .1192318    50.95   0.000     5.841365    6.308745
      hispm  |  -1.603621   .2187741    -7.33   0.000    -2.032411   -1.174832
   othracem  |   2.482386    .216284    11.48   0.000     2.058477    2.906295
       _cons |   6.210914    .398797    15.57   0.000     5.429286    6.992542
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   agem1 agefstm boy1st boy2nd blackm hispm othracem multi2nd

34 .
35 . *2sls: weeks (samesex, w/covs)
36 . ivregress 2sls weeksm1 (morekids = samesex) agem1 agefstm boy1st boy2nd blackm hispm othracem, r

Instrumental variables (2SLS) regression        Number of obs     =    394,840
                                                Wald chi2(8)      =   18252.28
                                                Prob > chi2       =     0.0000
                                                R-squared         =     0.0726
                                                Root MSE          =     21.462

------------------------------------------------------------------------------
             |               Robust
   weeksm1   |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   morekids  |   -5.55877   1.117829    -4.97   0.000    -7.749673   -3.367866
      agem1  |   1.362872   .0352894    38.62   0.000     1.293706    1.432038
    agefstm  |  -1.269755   .0520245   -24.41   0.000    -1.371722   -1.167789
     boy1st  |  -.0927166   .0687273    -1.35   0.177    -.2274196    .0419864
     boy2nd  |  -.1521926   .0688292    -2.21   0.027    -.2870953   -.0172899
     blackm  |   6.207114   .1364545    45.49   0.000     5.939668    6.47456
      hispm  |  -1.315178   .2625227    -5.01   0.000    -1.829713   -.8006428
   othracem  |   2.614926   .2260306    11.57   0.000     2.171914    3.057938
       _cons |   6.936651   .5431087    12.77   0.000     5.872177    8.001124
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   agem1 agefstm boy1st boy2nd blackm hispm othracem samesex

37 .
38 . *2sls: weeks (overid, w/covs)
39 . ivregress 2sls weeksm1 (morekids = multi2nd samesex) agem1 agefstm boy1st boy2nd blackm hispm othracem, r

Instrumental variables (2SLS) regression        Number of obs     =    394,840
```

```
                                                Wald chi2(8)      =   18224.63
                                                Prob > chi2       =     0.0000
                                                R-squared         =     0.0674
                                                Root MSE          =     21.522

------------------------------------------------------------------------------
             |               Robust
    weeksm1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -4.141475   .5311981    -7.80   0.000    -5.182605   -3.100346
       agem1 |   1.320112   .0190752    69.21   0.000     1.282725    1.357499
    agefstm |    -1.20586   .0271733   -44.38   0.000    -1.259118   -1.152601
      boy1st |  -.0833355   .0686168    -1.21   0.225    -.2178218    .0511509
      boy2nd |  -.1417591   .0686413    -2.07   0.039    -.2762936   -.0072245
      blackm |    6.10575   .1173185    52.04   0.000     5.87581     6.33569
       hispm |  -1.536577   .2138234    -7.19   0.000    -1.955663   -1.117491
    othracem |   2.513193   .2151399    11.68   0.000     2.091526    2.934859
       _cons |   6.379599   .3822657    16.69   0.000     5.630372    7.128826
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   agem1 agefstm boy1st boy2nd blackm hispm othracem multi2nd samesex

40 .
41 . *manual 2SLS
42 . reg morekids multi2nd samesex agem1 agefstm boy1st boy2nd blackm hispm othracem

      Source |       SS           df       MS      Number of obs   =   394,840
-------------+----------------------------------   F(9, 394830)    =   4708.57
       Model |  9200.59068         9  1022.28785   Prob > F        =     0.0000
    Residual |  85722.3271   394,830   .21711199   R-squared       =     0.0969
-------------+----------------------------------   Adj R-squared   =     0.0969
       Total |  94922.9177   394,839  .240409174   Root MSE        =     .46595

------------------------------------------------------------------------------
    morekids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    multi2nd |   .6049071   .0080499    75.14   0.000     .5891295    .6206847
     samesex |   .0614735   .0014839    41.43   0.000     .0585652    .0643819
       agem1 |   .0301281   .0002318   129.97   0.000     .0296738    .0305825
    agefstm |  -.0452589   .0002801  -161.58   0.000    -.0458079   -.0447099
      boy1st |  -.0080449   .0014839    -5.42   0.000    -.0109533   -.0051366
      boy2nd |  -.0084413   .0014839    -5.69   0.000    -.0113497   -.0055329
      blackm |   .0696438   .0023467    29.68   0.000     .0650443    .0742433
       hispm |   .1565985    .004367    35.86   0.000     .1480392    .1651578
    othracem |   .0729161   .0044574    16.36   0.000     .0641797    .0816525
       _cons |   .3630539   .0072301    50.21   0.000     .3488831    .3772248
------------------------------------------------------------------------------

43 . predict more_hat if e(sample)
   (option xb assumed; fitted values)

44 . reg weeksm1 more_hat agem1 agefstm boy1st boy2nd blackm hispm othracem, r

Linear regression                               Number of obs     =    394,840
                                                F(8, 394831)      =    2227.06
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0420
                                                Root MSE          =     21.813

------------------------------------------------------------------------------
             |               Robust
    weeksm1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    more_hat |  -4.141476   .5336062    -7.76   0.000    -5.187328   -3.095624
       agem1 |   1.320112   .0192077    68.73   0.000     1.282466    1.357759
    agefstm |    -1.20586   .0273376   -44.11   0.000    -1.25944    -1.152279
      boy1st |  -.0833355   .0695413    -1.20   0.231    -.2196343    .0529634
      boy2nd |  -.1417591   .0695644    -2.04   0.042    -.2781032   -.0054149
      blackm |    6.10575    .118865    51.37   0.000     5.872778    6.338722
       hispm |  -1.536577   .2171298    -7.08   0.000    -1.962145   -1.111009
    othracem |   2.513193   .2175725    11.55   0.000     2.086757    2.939628
       _cons |   6.379599   .3862117    16.52   0.000     5.622636    7.136563
------------------------------------------------------------------------------

45 .
46 . log close
      name:  <unnamed>
       log:  /Users/joshangrist/Documents/teaching/14.32/2020/notes/LN14/AE98/AE98for1432.smcl
  log type:  smcl
 closed on:  17 Apr 2020, 21:26:35
-------------------------------------------------------------------------------
```

5.2 The Quantity-Quality Trade-Off (Angrist, Lavy, and Schlosser, 2010)

- In the 1970s and 1980s, governments around the world discouraged childbearing in the belief that small families increase living standards
  - China's One Child Policy is the most (in)famous of these anti-natalist policies
  - Economists call the relationship between family size and living standards the *quantity-quality tradeoff*
- Are larger families really impoverished by their size? If only we could randomize the number of children and find out!
  - Angrist, Lavy, and Schlosser (2010) exploit AE98-style natural experiments for family size in samples of women with 2 or more children
    - *twins instrument*s, Z_{1i} now indicate multiples at various births (e.g., twins@2, twins@3)
    - *samesex instrument*s, Z_{2i} now indicate mothers of various samesex sibships (e.g., 3 girls@3)
  - Z_{1i} and Z_{2i} are both highly predictive of the number of children born in family i
  - They're arguably independent of the *potential* human capital of the *first-borns* in these families (samples used to construct the tables below consists of first-born non-twin Israeli Jews aged 18-60 in the Census, whose mothers were born after 1930 and had their first birth between the ages of 15-45)

TABLE 3.4
Quantity-quality first stages

	Twins instruments		Same-sex instruments		Twins and same-sex instruments
	(1)	(2)	(3)	(4)	(5)
Second-born twins	.320	.437			.449
	(.052)	(.050)			(.050)
Same-sex sibships			.079	.073	.076
			(.012)	(.010)	(.010)
Male		−.018		−.020	−.020
		(.010)		(.010)	(.010)
Controls	No	Yes	No	Yes	Yes

Notes: This table reports coefficients from a regression of the number of children on instruments and covariates. The sample size is 89,445. Standard errors are reported in parentheses.

TABLE 3.5
OLS and 2SLS estimates of the quantity-quality trade-off

Dependent variable	OLS estimates (1)	2SLS estimates: Twins instruments (2)	2SLS estimates: Same-sex instruments (3)	2SLS estimates: Twins and same-sex instruments (4)
Years of schooling	−.145	.174	.318	.237
	(.005)	(.166)	(.210)	(.128)
High school graduate	−.029	.030	.001	.017
	(.001)	(.028)	(.033)	(.021)
Some college (for age ≥ 24)	−.023	.017	.078	.048
	(.001)	(.052)	(.054)	(.037)
College graduate (for age ≥ 24)	−.015	−.021	.125	.052
	(.001)	(.045)	(.053)	(.032)

Notes: This table reports OLS and 2SLS estimates of the effect of family size on schooling. OLS estimates appear in column (1). Columns (2), (3), and (4) show 2SLS estimates constructed using the instruments indicated in column headings. Sample sizes are 89,445 for rows (1) and (2); 50,561 for row (3); and 50,535 for row (4). Standard errors are reported in parentheses.

6 Sampling Variance of 2SLS Estimates

- Here's equation (15) without controls and with the second-stage residual written out:

$$Y_i = \alpha + \lambda_{2SLS}\hat{D}_i + [\eta_i + \lambda_{2SLS}(D_i - \hat{D}_i)], \tag{18}$$

- 2SLS is OLS on this second-stage equation:

$$\hat{\lambda}_{2SLS} = \frac{\sum Y_i(\hat{D}_i - \bar{D})}{\sum(\hat{D}_i - \bar{D})^2},$$

Substituting for Y_i:

$$\begin{aligned}\hat{\lambda}_{2SLS} &= \lambda_{2SLS}\frac{\sum \hat{D}_i(\hat{D}_i - \bar{D})}{\sum(\hat{D}_i - \bar{D})^2} + \frac{\sum \hat{D}_i\eta_i}{\sum(\hat{D}_i - \bar{D})^2} + \lambda_{2SLS}\frac{\sum \hat{D}_i(D_i - \hat{D}_i)}{\sum(\hat{D}_i - \bar{D})^2} \\ &= \lambda_{2SLS} + \frac{\sum \hat{D}_i\eta_i}{\sum(\hat{D}_i - \bar{D})^2}\end{aligned} \tag{19}$$

- When $\hat{D}_i$ is the fitted value estimated in your sample, the last term in the first line above is exactly zero in your sample (why?)

$\frac{\sum \hat{D}_i(D_i - \hat{D}_i)}{\sum(\hat{D}_i - \bar{D})^2} = \frac{\sum \hat{D}_i(e_{1i})}{\sum(\hat{D}_i - \bar{D})^2} = 0$ as fitted values are uncorrelated with residuals

- Assuming η_i is homoscedastic with variance σ_η^2, the asymptotic standard error of $\hat{\lambda}_{2SLS}$ is therefore:

$$SE(\hat{\lambda}_{2SLS}) = \frac{1}{\sqrt{n}} \frac{\sigma_\eta}{\sigma_{\hat{D}}}$$

  where σ_η is the std dev of residual η_i and $\sigma_{\hat{D}}$ is the std dev of first-stage fitted values, $\hat{D}_i$

Notes

- The standard errors generated by OLS estimation of (18) are wrong (why?)
  - Stata `ivregress` gets 'em right
- $SE(\hat{\lambda}_{2SLS})$ is an asymptotic formula, derived under something like classical assumptions, but even given these assumptions, valid only in large samples
  - Robust, clustered, and Newey-West standard errors for 2SLS are known to Stata (again, valid only in large samples)
  - For more on 2SLS inference, see the MM Chapter 3 appendix and MHE Chapter 4
- We can say only that $\hat{\lambda}_{2SLS}$ is consistent and asymptotically Normally distributed; as a rule 2SLS estimates are biased
  - The bias of 2SLS is proportional to the number of instruments in an over-identified model and inversely proportional to the F statistic that tests instrument relevant in the first stage
    - With many weak instruments, 2SLS estimates are likely to be misleadingly close to the corresponding OLS estimates
    - Given a reasonably strong first stage, *just-identified* 2SLS estimates (one instrument for one endogenous regressor) are approximately unbiased

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 12

FE and ME, Mastered by IV

This note recounts a 'metrics drama in three acts (see also MM Section 6.2 and the appendix to Chapter 6). First, we see how data on siblings can be used to control for omitted variables bias in estimates of the economic returns to schooling. The key idea here is to use *panel data* to control for *unobserved individual effects*, also known as "fixed effects" (FEs). Their invisibility notwithstanding, the fixedness of these effects allows us to control for them. Act II reveals, however, that the news is not all good: *attenuation bias* due to measurement error (ME) tends to shrink regression coefficients towards zero, and attenuation bias is greatly aggravated in regression models with fixed effects. Models with fixed effects may therefore suggest the returns to schooling are low simply because schooling is measured imperfectly. Finally, Act III shows how instrumental variables methods resolve the FE & ME conundrum.

1 Fixed Effects: Twins Double the Fun

Twinsburg (Ohio) embraces its zygotic heritage with an annual Twins Festival. Not wanting to miss the party, labor economists use exotic zygotic data from the Twins Festival to control for OVB.

- The long regression that motivates a twins analysis of the economic returns to schooling can be written:

$$\ln Y_{if} = \alpha^l + \rho^l S_{if} + \lambda A_{if} + e^l_{if}. \tag{1}$$

  Here, subscript f stands for family, while subscript $i = 1, 2$ indexes twin siblings, say Karen and Sharon or Ronald and Donald.

- Control variable A_{if} is a measure of ability, motivation, or talent; conditional on this, we'd be prepared to assume that schooling, S_{if}, is as good as randomly assigned.
  - Alas, A_{if} is not collected in the Current Population Survey.

- Since Ronald and Donald have the same parents, were mostly raised together, and may even have the same genes, we might reasonably assume $A_{if} = A_f$. In other words, ability is a family *fixed effect*. Given this fixedness, we can write:

$$\ln Y_{1f} = \alpha^l + \rho^l S_{1f} + \lambda A_f + e^l_{1f}$$
$$\ln Y_{2f} = \alpha^l + \rho^l S_{2f} + \lambda A_f + e^l_{2f}.$$

  Subtracting the equation for Donald from that for Ronald gives:

$$\ln Y_{1f} - \ln Y_{2f} = \rho^l (S_{1f} - S_{2f}) + (e^l_{1f} - e^l_{2f}), \tag{2}$$

  a differenced regression model that captures the coefficient of interest and from which unobserved ability disappears!
  - From this we learn that when unobserved ability is constant within twin pairs, a regression of the *difference* in twins' earnings on the *difference* in their schooling recovers the long regression coefficient, ρ^l.

- Column 1 in MM Table 6.2 reports estimates of a short regression in levels ("short" because the model omits A_{if}; "levels" because the model isn't differenced):

$$\ln Y_{if} = \alpha^s + \gamma' X_f + \rho^s S_{if} + e^s_{if}. \tag{3}$$

This model controls for age, race, and sex in vector X_f. Estimates of the differenced equation (2) appear in column 2 (why does X_f disappear from equation 2?) same age, race and sex for twins

TABLE 6.2
Returns to schooling for Twinsburg twins

	Dependent variable			
	Log wage (1)	Difference in log wage (2)	Log wage (3)	Difference in log wage (4)
Years of education	.110 (.010)		.116 (.011)	
Difference in years of education		.062 (.020)		.108 (.034)
Age	.104 (.012)		.104 (.012)	
Age squared/100	−.106 (.015)		−.106 (.015)	
Dummy for female	−.318 (.040)		−.316 (.040)	
Dummy for white	−.100 (.068)		−.098 (.068)	
Instrument education with twin report	No	No	Yes	Yes
Sample size	680	340	680	340

Notes: This table reports estimates of the returns to schooling for Twinsburg twins. Column (1) shows OLS estimates from models estimated in levels. OLS estimates of models for cross-twin differences appear in column (2). Column (3) reports 2SLS estimates of a levels regression using sibling reports as instruments for

- The estimate of just over 6% in the differenced equation (reported in column 2 of Table 6.2) is substantially below the estimate of 11% in column 1. This decline suggests the short-regression estimate of ρ^s indeed suffers muchly from ability bias

2 Measurement Error Messes Things Up

Of 340 twin pairs interviewed for the Ashenfelter and Rouse (1998) study, about half report *identical* educational attainment.

- If my brother and I are so similar, why then should our schooling differ? Good question! (My middle brother, Misha, has a Ph.D. just like me - and we're not twins.)

- Yet, if most twins really have the same schooling, then a fair number of the non-zero differences in *reported* schooling may reflect misleading or mistaken reports (Misha may not *tell* you he has a a Ph.D. - he doesn't remember his graduate work fondly)
- Master of 'metrics refer to mistakes and misreporting in data as *measurement error*. Most people probably report their schooling correctly, but a few get it wrong. The fact that a few people report their schooling incorrectly sounds unimportant. Yet, when it comes to regression, the consequences of even minor mismeasurement can be major.
- And then, there's this: Mismeasured schooling affects (2) much more than it does (1).

Interlude: Attenuation Bias

Let's simplify for a moment: forget ability bias and twins, focus only on measurement. Suppose you've dreamed of running the regression:

$$Y_i = \alpha + \beta S_i^* + e_i, \tag{4}$$

but data on S_i^*, the regressor of your dreams, are unavailable.

- You see only a mismeasured version, S_i:

$$S_i = S_i^* + u_i, \tag{5}$$

  where u_i is the measurement error in S_i

- Assume that:

$$E[u_i] = 0 \tag{6}$$
$$C(S_i^*, u_i) = C(e_i, u_i) = 0 \tag{7}$$

  These assumptions are said to describe "classical measurement error" . Note that the first part of (7) implies:

$$V(S_i) = V(S_i^*) + V(u_i).$$

- The regression coefficient we're after, β in (4), is given by:

$$\beta = \frac{C(Y_i, S_i^*)}{V(S_i^*)}. \tag{8}$$

  Alas, we see only the mismeasured regressor, S_i, instead of S_i^*. Regressing Y_i on S_i yields slope coefficient:

$$\begin{aligned}\beta_b &= \frac{C(Y_i, S_i)}{V(S_i)} \\ &= \frac{C(\alpha + \beta S_i^* + e_i, S_i^* + u_i)}{V(S_i)} \\ &= \frac{C(\alpha + \beta S_i^* + e_i, S_i^*)}{V(S_i)} = \beta \frac{V(S_i^*)}{V(S_i)}\end{aligned}$$

The 3rd equals above uses the classical assumptions, (7); be sure you can see how.

- We can know write:

$$\beta_b = r\beta, \tag{9}$$

where

$$r = \frac{V\left(S_i^*\right)}{V\left(S_i\right)} = \frac{V\left(S_i^*\right)}{V\left(S_i^*\right) + V\left(u_i\right)},$$

is a number between zero and one

  - Fraction r is called the *reliability* of S_i (sometimes also the "signal-to-noise ratio.")
  - Reliability reveals the extent of proportional *attenuation bias* in β_b:

$$\frac{\beta_b}{\beta} = r$$

  - β_b is closer to zero than β unless $r = 1$ (in which case, there's no measurement error after all)

Covariates and Differencing Aggravate Attenuation Bias

The addition of covariates to a model with mismeasured regressors exacerbates attenuation bias.

- Suppose the regression of interest is:

$$Y_i = \alpha + \gamma X_i + \beta S_i^* + e_i, \tag{10}$$

where X_i is a control variable, perhaps IQ or a test score. Regression anatomy says:

$$\beta = \frac{C(Y_i, \widetilde{S}_i^*)}{V(\widetilde{S}_i^*)},$$

where $\widetilde{S}_i^*$ is the residual from a regression of S_i^* on X_i

- Replacing S_i^* with S_i in (10), the coefficient on S_i becomes:

$$\beta_b = \frac{C(Y_i, \widetilde{S}_i)}{V(\widetilde{S}_i)},$$

where $\widetilde{S}_i$ is the residual from a regression of S_i on X_i

- In models with covariates, it's customary to assume measurement error and covs are uncorrelated, that is, $E[X_i u_i] = 0$ (we've already assumed $E[S_i^* u_i] = 0$, so this seems a natural extension). Given this assumption, we have:

$$\widetilde{S}_i = \widetilde{S}_i^* + u_i, \tag{11}$$

where u_i and $\widetilde{S}_i^*$ are uncorrelated with each other (show this). We therefore have:

$$V(\widetilde{S}_i) = V(\widetilde{S}_i^*) + V(u_i).$$

Note also that $V(\widetilde{S}_i^*) < V\left(S_i^*\right)$ when covariates predict true schooling, as seems likely

- Applying the same logic used to establish (9), we get:

$$\begin{aligned}\beta_b &= \frac{C(Y_i, \widetilde{S}_i)}{V(\widetilde{S}_i)} \\ &= \frac{V(\widetilde{S}_i^*)}{V(\widetilde{S}_i^*) + V(u_i)}\beta = \tilde{r}\beta,\end{aligned} \tag{12}$$

  where

$$\tilde{r} = \frac{V(\widetilde{S}_i^*)}{V(\widetilde{S}_i^*) + V(u_i)} < \frac{V(S_i^*)}{V(S_i^*) + V(u_i)} = r.$$

  Because covariates reduce the variance of the signal in S_i, while leaving the variance of the noise unchanged, ***covariates aggravate attenuation bias.***

- ***Fixed effects are a worst-case scenario for covariate-aggravated attenuation bias***
  - To see why, consider a panel model for the effects of true schooling:

$$Y_{if} = \alpha_f + \beta S_{if}^* + e_{if}, \tag{13}$$

    where $\alpha_f = \alpha^l + \lambda A_f$ and A_f is unobserved ability, as before
  - As noted in Act I, we can eliminate the fixed effect by differencing:

$$Y_{1f} - Y_{2f} = \beta\left(S_{1f}^* - S_{2f}^*\right) + e_{1f} - e_{2f}, \tag{14}$$

- In this scenario, we might imagine that true schooling is also similar within families, so that within-twin differences are mostly noise. We can describe this extreme scenario by modeling observed schooling in the twins panel as:

$$S_{if} = S_f^* + u_{if} \tag{15}$$

  where S_f^* is true schooling, fixed within families, and u_{if} is twin-specific reporting error.
  - In this extreme case, the observed difference in schooling is *entirely* noise:

$$S_{1f} - S_{2f} = u_{1f} - u_{2f} \tag{16}$$

    (what, then, will we get from estimation of differenced equation (2)?)
  - In practice, $S_{1f} - S_{2f}$ is probably not *all* noise. More realistically, we have:

$$S_{1f} - S_{2f} = (S_{1f}^* - S_{2f}^*) + (u_{1f} - u_{2f}). \tag{17}$$

    Even so, because S_{1f}^* and S_{2f}^* are so similar, the difference between them, $S_{1f}^* - S_{2f}^*$, has variance well below that of the difference in measured schooling, $S_{1f} - S_{2f}$

- Attenuation bias in differenced equation (2) is likely worse than attenuation bias in the levels equation (3). Aggravated attenuation bias provides an alternative explanation (besides ability bias) for the sharp decline in schooling coefficients as we move from column 1 to column 2 in Table 6.2

3 IV to the Rescue

Perhaps attenuation bias makes the differencing cure (for ability bias) worse than the disease. But all is not lost.

- Recall from LN11 that the IV estimator of the coefficient on S_i in a bivariate regression of Y_i on S_i is the sample analog of:

$$\beta_{IV} = \frac{C(Y_i, Z_i)}{C(S_i, Z_i)}, \tag{18}$$

  where the instrumental variable is Z_i.

  - In the measurement error version of the IV story, we use Z_i to instrument for mismeasured S_i
  - In the context of an equation like (4), this works when *Z_i is correlated with S_i^*, but uncorrelated with both measurement error, u_i, and the residual, e_i*

- To see how IV gives us what we want, use (4) and (5) to substitute for Y_i and S_i in (18):

$$\begin{aligned}\beta_{IV} &= \frac{C(Y_i, Z_i)}{C(S_i, Z_i)} = \frac{C(\alpha + \beta S_i^* + e_i, Z_i)}{C(S_i^* + u_i, Z_i)} \\ &= \frac{\beta C(S_i^*, Z_i) + C(e_i, Z_i)}{C(S_i^*, Z_i) + C(u_i, Z_i)}.\end{aligned}$$

- Assuming $C(e_i, Z_i) = C(u_i, Z_i) = 0$, we have:

$$\beta_{IV} = \beta \frac{C(S_i^*, Z_i)}{C(S_i^*, Z_i)} = \beta.$$

  Attenuation bias begone!

- IV solutions to measurement error problems often exploit multiple measures of the same underlying construct. If only, we had two measures of schooling!

  - We do: the Twinsburg sample survey asked each twin to report not only his or her own schooling but also that of their sibling. We therefore have two measures of schooling for each twin, one self-report and one sibling report.

- This extra info is especially valuable for the measurement-error-afflicted differenced equation. Assuming the measurement error in self- and sibling-reports is uncorrelated (i.e., mistakes I make in reporting my own schooling are uncorrelated with mistakes my sibling makes in reporting my schooling), the difference in sibling reports can be used to instrument the difference in self-reports in equation (2)

  - In the IV formula, (18), the variable to be instrumented is

$$S_i \equiv (S_{1f} - S_{2f}),$$

    while $Y_i \equiv (\ln Y_{1f} - \ln Y_{2f})$,
  - The instrument is

$$Z_i \equiv \left(S_{1f}^2 - S_{2f}^1\right),$$

    where S_{if}^j is sibling j's report of sibling i's schooling

- **The resulting IV estimates, reported in cols 3-4 in Table 6.2, suggest the decline in returns to schooling from columns 1 to 2 is indeed due to ME rather than OVB.**

  And so the curtain falls on our story.

MIT 14.32/14.320 **Spring 2021**
J. Angrist

Lecture Note 13

Doing Differences-in-Differences (DD)

1 Oh No, Not Another Liquidity Crisis!

On the eve of the Great Depression, Caldwell and Company was the largest Southern banking chain. Alas, in November 1930, mismanagement and the October 1929 stock market crash brought the Caldwell empire down. Within days, Caldwell's collapse felled closely tied banking networks in Tennessee, Arkansas, Illinois, and North Carolina, precipitating a run on Mississippi banks in December 1930.

Policymakers facing a bank have a choice: open the flow of credit or turn off the tap. On one hand, lending to troubled banks may allow them to meet increasingly urgent withdrawal demands and stave off depositor panic. On the other hand, support for bad banks raises the specter of moral hazard. If bankers know that the central bank will lend cheaply during a crisis, they needn't take care to avoid crises in the first place. Which strategy is better?

Richardson and Troost (2009) tackle this question by exploiting the differential response of regional Federal Reserve Banks to the Caldwell collapse:

- The 6th (Atlanta) District favored lending to troubled banks and increased bank lending by about 40% within 4 weeks of Caldwell's collapse
- The 8th (St. Louis) District decreased bank lending by 10% in the same period
- As it happens, the 6th/8th District border runs smack through the middle of Miss., so we get a well-controlled within-state contrast in lending policy
- Think of the 8th as a passive control group, and the 6th as the treatment group, experimenting with increased lending
- 8 months into the crisis:
  - 132 banks were open in the 8th
  - 121 banks were open in the 6th
  - A deficit of 11 banks in the 6th*: The easy-money treatment effect is negative!*
- Look again - On July 1, 1930 (*before* the Caldwell crisis):
  - 165 banks were open in the 8th
  - 135 banks were open in the 6th

1.1 Parallel worlds

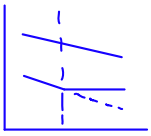
DD uses parallelism to adjust for differences across districts in the pre-treatment period:

- Let $Y_{d,t}$ denote the number of banks observed operating in District d in year t

– The DD effect of loose money in the 6th (treatment) District during the Caldwell crisis compares post-crisis differences with the baseline difference between the two districts:

$$\begin{aligned}\delta_{DD} &= (Y_{6,1931} - Y_{8,1931}) - (Y_{6,1930} - Y_{8,1930}) \\ &= (121 - 132) - (135 - 165) \\ &= -11 - (-30) = 19.\end{aligned}$$

– Equivalently, DD contrasts the *change* in the number of banks operating in the two districts:



$$\begin{aligned}\delta_{DD} &= (Y_{6,1931} - Y_{6,1930}) - (Y_{8,1931} - Y_{8,1930}) \\ &= (121 - 135) - (132 - 165) \\ &= -14 - (-33) = 19.\end{aligned}$$

– 19 more banks failed in the 8th, yo

- Either way you look at it, DD controls confounding from fixed (and therefore pre-treatment) differences in levels that arise even in the absence of treatment

1.1.1 Parallelism Meets Potentials

- The heart of the DD setup is an *additive model for potential outcomes* in the no-treatment state:

$$\mathrm{Y}_{d,t}(0) = \beta_d + \gamma_t \tag{1}$$

where $\mathrm{Y}_{d,t}(0)$ is notation for the *potential outcome* describing what happens in district d and period t in the absence of an intervention (defined for all d and t)

– Parameter β_d is called a "district effect"; parameter γ_t is called a "year effect"

– Assuming that the causal effect $\mathrm{Y}_{d,t}(1) - \mathrm{Y}_{d,t}(0)$ is constant:

$$\mathrm{Y}_{d,t}(1) = \beta_d + \gamma_t + \delta_{DD}.$$

So,

6 is treated
8th is not

$$\mathrm{Y}_{6,1931} - \mathrm{Y}_{6,1930} = (\gamma_{1931} + \delta_{DD}) - \gamma_{1930}$$ Time difference plus causal effect

$$\mathrm{Y}_{8,1931} - \mathrm{Y}_{8,1930} = \gamma_{1931} - \gamma_{1930}$$ – Time difference

Causal Effect

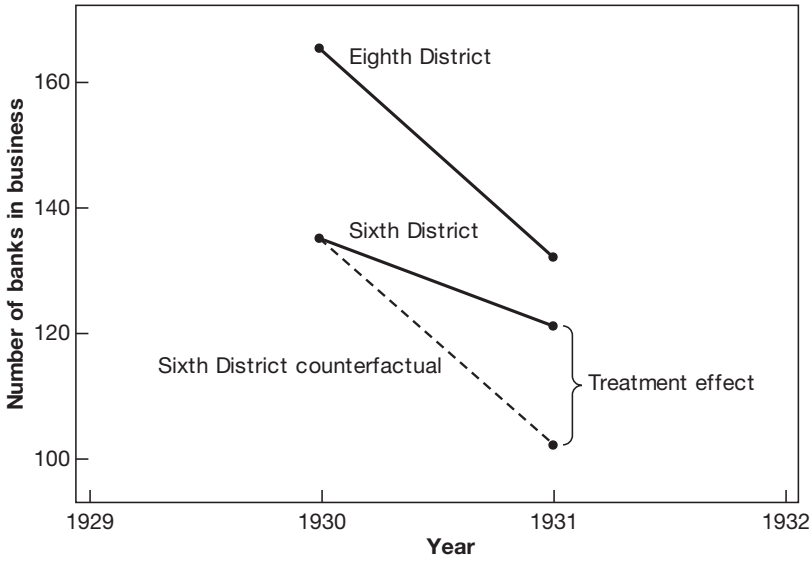
– The double-diff (DD) captures causal effect δ:

$$\begin{aligned}&\{\mathrm{Y}_{6,1931} - \mathrm{Y}_{6,1930}\} - \{\mathrm{Y}_{8,1931} - \mathrm{Y}_{8,1930}\} \\ &= (Y_{6,1931} - Y_{8,1931}) - (Y_{6,1930} - Y_{8,1930}) = \delta_{DD}\end{aligned}$$

Basically difference between outcome had treatment not occurred and outcome with treatment.
Use parallel trend to estimate counterfactual w/o treatment outcome

- Every picture tells a story

FIGURE 5.1
Bank failures in the Sixth and Eighth Federal Reserve Districts



Notes: This figure shows the number of banks in operation in Mississippi in the Sixth and Eighth Federal Reserve Districts in 1930 and 1931. The dashed line depicts the counterfactual evolution of the number of banks in the Sixth District if the same number of banks had failed in that district in this period as did in the Eighth.

TABLE 5.1
Wholesale firm failures and sales in 1929 and 1933

	1929	1933	Difference (1933–1929)
Panel A. Number of wholesale firms			
Sixth Federal Reserve District (Atlanta)	783	641	−142
Eighth Federal Reserve District (St. Louis)	930	607	−323
Difference (Sixth–Eighth)	−147	34	181
Panel B. Net wholesale sales ($ million)			
Sixth District Federal Reserve (Atlanta)	141	60	−81
Eighth District Federal Reserve (St. Louis)	245	83	−162
Difference (Sixth–Eighth)	−104	−23	81

Notes: This table presents a DD analysis of Federal Reserve liquidity effects on the number of wholesale firms and the dollar value of their sales, paralleling the DD analysis of liquidity effects on bank activity in Figure 5.1.

1.2 Common Trends

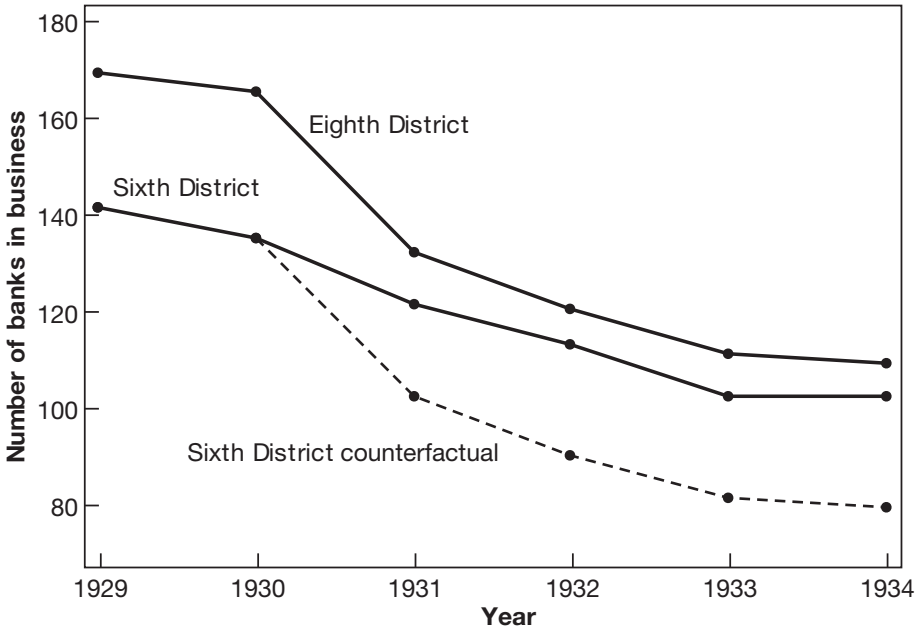
- The DD model for counterfactual no-treatment outcomes in both districts allows for:
  - Time-invariant district effects, β_d
  - Common period (year) effects, γ_t
- The key DD assumption is parallel outcome trends in treatment and control units (districts)
- Common trends can be applied to transformed data, e.g.,

$$\log Y_{d,t}(0) = \beta_d + \gamma_t$$

  But common trends in logs does not imply (indeed, contradicts) common trends in levels
  - DD identification is fickle

- Luckily, Mis-sis-sippi is DD heaven:

FIGURE 5.3
Trends in bank failures in the Sixth and Eighth Federal Reserve Districts, and the Sixth District's DD counterfactual



Notes: This figure adds DD counterfactual outcomes to the banking data plotted in Figure 5.2. The dashed line depicts the counterfactual evolution of the number of banks in the Sixth District if the same number of banks had failed in that district after 1930 as did in the Eighth.

1.3 Regression DD Heads South

- Stack data on districts and years in a sample of size 12 (with 6 years for each district); call this Y_{dt} for the number of banks operating in district d in year t
- Let $TREAT_d$ indicate data from the 6th District and let $POST_t$ indicate post-treatment years (Dummy Variables)
- The regression DD estimator, δ_{rDD}, comes from fitting:

$$\begin{aligned} Y_{dt} &= \alpha + \beta TREAT_d + \gamma POST_t \\ &\quad + \delta_{rDD}(TREAT_d \times POST_t) + e_{dt} \end{aligned}$$

$Y_{dt} - (\beta TREAT_d + \gamma POST_t + \alpha + e_{dt}) = \delta_{rDD}(TREAT_d \times POST_t)$

1 if both are true i.e. observation is treated

- With only two periods, estimates of δ_{DD} and δ_{rDD} coincide. With more, δ_{rDD} is more precise than the simple four-number DD recipe (using the data in Fig 5.3 generates an estimate of 21 banks saved, with a standard error of about 11)
- Regression DD
  - generates SEs (beware of clustering and serial correlation)
  - facilitates specification testing
- Compare DD policy analysis with the fixed effects panel-data estimator we used to estimate the returns to schooling in LN12:
  - Same econometric idea (differencing eliminates unobserved individual effects, here, for districts; before, for twins)
  - Earlier, we analyzed a large sample of microdata; policy DD often uses aggregate data such as for states and regions (this complicates statistical inference)

2 Does the Min Matter?

- On January 1 2018, 18 states and 20 cities raised their minimum wages (Pres. Biden just hiked the min to \$15/hour for federal contractors; Dems hope to make this universal)
  - Nice work if you can get it! But can you indeed get it – that's the \$30,000 question.
  - What does history teach us?
- On April 1, 1992, New Jersey imposed a state minimum wage of \$5.05. The Federal minimum wage was then \$4.25
  - Card and Krueger (1994) surveyed fast food restaurants in NJ and Eastern PA before this change (February 1992) and after (November 1992).
  - A DD classic!

TABLE 2—MEANS OF KEY VARIABLES

Variable	Stores in: NJ	PA	t[a]
1. *Distribution of Store Types (percentages):*			
a. Burger King	41.1	44.3	−0.5
b. KFC	20.5	15.2	1.2
c. Roy Rogers	24.8	21.5	0.6
d. Wendy's	13.6	19.0	−1.1
e. Company-owned	34.1	35.4	−0.2
2. *Means in Wave 1:*			
a. FTE employment	20.4	23.3	−2.0
	(0.51)	(1.35)	
b. Percentage full-time employees	32.8	35.0	−0.7
	(1.3)	(2.7)	
c. Starting wage	4.61	4.63	−0.4
	(0.02)	(0.04)	
d. Wage = $4.25 (percentage)	30.5	32.9	−0.4
	(2.5)	(5.3)	
e. Price of full meal	3.35	3.04	4.0
	(0.04)	(0.07)	
f. Hours open (weekday)	14.4	14.5	−0.3
	(0.2)	(0.3)	
g. Recruiting bonus	23.6	29.1	−1.0
	(2.3)	(5.1)	
3. *Means in Wave 2:*			
a. FTE employment	21.0	21.2	−0.2
	(0.52)	(0.94)	
b. Percentage full-time employees	35.9	30.4	1.8
	(1.4)	(2.8)	
c. Starting wage	5.08	4.62	10.8
	(0.01)	(0.04)	
d. Wage = $4.25 (percentage)	0.0	25.3	—
		(4.9)	
e. Wage = $5.05 (percentage)	85.2	1.3	36.1
	(2.0)	(1.3)	
f. Price of full meal	3.41	3.03	5.0
	(0.04)	(0.07)	
g. Hours open (weekday)	14.4	14.7	−0.8
	(0.2)	(0.3)	
h. Recruiting bonus	20.3	23.4	−0.6
	(2.3)	(4.9)	

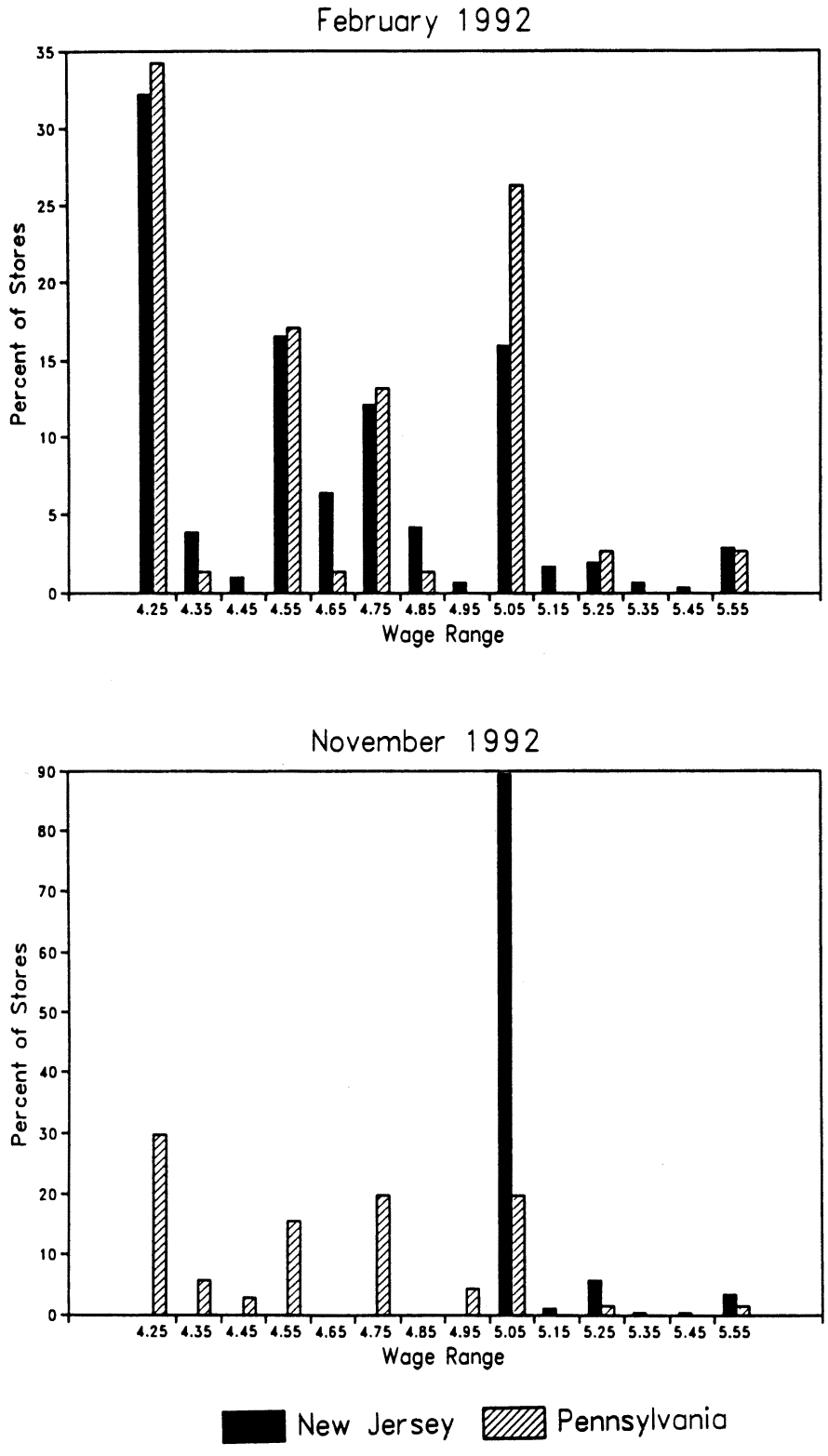


FIGURE 1. DISTRIBUTION OF STARTING WAGE RATES

- The simplest DD analysis involves just 4 numbers, repeated in MHE Table 5.2.1

TABLE 5.2.1
Average employment in fast food restaurants before and after the New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ – PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (.94)	21.03 (.52)	−.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	.59 (.54)	2.76 (1.36)

Notes: Adapted from Card and Krueger (1994), table 3. The table reports average full-time-equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all restaurants with data on employment. Employment at six closed restaurants is set to zero. Employment at four temporarily closed restaurants is treated as missing. Standard errors are reported in parentheses.

3 Regression DD Reprise

- Stack the data from the two waves of the CK survey and let y_{ist} be employment in restaurant i in state s in year t; code dummy variables as follows:

$$\begin{aligned} NJ_s &= 1 && \text{if s = New Jersey (= 0 for PA)} \\ D_t &= 1 && \text{if t = 2 (= 0 in wave 1)} \\ M_{st} &= 1 && \text{if s = NJ and t=2 (=0 for all other cases)} \end{aligned}$$

- In other words, $M_{st} = NJ_s \times D_t$
- Regress employment at restaurant i in state s and period t on these three variables:

$$Y_{ist} = \alpha + \beta NJ_s + \gamma D_t + \delta_{rDD} M_{st} + \epsilon_{ist} \tag{2}$$

↳ Indicates treated observations $NJ_s \times D_t$

- DD regression talk
  - The NJ_s coefficient is a *state effect*
  - The D_t is a *time effect*
  - The variable $M_{st} = NJ_s \times D_t$ is an *interaction term*

3.1 DD in Practice

- Estimates based on (2) control for:
  1. fixed differences between states (NJ_s in equation (2))
  2. common period effects (D_t in equation (2))
- State-time interactions (differences in employment trends) are then seen as causal effects

Simple DD lives or dies on the common trends assumption

Alas, reality is uncommonly messy:

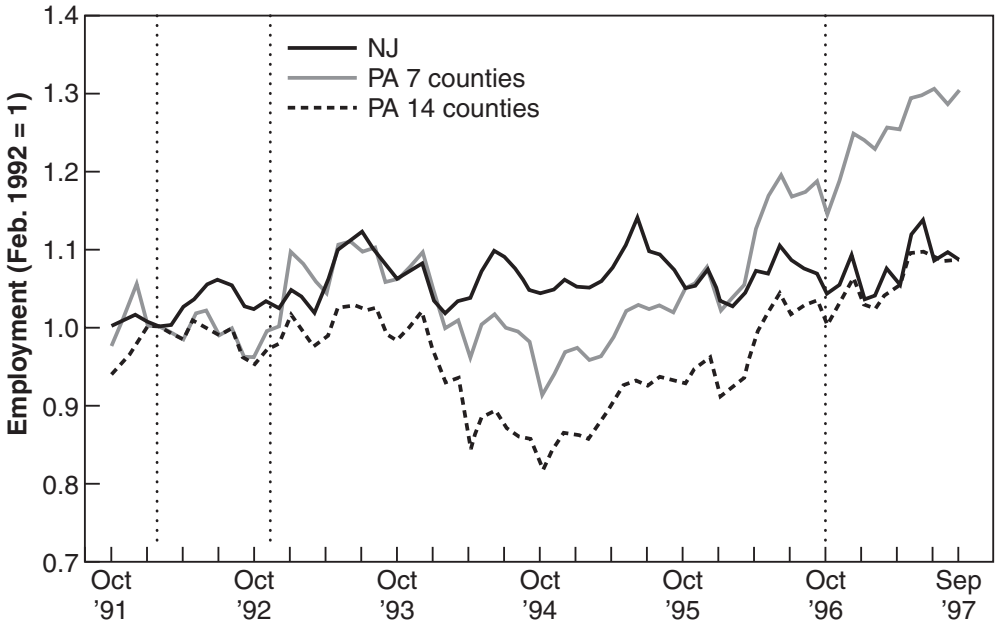


Figure 5.2.2 Employment in New Jersey and Pennsylvania fast food restaurants, October 1991 to September 1997 (from Card and Krueger 2000). Vertical lines indicate dates of the original Card and Krueger (1994) survey and the October 1996 federal minimum

- But we can look within states too

TABLE 3—AVERAGE EMPLOYMENT PER STORE BEFORE AND AFTER THE RISE IN NEW JERSEY MINIMUM WAGE

Variable	Stores by state			Stores in New Jersey[a]			Differences within NJ[b]	
	PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Wage = $4.25 (iv)	Wage = $4.26–$4.99 (v)	Wage ≥ $5.00 (vi)	Low–high (vii)	Midrange–high (viii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	−2.69 (1.37)	−2.17 (1.41)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)	0.75 (1.27)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)	3.36 (1.48)	2.91 (1.41)
4. Change in mean FTE employment, balanced sample of stores[c]	−2.28 (1.25)	0.47 (0.48)	2.75 (1.34)	1.21 (0.82)	0.71 (0.69)	−2.16 (1.01)	3.36 (1.30)	2.87 (1.22)
5. Change in mean FTE employment, setting FTE at temporarily closed stores to 0[d]	−2.28 (1.25)	0.23 (0.49)	2.51 (1.35)	0.90 (0.87)	0.49 (0.69)	−2.39 (1.02)	3.29 (1.34)	2.88 (1.23)

Notes: Standard errors are shown in parentheses. The sample consists of all stores with available data on employment. FTE (full-time-equivalent) employment counts each part-time worker as half a full-time worker. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing.

[a]Stores in New Jersey were classified by whether starting wage in wave 1 equals $4.25 per hour ($N = 101$), is between $4.26 and $4.99 per hour ($N = 140$), or is $5.00 per hour or higher ($N = 73$).

[b]Difference in employment between low-wage ($4.25 per hour) and high-wage (≥ $5.00 per hour) stores; and difference in employment between midrange ($4.26–$4.99 per hour) and high-wage stores.

4 Drink, Drank, Dead: DD With Many States and Years

Alcohol is the most widely abused intoxicant. How should it be regulated? MADD lobbies to limit access for youth, while college presidents call to liberalize. Who's right?

- State MLDAs running from age 18-21 generate up to three treatment effects relative to age 21: the effects of legal drinking at age 18, 19, and 20.
  - We capture this with a single variable called $LEGAL_{st}$, the fraction of 18-20 year olds allowed to drink in state s and year t. In some states, no one under 21 is allowed to drink, while in states with an age-19 MLDA, roughly two thirds of under-21 year olds can drink, and in states with an age-18 MLDA, all 18-21 year olds can drink.
- Because $LEGAL_{st}$ varies by both state and year, we can use a generalized DD model to capture its effects.
- Using data on death rates from 1971-84 in the 50 states plus DC (denoted M_{st}), a multi-state regression DD model looks like this:

Dummy for each state Dummy for each year

$$Y_{st} = \alpha + \delta_{rDD} LEGAL_{st} + \sum_{k=AL}^{WY} \beta_k STATE_{ks} + \sum_{j=1971}^{1983} \gamma_j YEAR_{jt} + e_{st}$$

Dummy variables $STATE_{ks}$ switch on when state k in the summation equals state s on the left hand side. State effects, β_k, are the coefficients on these dummies (DC is the reference state). Similarly, the year effects, γ_j, are coefficients on dummies $YEAR_{jt}$ that switch on when year j in the summation equals year t on the left hand side (1970 is the reference year)

– This can be written more compactly as

$$Y_{st} = \beta_s + \gamma_t + \delta_{rDD} LEGAL_{st} + e_{st},$$

where β_s is again a state effect and γ_t is again a year effect

- Estimates of δ_{rDD} suggest that legal alcohol access causes about 10 additional motor vehicle deaths among 18-20 year olds, of which about 7 are the result of motor vehicle accidents. The estimated MVA effect is reasonably precise, with a standard error of about 2.5. Regression DD generates little evidence of an impact of legal drinking on deaths from internal causes.

TABLE 5.2
Regression DD estimates of MLDA effects on death rates

Dependent variable	(1)	(2)	(3)	(4)
All deaths	10.80	8.47	12.41	9.65
	(4.59)	(5.10)	(4.60)	(4.64)
Motor vehicle accidents	7.59	6.64	7.50	6.46
	(2.50)	(2.66)	(2.27)	(2.24)
Suicide	.59	.47	1.49	1.26
	(.59)	(.79)	(.88)	(.89)
All internal causes	1.33	.08	1.89	1.28
	(1.59)	(1.93)	(1.78)	(1.45)
State trends	No	Yes	No	Yes
Weights	No	No	Yes	Yes

Notes: This table reports regression DD estimates of minimum legal drinking age (MLDA) effects on the death rates (per 100,000) of 18–20-year-olds. The table shows coefficients on the proportion of legal drinkers by state and year from models controlling for state and year effects. The models used to construct the estimates in columns (2) and (4) include state-specific linear time trends. Columns (3) and (4) show weighted least squares estimates, weighting by state population. The sample size is 714. Standard errors are reported in parentheses.

Regression DD Estimates of MLDA-Induced Deaths among 18-20 Year Olds, from 1970 - 1983

- Results here are close to those from an MLDA regression discontinuity design (as we'll soon see)

4.1 Worried About Common Trends? Relax, Have a Drink!

- With many states and years, we can relax the common trends assumption and allow separate linear trends for each state. Regression DD with state-specific trends looks like this:

$$Y_{st} = \alpha + \delta_{rDD} LEGAL_{st} + t\left[\sum_{k=AL}^{WY} \theta_k STATE_{ks}\right] + \sum_{k=AL}^{WY} \beta_k STATE_{ks} + \sum_{j=1971}^{1983} \gamma_j YEAR_{jt} + e_{st} \quad (3)$$

Coefficient θ_k captures the linear trend for state k

- We can also write

$$Y_{st} = \beta_s + \gamma_t + \theta_s t + \delta_{rDD} LEGAL_{st} + e_{st},$$

where state-specific trends are denoted θ_s

- Figs 5.4-5.6 shows how this works for the state of Allatsee, which reduced its MLDA to 18 in 1975 and neighboring Alabaster, which held the line at 21.
  - Fig. 5.4 is the ideal parallel trends scenario:

FIGURE 5.4
An MLDA effect in states with parallel trends

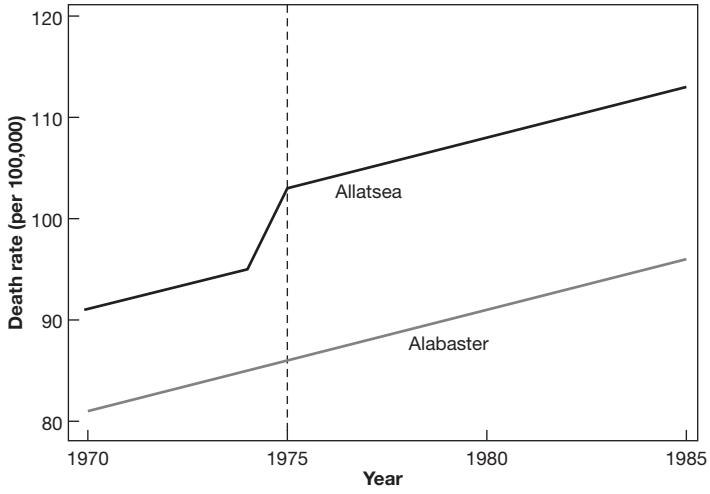
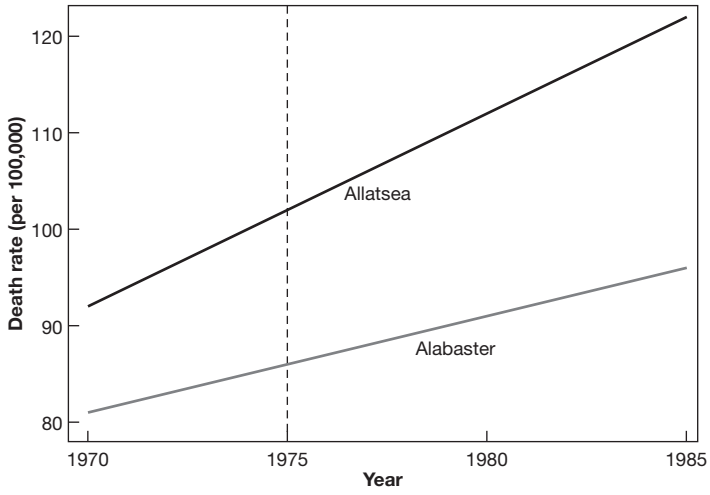


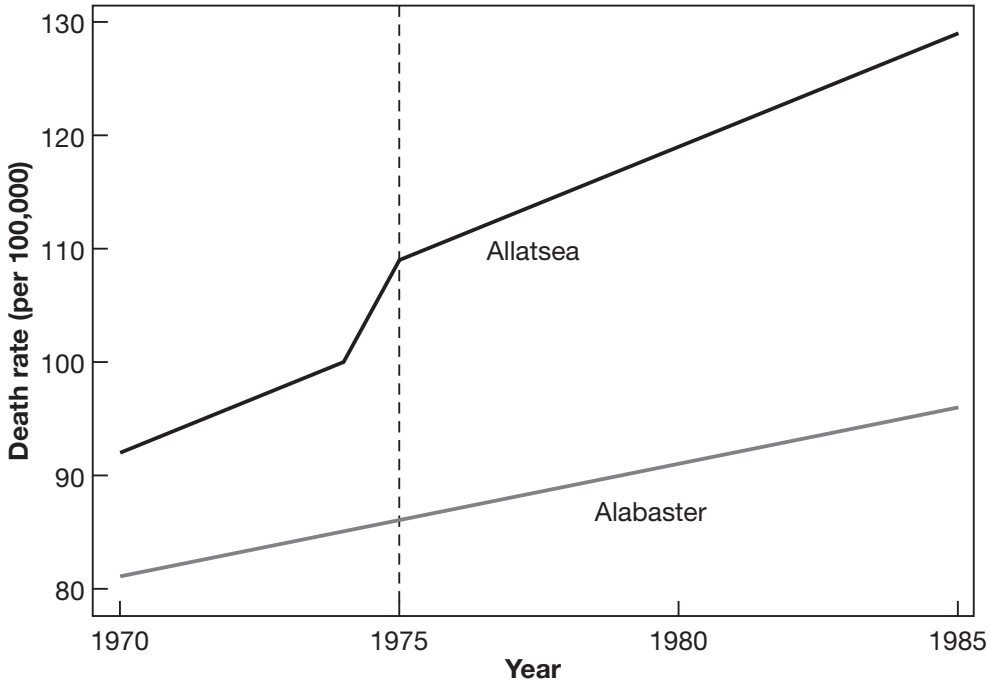
FIGURE 5.5
A spurious MLDA effect in states where trends are not parallel



- Fig. 5.5 is worrying: trends are unchanged in 1975, yet regression DD is likely to generate estimates suggesting an MLDA effect!

- Luckily, the trends here are linear. The model in (3) allows for this, recovering the correct treatment effect in the presence of state-specific trends
- This picture tells the story:

FIGURE 5.6
A real MLDA effect, visible even though trends are not parallel



- As it turns out, the MLDA estimates shown in Table 5.2 aren't highly sensitive to control for trends.
  - That's DD heaven!
- But weight! (Happily this doesn't matter much, either)
  - Read all about it MM Section 5.2

4.2 OVBeer Taxes

- State policy is a messy business, with frequent changes on many fronts
- An important consideration in research on drinking is the price of a drink; this price, it turns out, is mostly state taxes
- Data on many states and years allow us to explore the effects of similarly-timed policy innovations such as coincident tax changes and MLDA changes
- Regression DD results from models including state beer taxes appear in Table 5.3

TABLE 5.3
Regression DD estimates of MLDA effects controlling for beer taxes

	Without trends		With trends	
Dependent variable	Fraction legal (1)	Beer tax (2)	Fraction legal (3)	Beer tax (4)
All deaths	10.98	1.51	10.03	−5.52
	(4.69)	(9.07)	(4.92)	(32.24)
Motor vehicle accidents	7.59	3.82	6.89	26.88
	(2.56)	(5.40)	(2.66)	(20.12)
Suicide	.45	−3.05	.38	−12.13
	(.60)	(1.63)	(.77)	(8.82)
Internal causes	1.46	−1.36	.88	−10.31
	(1.61)	(3.07)	(1.81)	(11.64)

Notes: This table reports regression DD estimates of minimum legal drinking age (MLDA) effects on the death rates (per 100,000) of 18–20-year-olds, controlling for state beer taxes. The table shows coefficients on the proportion of legal drinkers by state and year and the beer tax by state and year, from models controlling for state and year effects. The fraction legal and beer tax variables are included in a single regression model, estimated without trends to produce the estimates in columns (1) and (2) and estimated with state-specific linear trends to produce the estimates in columns (3) and (4). The sample size is 700. Standard errors are reported in parentheses.

Getting a Little Jumpy: Rockin' RD

Master Joshway

MIT 14.32/14.320 LN14 (Spring 2021)

Doin' RD: Sharpish!

Natura non facit saltus - Leibniz, Darwin, Marshall

- RD methods come from the paradoxical idea that *rules* – which at first appear to reduce or even eliminate the scope for randomness – reveal causal effects
- Sharp RD designs arise when treatment, D_i, is a deterministic and discontinuous function of a covariate, x_i, sometimes called the *running variable*:

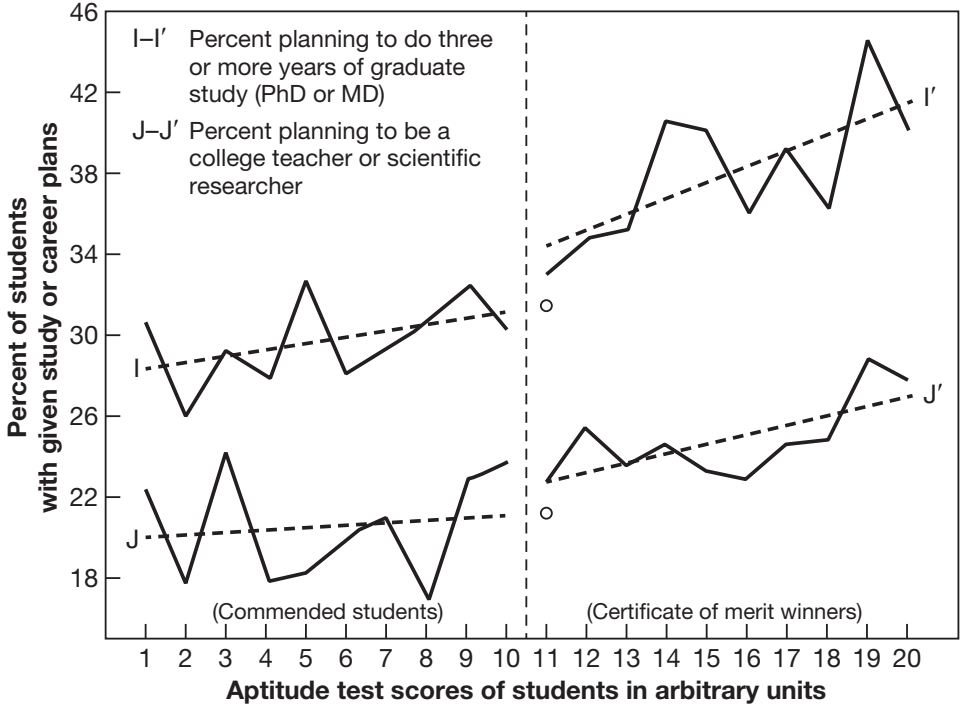
$$\mathrm{D}_i = \begin{cases} 1 & \text{if } x_i \geq x_0 \\ 0 & \text{if } x_i < x_0 \end{cases} \tag{1}$$

where x_0 is a known threshold or cutoff value

  - D_i is a *deterministic* function of x_i: if you know x_i, you know D_i
  - D_i is *discontinuous* because no matter how close x_i gets to x_0, treatment is unchanged until $x_i = x_0$
  - Relationships are otherwise smooth

RD is Born: Thistlethwaite and Campbell (1960)

FIGURE 4.10
Thistlethwaite and Campbell's Visual RD



Notes: This figure plots PSAT test takers' plans for graduate study (line I–I′) and a measure of test takers' career plans (line J–J′) against the running variable that determines National Merit recognition.

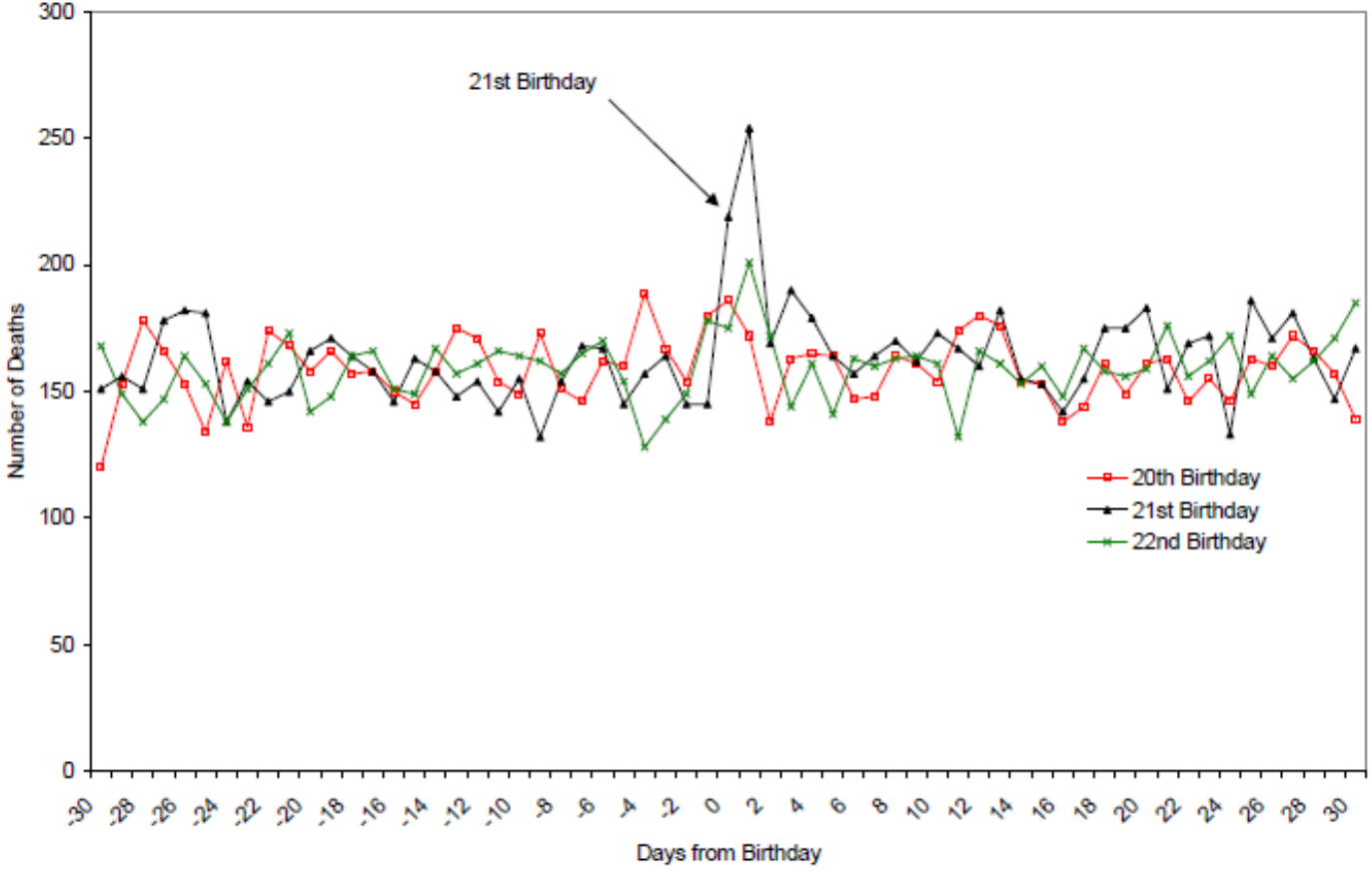
Birthdays and Funerals

Animal House Comes Home

Katy: Is this really what you're gonna do for the rest of your life?
Boon: What do you mean?
Katy: I mean hanging around with a bunch of animals getting drunk every weekend.
Boon: No! After I graduate, I'm gonna get drunk every night!

- **Binge drinking** killed MIT fraternity member Scott Krueger in 1997; a few years later, BU freshman fraternity pledge Anthony Barksdale died similarly
- Through the **Amethyst Initiative**, American college presidents lobby states to *reduce* the minimum legal drinking age (MLDA), arguing that legal drinking reduces pathological drinking
- Carpenter and Dobkin (2009, 2011) estimate the effect of the MLDA using DD and RD
  - As it turns out, **the MLDA saves lives**

Appendix A: Deaths by Days to Birthday



Doing Sharp RD

- Suppose potential outcomes can be described by a linear, constant-effects model

$$E[Y_{0i}|x_i] = \alpha + \beta x_i \quad (2)$$
$$Y_{1i} = Y_{0i} + \rho, \quad (3)$$

leading to the regression,

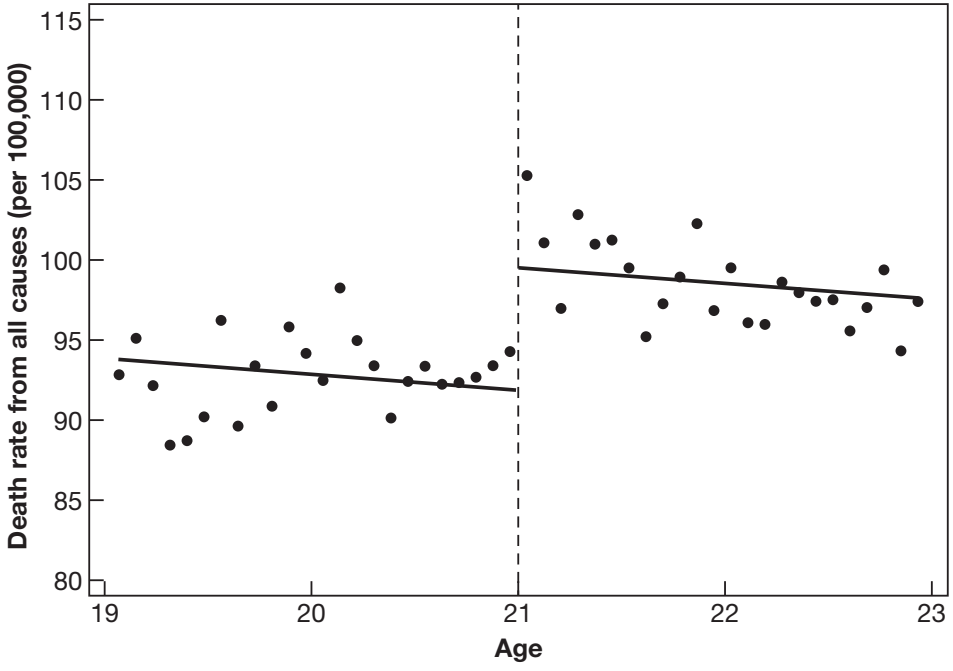
$$Y_i = \alpha + \beta x_i + \rho D_i + \eta_i, \quad (4)$$

where ρ is the causal effect of interest

- D_i is not only correlated with x_i, it's *determined* by x_i
- RD distinguishes the nonlinear and discontinuous function, $D_i = 1(x_i \geq x_0)$, from the smooth function, βx_i
- Given (1) and (2), there isn't – cannot be! – any OVB in OLS estimates of (4) [why?] D_i is uniquely determined by X_i

Picturing MLDA RD

FIGURE 4.2
A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the minimum legal drinking age (MLDA) cutoff).

Nonlinear Running Var Control

- **Figure 6.1.1** suggests causal effects are identified even if $E[Y_{0i}|x_i] = f(x_i)$ for a smooth but nonlinear function, $f(x_i)$
- In this case, we do RD by fitting:

$$Y_i = f(x_i) + \rho D_i + \eta_i \tag{5}$$

- Typically, $f(x_i)$ is polynomial,

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \rho D_i + \eta_i \tag{6}$$

- Slopes can differ on either side of the cutoff

$$\begin{aligned} Y_i &= \alpha + \beta_1 \tilde{x}_i + \beta_2 \tilde{x}_i^2 + ... + \beta_p \tilde{x}_i^p \\ &\quad + \rho D_i + \beta_1^* D_i \tilde{x}_i + \beta_2^* D_i \tilde{x}_i^2 + ... + \beta_p^* D_i \tilde{x}_i^p + \eta_i \end{aligned} \tag{7}$$

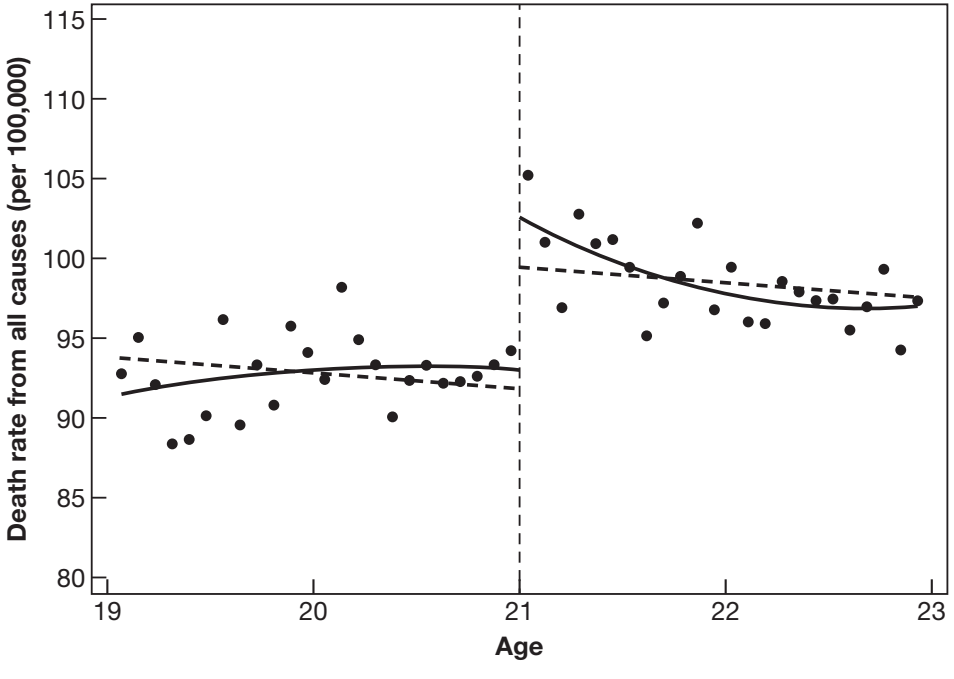
(Different coefficient for interaction term i.e. on the right side of the cut-off.)

where $\tilde{x}_i = x_i - x_0$ and β_j^* is the coefficient on the jth poly interaction

  - The interacted model implicitly extrapolates treatment effects away from the cutoff, but RD pros rarely use this

Quadratic Control

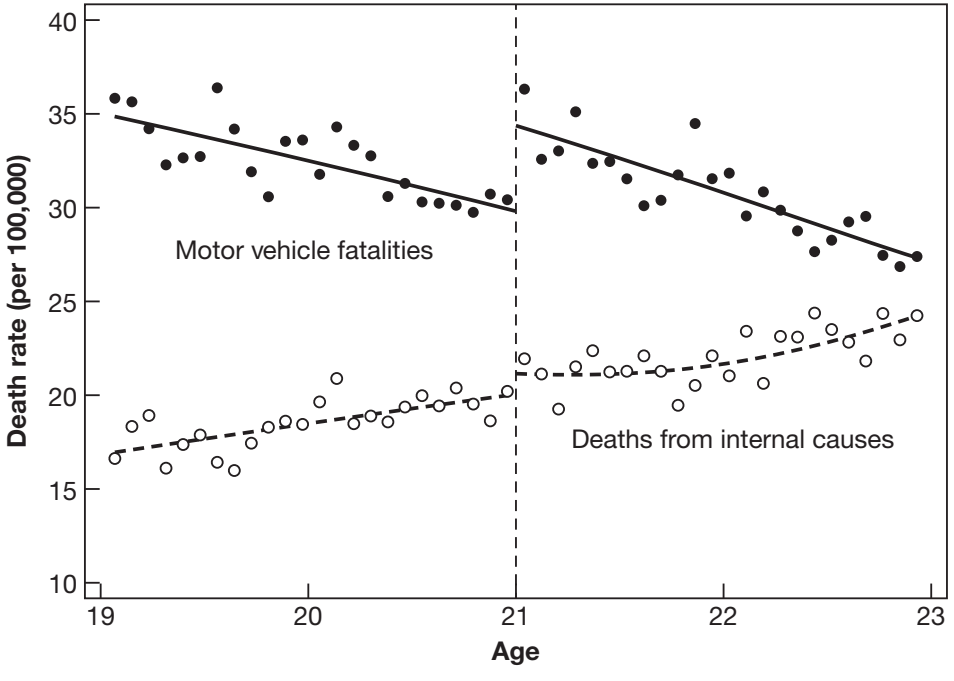
FIGURE 4.4
Quadratic control in an RD design



Notes: This figure plots death rates from all causes against age in months. Dashed lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months. The solid lines plot fitted values from a regression of mortality on an over-21 dummy and a quadratic in age, interacted with the over-21 dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

MVA vs Internal Causes

FIGURE 4.5
RD estimates of MLDA effects on mortality by cause of death



Notes: This figure plots death rates from motor vehicle accidents and internal causes against age in months. Lines in the figure plot fitted values from regressions of mortality by cause on an over-21 dummy and a quadratic function of age in months, interacted with the dummy (the vertical dashed line indicates the minimum legal drinking age [MLDA] cutoff).

TABLE 4.1
Sharp RD estimates of MLDA effects on mortality

Dependent variable	Ages 19–22		Ages 20–21	
	(1)	(2)	(3)	(4)
All deaths	7.66	9.55	9.75	9.61
	(1.51)	(1.83)	(2.06)	(2.29)
Motor vehicle accidents	4.53	4.66	4.76	5.89
	(.72)	(1.09)	(1.08)	(1.33)
Suicide	1.79	1.81	1.72	1.30
	(.50)	(.78)	(.73)	(1.14)
Homicide	.10	.20	.16	−.45
	(.45)	(.50)	(.59)	(.93)
Other external causes	.84	1.80	1.41	1.63
	(.42)	(.56)	(.59)	(.75)
All internal causes	.39	1.07	1.69	1.25
	(.54)	(.80)	(.74)	(1.01)
Alcohol-related causes	.44	.80	.74	1.03
	(.21)	(.32)	(.33)	(.41)
Controls	age	age, age^2, interacted with over-21	age	age, age^2, interacted with over-21
Sample size	48	48	24	24

Notes: This table reports coefficients on an over-21 dummy from regressions of month-of-age-specific death rates by cause on an over-21 dummy and linear or interacted quadratic age controls. Standard errors are reported in parentheses.


21

Negotiating Sharp Turns

- **Figure 6.1.1 (Panel C)** shows how a sharp turn in $E[Y_{0i}|x_i]$ might be mistaken for a jump from one CEF to another
- Avoid such mistakes; look only at data near cutoffs:

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

so that

$$\begin{aligned} & \lim_{\delta\to 0} E\left[Y_i|x_0<x_i<x_0+\delta\right]-E\left[Y_i|x_0-\delta<x_i<x_0\right] \\ = \; & E[Y_{1i}|x_i=x_0]-E[Y_{0i}|x_i=x_0] \qquad (8) \\ = \; & E[Y_{1i}-Y_{0i}|x_i=x_0]\equiv\tau(x_0) \qquad (9) \end{aligned}$$

- This limiting argument obviates the need to model $E[Y_{0i}|x_i]$
- $\tau(x_0)$ is an average treatment effect near the cutoff, x_0

All About that Bandwidth

- To estimate, try kernel-weighting:

$$\hat{E}_n\left[\mathrm{Y}_{0i}|x_i = x_0\right] = \frac{\sum_i k_{\delta_n}(\tilde{x}_i)1[\tilde{x}_i < 0]\mathrm{Y}_i}{\sum_i k_{\delta_n}(\tilde{x}_i)1[\tilde{x}_i < 0]}, \tag{10}$$

  (similarly to the right of x_0) where $k_{\delta_n}(\tilde{x}_i)$ is a weight ("kernel") with data-dependent *bandwidth* δ_n that shrinks as n grows

  - **The Edge**, perhaps: $k_\delta(\tilde{x}_i) = 1[|\frac{\tilde{x}_i}{\delta}| < 1] \cdot (1 - |\frac{\tilde{x}_i}{\delta}|)$
  - Assuming $\delta_n \to 0$ (not too fast) as $n \to \infty$ (not too slow), The Edge's nonparametric $\hat{E}_n\left[\mathrm{Y}_{ji}|x_i = x_0\right] \underset{p}{\longrightarrow} E\left[\mathrm{Y}_{ji}|x_i = x_0\right]$

- Nonpara-'metrics amounts to a comparison of weighted averages near the cutoff: pick your bw and go!
  - Alas, local averaging is big-time biased at boundary points
  - We improve on this by estimating regs like (7) inside the bandwidth
  - For example: +/-12 months, as in columns 3-4 of **MM Table 4.1**

Modern Times: Nonparametric RD via Local Linear Regression

- *Local averaging is biased at boundary points*.
  - Hahn, Todd, and van der Klaauw (2001) use local linear regression (LLR) to reduce bias; this has become the RD standard
- Local linear regression (LLR) is weighted least squares, minimizing:

$$\sum_i k_\delta(\tilde{x}_i)(\mathrm{Y}_i - \alpha - \tau_0 \mathrm{D}_i - \beta_{01}\tilde{x}_i - \beta_1^* \mathrm{D}_i \tilde{x}_i)^2$$

  This model is "local" because: kernel weighting discards obs outside the bandwidth, weights those inside in proportion to cutoff proximity, under shrinking bw as sample size grows
- Nonparametric RD doesn't fix RD designs compromised by running variable manipulation

The IK Bandwidth

- On the journey to nonparametric asymptopia, bandwidth shrinks, and so the data grow thin. This increases standard errors!
  - RD 'metrics masters hope for stable, precise estimates as bw narrows
- Imbens and Kalyanaraman (2011) propose a data-driven bw that minimizes MSE
  - Let $\hat{\tau}(\delta)$ be the nonparametric estimator based on a bandwidth δ and fixed n. Define MSE as a function of bw:

$$\begin{aligned} MSE(\delta) &\equiv E[\hat{\tau}(\delta) - \tau]^2 = (E[\hat{\tau}(\delta)] - \tau)^2 + E(\hat{\tau}(\delta) - E[\hat{\tau}(\delta)])^2 \\ &= BIAS_\delta^2(\hat{\tau}) + VAR_\delta(\hat{\tau}) \end{aligned}$$

  - Approximate $MSE(\delta)$ is minimized by a function of the conditional variance of $\mathrm{Y}_i|x_i$ and second derivatives of $E[\mathrm{Y}_i|x_i]$
- Calonico, Cattaneo, and Titiunik (2014), Armstrong and Kolesar (2017), Imbens and Wager (2018) improve on this

Spec Checks and Worries

- Manipulation messes with RD
  - When those in control strive to avoid or cross the threshold, potential outcomes to the left and right of a cutoff are unlikely to be similar
- Spec checks
  - Balance tests: baseline covariates should be similar across the cutoff
  - Density smoothness: manipulation may induce spikes and discontinuities in running variable densities when someone tries to push across the cutoff (McCrary 2008 introduces a formal test for this)
    - Angrist, Lavy, Leder-Luis, and Shany (2019) uncover evidence of manipulated Israeli school enrollment (Israeli schools get to add a class when enrollment tops 40)
- *Heaps of trouble*: **Almond, et al.. (2010)** estimate effects of the extra medical care VLBW babies get; this is RD at a 1500 gram cutoff
- **Barecca, et al. (2011)** argue that heaping biases these estimates

ADKW (2010) First Stage

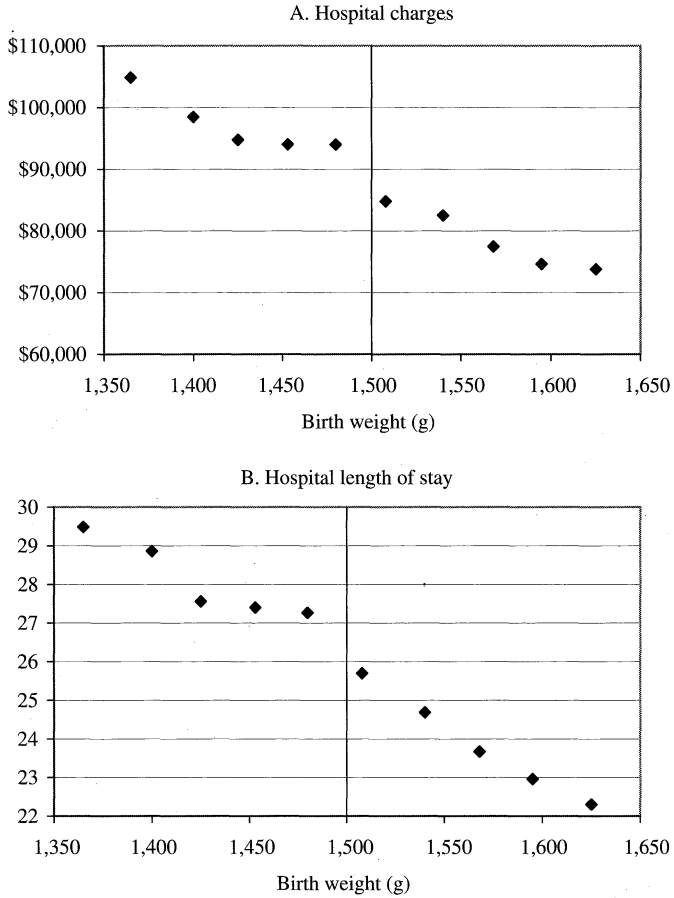
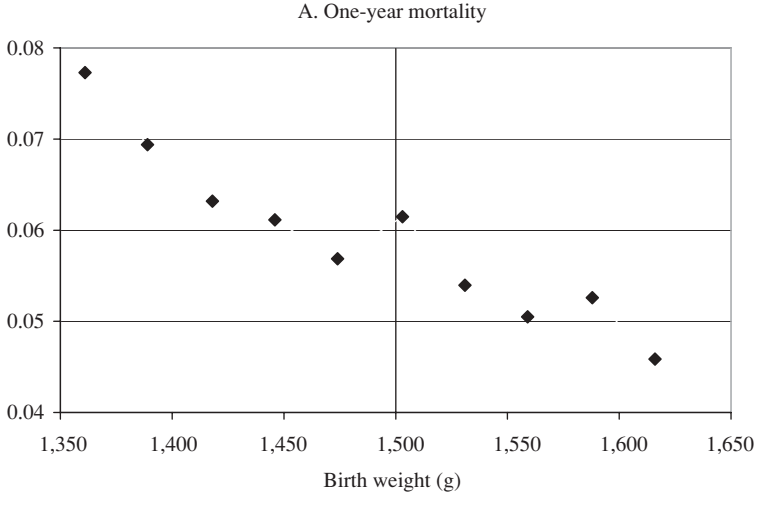


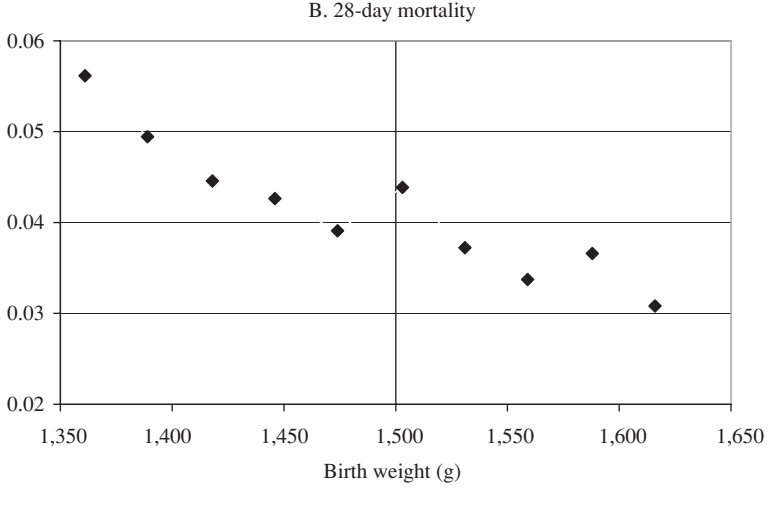
FIGURE III

Summary Treatment Measures around 1,500 g

Data are all births in the five-state sample (AZ, CA, MD, NY, and NJ), as described in the text. Charges are in 2006 dollars. Points represent gram-equivalents of ounce intervals, with births grouped into one-ounce bins radiating from 1,500 g; the estimates are plotted at the median birth weight in each bin.

ADKW RF


A. One-year mortality
0.08
0.07
0.06
0.05
0.04
1,350
1,400
1,450
1,500
1,550
1,600
1,650
Birth weight (g)


B. 28-day mortality
0.06
0.05
0.04
0.03
0.02
1,350
1,400
1,450
1,500
1,550
1,600
1,650
Birth weight (g)

Heaps of Birth Weight

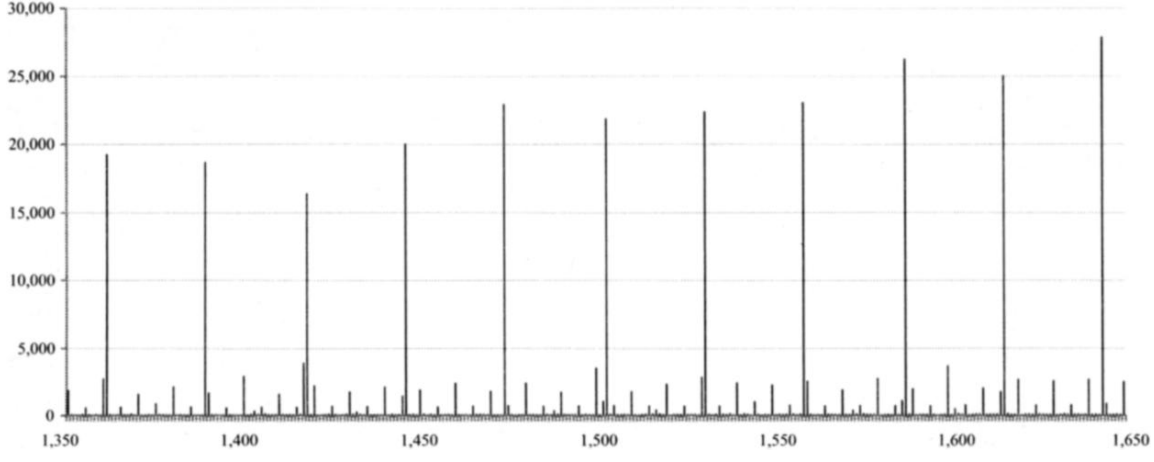


FIGURE I
Frequency of Births by Gram: Population of U.S. Births between 1,350 and 1,650 g
NCHS birth cohort linked birth/infant death files, 1983–1991 and 1995–2003, as described in the text.

Doin' Donuts

TABLE I

REPLICATION OF ADKW'S MAIN RESULTS ALONG WITH DONUT-RD ESTIMATES

Mortality Outcome	One-year (1)	28-Day (2)	7-Day (3)	24-Hour (4)
Panel A: Our replication of ADKW's estimates				
Weight < 1,500 g	−0.0071	−0.0071*	−0.0046	−0.0033
	(0.0041)	(0.0032)	(0.0028)	(0.0020)
Observations	202,078	202,078	202,078	202,078
Clusters	171	171	171	171
Panel B: Donut RD dropping those at 1,500 g				
Weight < 1,500 g	−0.0033*	−0.0042**	−0.0023	−0.0018
	(0.0014)	(0.0013)	(0.0013)	(0.0010)
Observations	198,534	198,534	198,534	198,534
Clusters	170	170	170	170
Panel C: Donut RD dropping those within 1 g of 1,500-g cutoff				
Weight < 1,500 g	−0.0035*	−0.0043**	−0.0024	−0.0018
	(0.0014)	(0.0012)	(0.0013)	(0.0010)
Observations	198,334	198,334	198,334	198,334
Clusters	168	168	168	168
Panel D: Donut RD dropping those within 2 g of 1,500-g cutoff				
Weight < 1,500 g	−0.0027*	−0.0037**	−0.0019	−0.0013
	(0.0014)	(0.0012)	(0.0012)	(0.0009)
Observations	197,135	197,135	197,135	197,135
Clusters	166	166	166	166
Panel E: Donut RD dropping those within 3 g of 1,500-g cutoff				
Weight < 1,500 g	−0.0018	−0.0026	−0.0018	−0.0011
	(0.0019)	(0.0015)	(0.0015)	(0.0014)
Observations	175,108	175,108	175,108	175,108

Fuzzy Logic

RD Meets IV

- In a fuzzy RD scenario, D_i denotes treatment as before, but this is no longer deterministically related to the threshold-crossing rule, $x_i \geq x_0$
- Instead, we use

$$\mathrm{T}_i = 1(x_i \geq x_0)$$

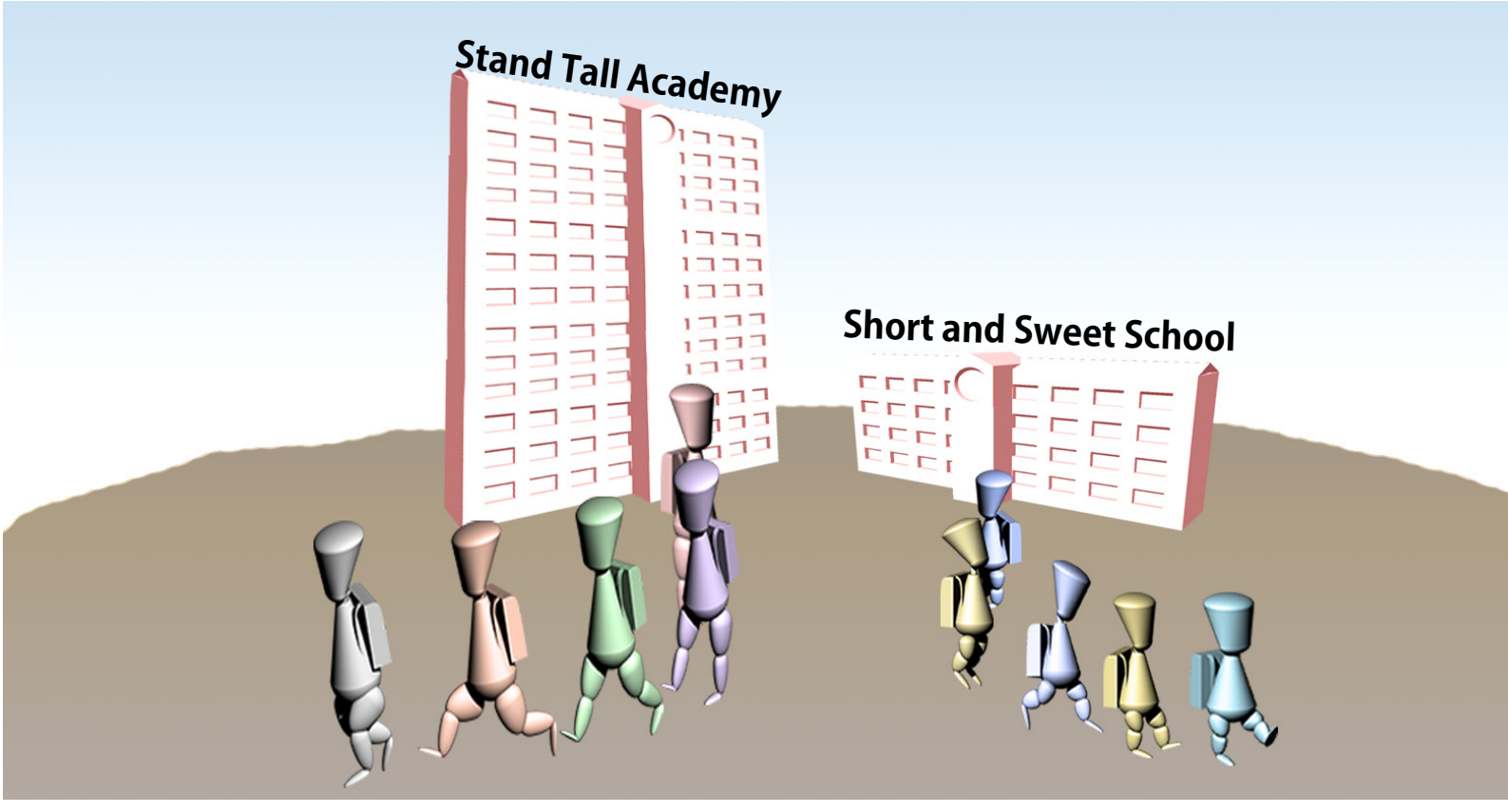
  to instrument D_i
- This works when treatment intensity changes discontinuously at x_0, and when cutoff-crossing changes outcomes only by virtue of this
- We implement nonparametric and fuzzy by estimating local linear first and second stages:

$$\begin{aligned} \mathrm{D}_i &= \gamma + \pi \mathrm{T}_i + \gamma_1 \tilde{x}_i + \gamma_2 \mathrm{T}_i \tilde{x}_i + \varepsilon_i \\ \mathrm{Y}_i &= \alpha + \lambda \mathrm{D}_i + \beta_1 \tilde{x}_i + \beta_2 \mathrm{T}_i \tilde{x}_i + \eta_i \end{aligned}$$

  Fuzzy nonparametric RD fits these by kernel-weighted least squares inside the bandwidth

Exam Time!

Examining Exam School Quality (**SEII**)



My kids gonna go Tall!

Simple Comparisons Mislead

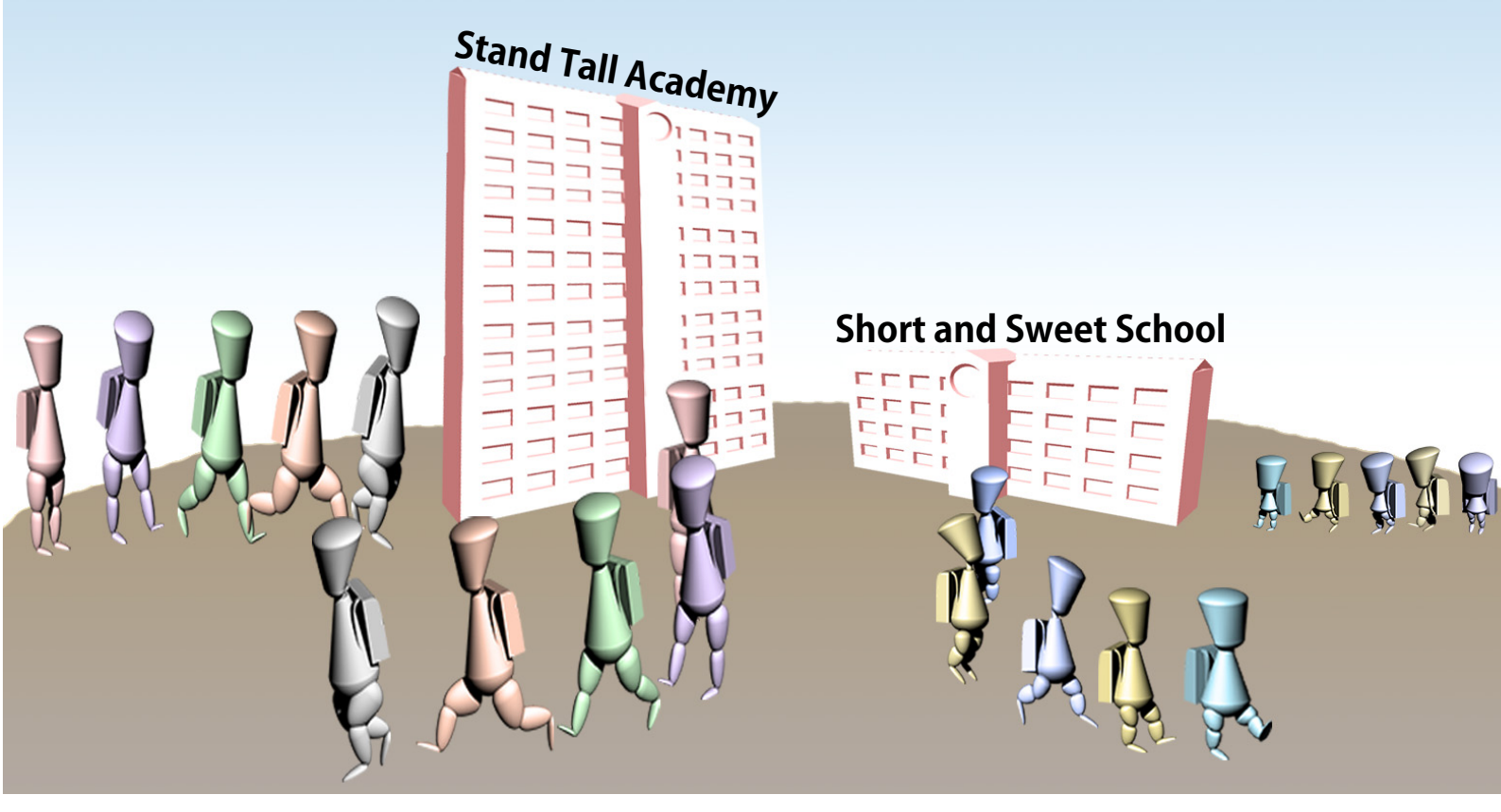
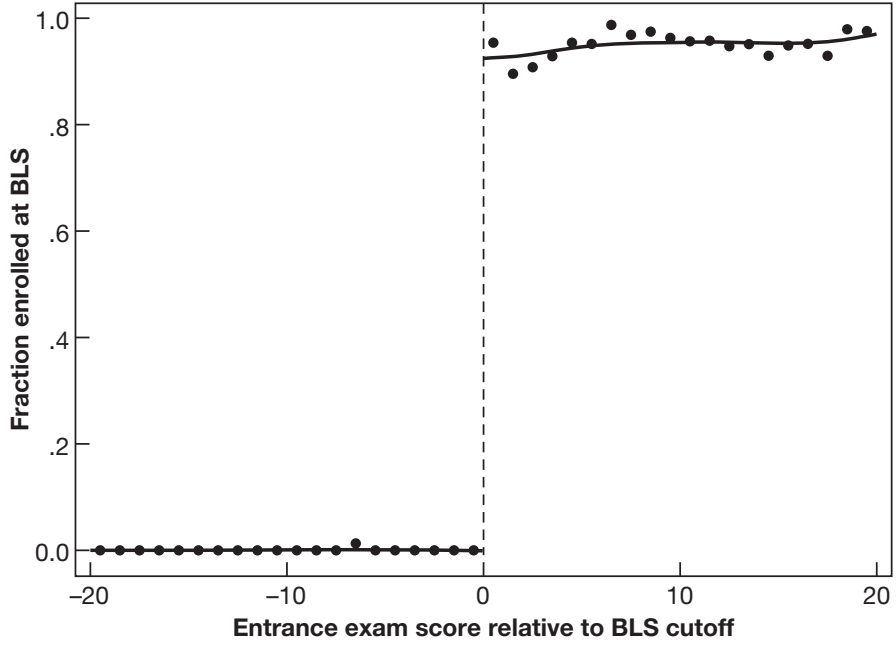
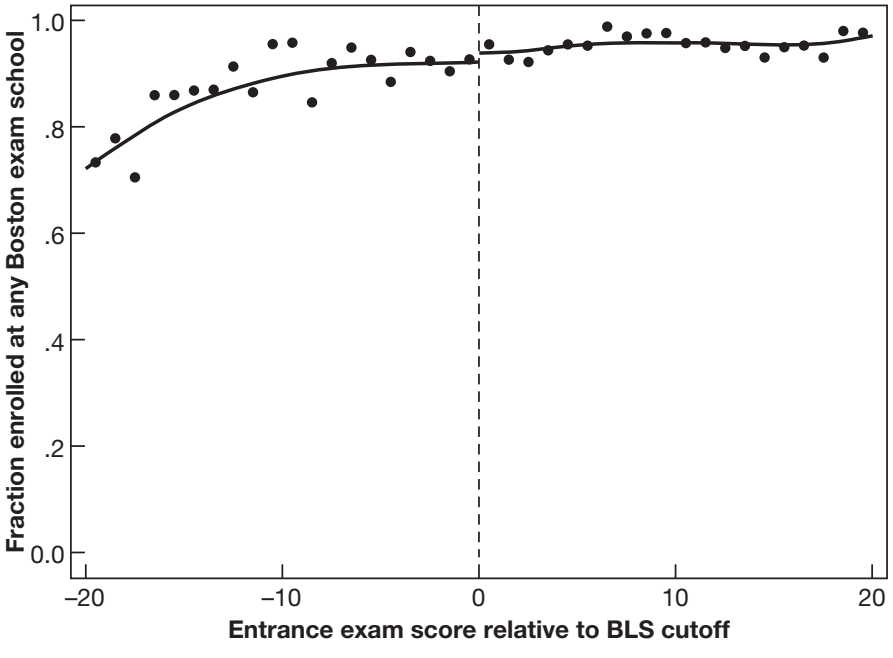


FIGURE 4.6
Enrollment at BLS



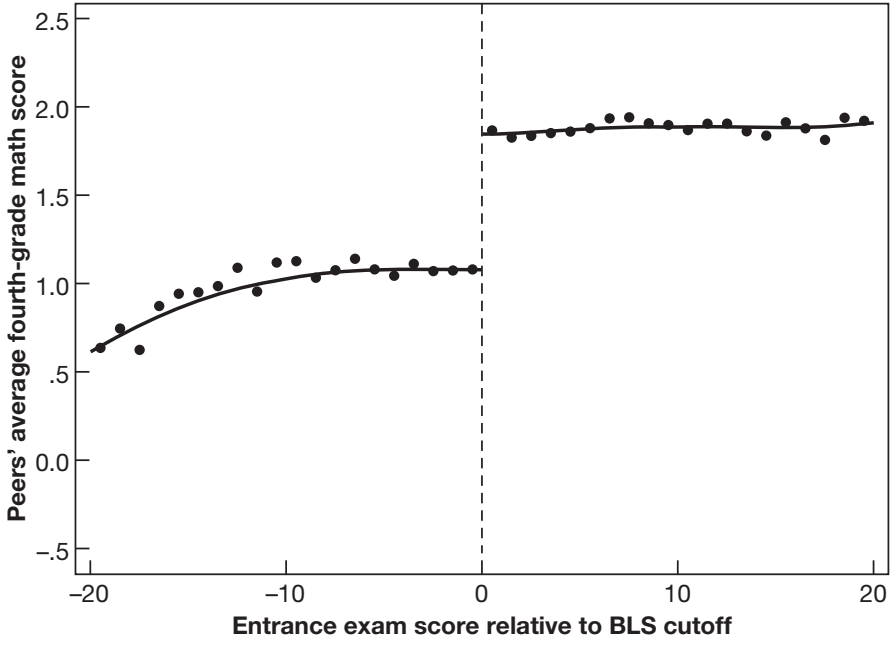
Notes: This figure plots enrollment rates at Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the BLS admissions cutoff. Solid lines show fitted values from a local linear regression estimated separately on either side of the cutoff (indicated by the vertical dashed line).

FIGURE 4.7
Enrollment at any Boston exam school



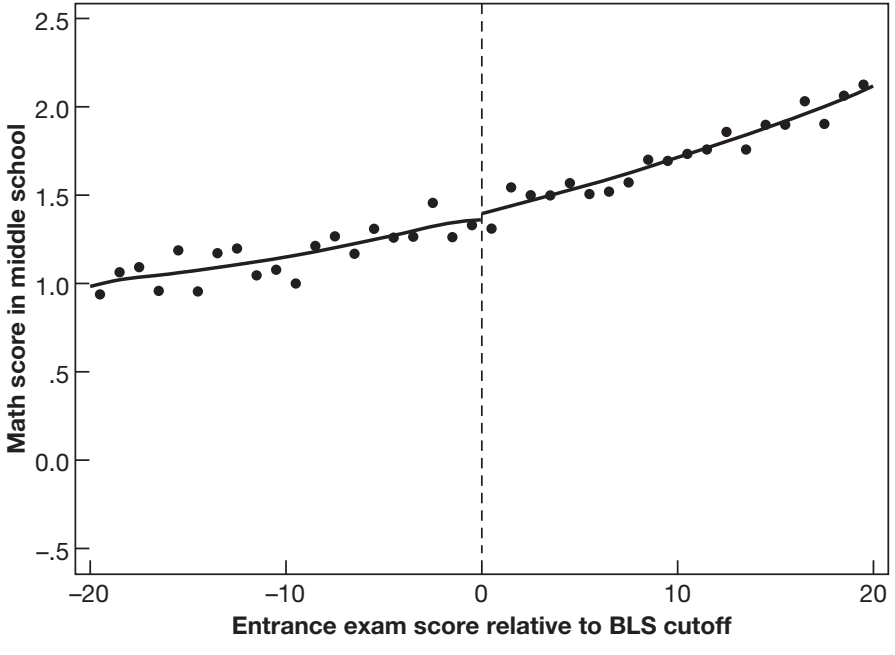
Notes: This figure plots enrollment rates at any Boston exam school, conditional on admissions test scores, for Boston Latin School (BLS) applicants scoring near the BLS admissions cutoff. Solid lines show fitted values from a local linear regression, estimated separately on either side of the cutoff (indicated by the vertical dashed line).

FIGURE 4.8
Peer quality around the BLS cutoff



Notes: This figure plots average seventh-grade peer quality for applicants to Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the admissions cutoff. Peer quality is measured by seventh-grade schoolmates' fourth-grade math scores. Solid lines show fitted values from a local linear regression, estimated separately on either side of the cutoff (indicated by the vertical dashed line).

FIGURE 4.9
Math scores around the BLS cutoff



Notes: This figure plots seventh- and eighth-grade math scores for applicants to the Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the admissions cutoff. Solid lines show fitted values from a local linear regression, estimated separately on either side of the cutoff (indicated by the vertical dashed line).

RF/First: It's Two-Stage Least Squares!

- The second stage model looks like this:

$$y_i = \Gamma' X_i + \psi' D_i + \eta_i$$

  - X_i includes controls (year of test, grade, application school/cohort, and the running variables that determine offers)
  - D_i is (instrumented) peer achievement and/or peer composition; ψ is a vector of causal effects
  - Data come from six schools: 3 Boston, 3 NYC

- This fuzzy RD setup allows for multiple causal channels
  - Variables to be instrumented are continuous rather than dummies
  - First stages include the controls in the second stage, plus exam-school offer dummies as instruments, in a six-school, two-city stack
  - To boost precision, interactions between offer dummies and application cohort are added to the instrument list

- **Table 9** reports estimates of ψ, along with first stage estimates

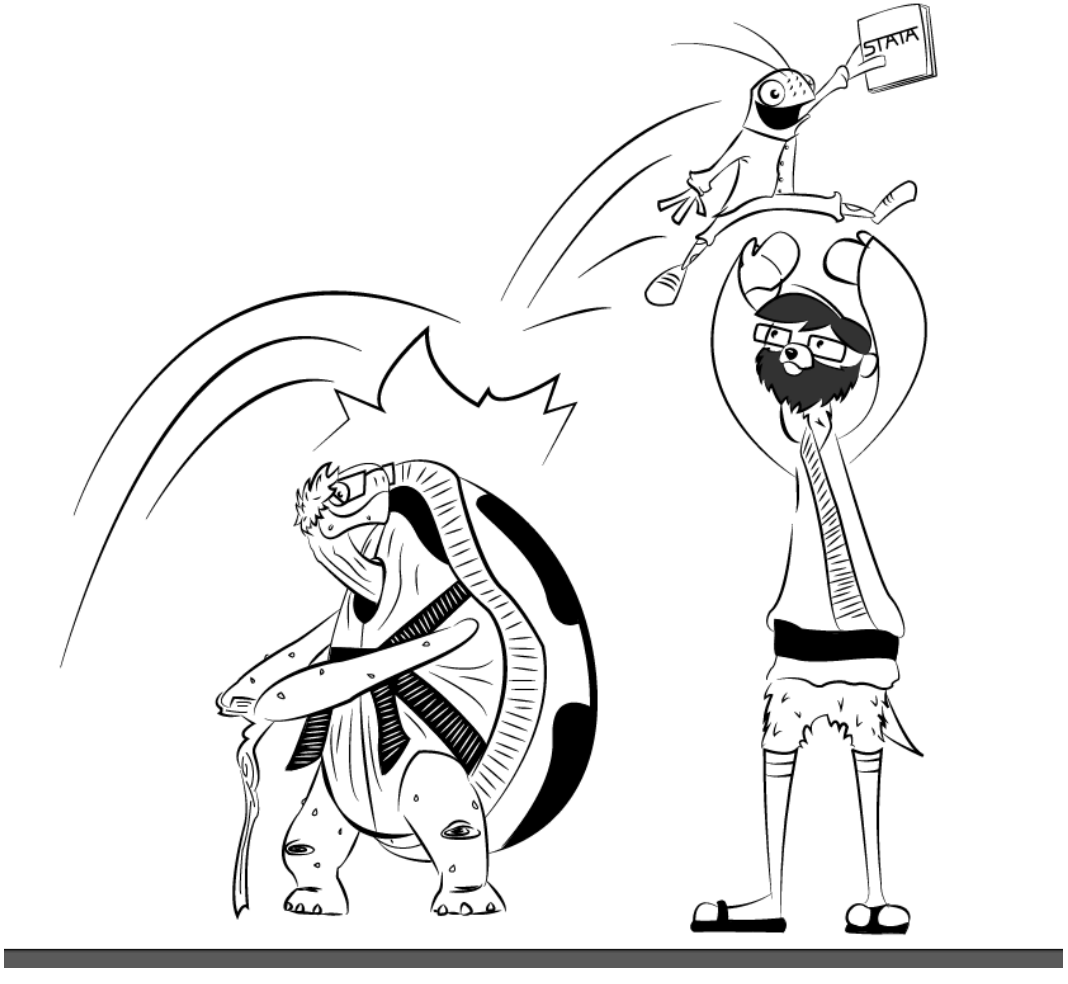
Table 9. 2SLS Estimates for Boston and New York

	Math				
	(1)	(2)	(3)	(4)	(5)
	2SLS (models with cohort interactions)				
Peer mean	-0.045		0.062	-0.046	
	(0.031)		(0.080)	(0.045)	
Proportion nonwhite		0.173	0.440		0.196
		(0.114)	(0.286)		(0.147)
Years in exam school				0.001	0.009
				(0.038)	(0.032)
	First stage Fs (models with cohort interactions)				
Peer mean	76.4		9.6	61.9	
Proportion nonwhite		65.6	16.8		58.9
Years in exam school				16.5	18.6
N	31862	33264	31862	31862	33264
	First Stage Estimates (models without cohort interactions)				
	Panel A: Boston				
O'Bryant	0.760***	-0.123***			
	(0.071)	(0.013)			
Latin Academy	0.347***	-0.208***			
	(0.075)	(0.014)			
Latin School	0.790***	-0.229***			
	(0.040)	(0.013)			
	Panel B: NYC				
Brooklyn Tech	0.494***	-0.137***			
	(0.074)	(0.024)			
Bronx Science	0.175***	-0.101***			
	(0.067)	(0.031)			
Stuyvesant	0.264***	-0.066***			
	(0.076)	(0.022)			

RD Frontiers

- Schemes to manage manipulation (Doing donuts as in Barreca, et al. 2011; bounds ala Gerard, Rokkanen, Rothe 2019)
- Fix those SEs, yo (Calonico, Cattaneo, and Titunik 2014)
- Discrete running variables (Rothe and Kolesar 2018)
- Inference for regression kink designs (Ganong and Jaeger 2018)
- Matching markets with mixed multiple tie-breaking (Abdulkadiroglu, et al.2021)
- RD away from the cutoff (Angrist and Rokkanen 2015)

You can master 'metrics too!



Tables and Figures

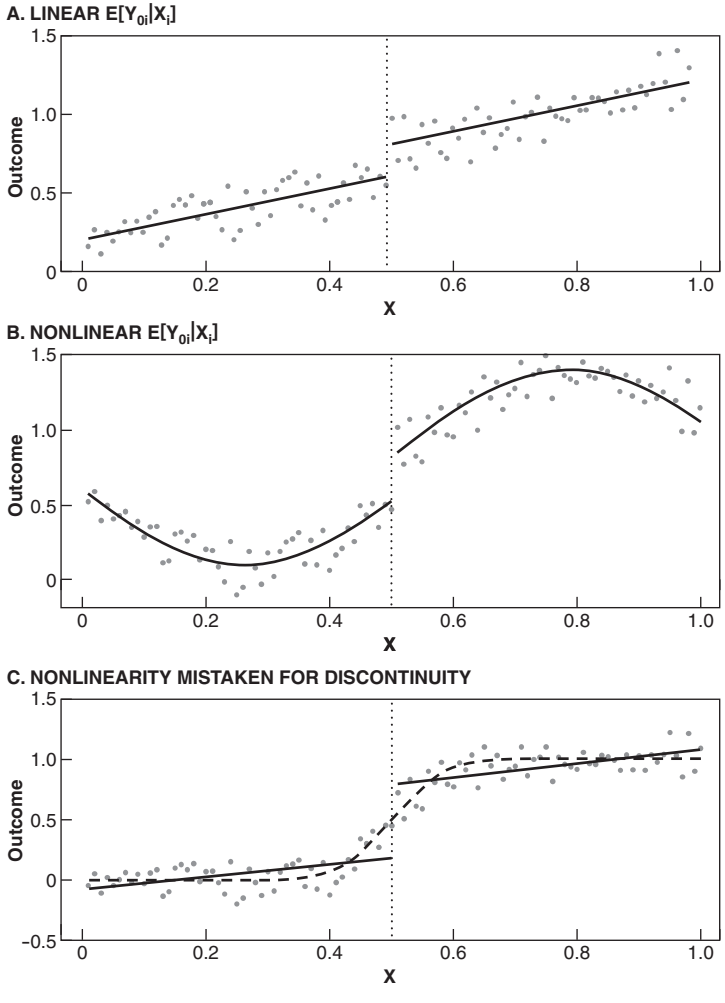


Figure 6.1.1 The sharp regression discontinuity design.

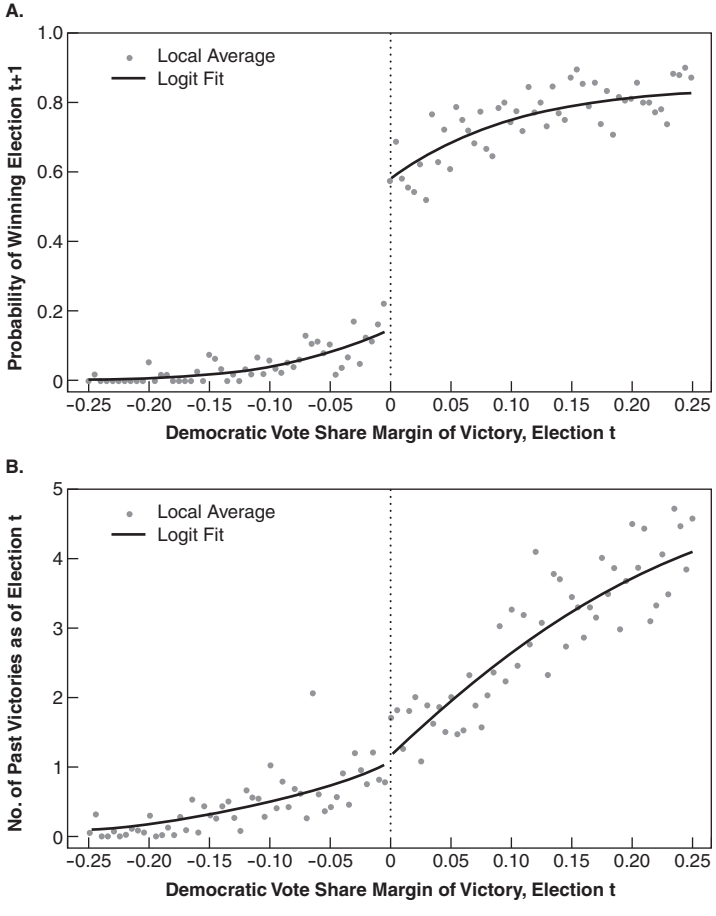


Figure 6.1.2 The probability of winning an election by past and future vote share (from Lee, 2008). (A) Candidate's probability of winning election $t + 1$, by margin of victory in election t: local averages and logit polynomial fit. (B) Candidate's accumulated number of past election victories, by margin of victory in election t: local averages and logit polynomial fit.