

Economics 142

D. Card

Today's agenda

1. quick overview of class

- course requirements: problem sets; midterm; course project
- course content

2. quick refresher on some basic stats

3. confidence intervals; minimum sample sizes

4. prep. for PS#1

Refresher on statistics (Appendix B, C of Wooldridge)

Y_1, Y_2, \dots, Y_n random sample from a pop, mean μ , variance σ^2

$\bar{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ is the sample mean; a “statistic” (a function of the sample that has no unknown parameters)

$s_n^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ is the sample variance (note $n-1$)

$E[\bar{Y}_n] = \mu$: the sample mean is “unbiased” for the pop. mean

$Var[\bar{Y}_n] = Var[\sum_{i=1}^n (Y_i/n)] = \sigma^2/n.$

$E[s_n^2] = \sigma^2$: the “d.f. corrected” sample var is unbiased for σ^2

convergence in probability: Z_1, Z_2, \dots is a sequence of r.v.'s *converges in probability* to b if for any $\varepsilon > 0$:

$$\lim_{n \rightarrow \infty} P(|Z_n - b| < \varepsilon) = 1$$

Write as $\text{plim } Z_n = b$.

Three famous results: Markov inequality; Chebyshev inequality; WLLN

1. *Markov:* if X is r.v., with $P(X > 0) = 1$ then for any $t > 0$:

$$P(X \geq t) \leq \frac{E[X]}{t}.$$

Proof: $E[X] = \int_0^\infty xf(x)dx = \int_0^t xf(x)dx + \int_t^\infty xf(x)dx$

$$\Rightarrow E[X] \geq \int_t^\infty xf(x)dx \geq t \int_t^\infty f(x)dx = tP(X \geq t).$$

Chebychev: If X is a random variable s.t. $Var[X]$ exists, then for any $t > 0$:

$$P(|X - E[X]| \geq t) \leq \frac{Var[X]}{t^2}.$$

Proof: consider r.v. $Y = (X - E[X])^2$. Note $E[Y] = Var[X]$. Using Markov $P(Y \geq \tau) \leq \frac{E[Y]}{\tau}$.

So, letting $\tau = t^2$,

$$P(Y \geq t^2) = P(|X - E[X]| \geq t) \leq \frac{E[Y]}{t^2} = \frac{Var[X]}{t^2}$$

WLLN. Suppose that Y_1, Y_2, \dots, Y_n is a random sample from a pop with mean μ , variance σ^2 (both finite). Then $\text{plim } \bar{Y}_n = \mu$.

Proof. Pick $\varepsilon > 0$. Applying Chebychev to \bar{Y}_n :

$$P(|\bar{Y}_n - \mu| \geq \varepsilon) \leq \frac{\text{Var}[\bar{Y}_n]}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$$

$$\Rightarrow P(|\bar{Y}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2}$$

So

$$\lim_{n \rightarrow \infty} P(|\bar{Y}_n - \mu| < \varepsilon) = 1.$$

WLLN says that the distribution of the sample mean “collapses” to the point μ as the sample size gets bigger, (i.e., $\text{plim}(\bar{Y}_n - \mu) = 0$).

WLLN says $\text{plim}(\bar{Y}_n - \mu) = 0$. The *Central Limit Theorem (CLT)* says that the distribution of \bar{Y}_n collapses to a *normal* at the rate $n^{1/2}$: if we consider the “scaled” r.v. $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$, this has a normal distribution $N(0, 1)$ as $n \rightarrow \infty$. A key idea of statistics is that for a given n we can step back from the limit and still be “approximately” OK.

$\{Z_n\}$, a sequence of r.v.’s, *converges in distribution to a r.v. Z with c.d.f. $F(x)$* if

$$\lim_{n \rightarrow \infty} P(Z_n \leq x) = F(x).$$

CLT: Let Y_1, Y_2, \dots, Y_n be a random sample from a population with mean μ , variance σ^2 . Then the “scaled” r.v. $\sqrt{n}(\bar{Y}_n - \mu)/\sigma$ converges in distribution to $N(0, 1)$. That is, for any fixed x :

$$\lim_{n \rightarrow \infty} P\left(\frac{\sqrt{n}(\bar{Y}_n - \mu)}{\sigma} \leq x\right) = \Phi(x),$$

where $\Phi()$ is the standard normal c.d.f. This is often written as

$$\sqrt{n}(\bar{Y}_n - \mu)/\sigma \approx N(0, 1)$$

$$\Rightarrow \bar{Y}_n \approx N(\mu, \sigma^2/n)$$

In fact, CLT remains true if, instead of scaling by σ , we scale by s_n (the *estimate* of σ):

$$\sqrt{n}(\bar{Y}_n - \mu)/s_n \approx N(0, 1).$$

Aside on sampling from a normal distribution.

CLT says that the sample mean is “asymptotically normal”, regardless of the underlying distribution that the Y ’s are drawn from (as long as μ and σ^2 are finite). Suppose that each Y_i is a draw from $N(\mu, \sigma^2)$. In this case,

$$\bar{Y}_n = \sum_{i=1}^n (Y_i/n).$$

Now, we know that if X and Z are independently distributed $X \sim N(\mu_x, \sigma_x^2)$ and $Z \sim N(\mu_z, \sigma_z^2)$ then

$$aX + bZ \sim N(a\mu_x + b\mu_z, a^2\sigma_x^2 + b^2\sigma_z^2)$$

Extending this result $\bar{Y}_n \sim N(\mu, \sigma^2/n)$ or $\sqrt{n}(\bar{Y}_n - \mu)/\sigma \sim N(0, 1)$. In this case, the distribution of \bar{Y}_n is exact.

Aside on sampling from a normal distribution, continued....

Also, the distribution when we use s_n instead of σ (which is unknown) to scale is known to be a so-called “t-distribution”:

$$\sqrt{n}(\bar{Y}_n - \mu)/s_n \sim t_{n-1}$$

where t_{n-1} is the t-distribution with $n-1$ degrees of freedom. For large n the t is very close to the standard normal. For smaller n the t distribution has fatter tails.

Confidence intervals.

Suppose $Z \sim N(0, 1)$. Then we know Z is symmetrically distributed around 0 with a “bell curve” distribution. Define $z_p > 0$ as the real number such that $\Phi(z_p) = 1 - p$ (for $p < .5$). This is the point such that $P(Z \geq z_p) = p$. We ask: what is the symmetric interval (around 0) such that a standard normal falls in the interval with probability $1 - \alpha$? This is the interval $(-z_{\alpha/2}, z_{\alpha/2})$. Why? Because the probability of falling above $z_{\alpha/2}$ is $\alpha/2$, and by symmetry the probability of falling below $-z_{\alpha/2}$ is also $\alpha/2$. So the probability of being outside the interval is α .

For $Z \sim N(0, 1)$ $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$. Suppose we have obtained a random sample of some Y 's and formed the estimated mean and standard deviation. By the CLT $\sqrt{n}(\bar{Y}_n - \mu)/s_n \approx N(0, 1)$, so (approximately):

$$P(-z_{\alpha/2} \leq \sqrt{n}(\bar{Y}_n - \mu)/s_n \leq z_{\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P\left(-\frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq \bar{Y}_n - \mu \leq \frac{s_n z_{\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-\bar{Y}_n - \frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq -\mu \leq -\bar{Y}_n + \frac{s_n z_{\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\bar{Y}_n - \frac{s_n z_{\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{Y}_n + \frac{s_n z_{\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha$$

This is interpreted as: if we kept repeating a sample of size n , $1 - \alpha$ percent of the time the interval $\bar{Y}_n \pm \frac{s_n z_{\alpha/2}}{\sqrt{n}}$ would “capture” the true mean μ . This is called the $(1 - \alpha)$ “confidence interval”.

Using these ideas.

Suppose we have to draw a sample of a binary random variable (e.g., the fraction of people who vote Democrat; or the fraction of mortgages in a “mortgage-backed security” portfolio that were improperly underwritten). Let p represent the true probability of the event of interest being “true”: so Y_i is a Bernoulli r.v. with mean p and variance $p(1-p)$. For a sample of size n we estimate the mean of the Y_i 's, which is the fraction of “1's” we get. For simplicity call this \bar{p}_n . Note that $E[\bar{p}_n] = p$. Also, in this case our estimate of the variance term is

$$s_n^2 = \frac{1}{n-1} \sum_i (Y_i - \bar{p}_n)^2 = \frac{1}{n-1} \sum (Y_i^2 - 2\bar{p}_n Y_i + \bar{p}_n^2) = \frac{n}{n-1} \bar{p}_n (1 - \bar{p}_n).$$

In the Bernoulli case, people often divide the sum by n so the estimate is $s_n = \sqrt{\bar{p}_n(1 - \bar{p}_n)}$.

How big a sample do we need?

“Margin of Error” – one way people decide the sample size is to set a “margin of error” with a given level of confidence. The margin of error is $1/2$ of the width of a $(1-\alpha)$ confidence interval. For a 95% confidence interval (the “industry standard”), $z_{\alpha/2} = 1.96$. Thus the width of half of the CI is

$$m = \frac{s_n z_{\alpha/2}}{\sqrt{n}}.$$

Now we don't know p so we don't know s_n : but a “worst case” is $p_n = 0.5$ which implies that $s_n = \sqrt{0.5(1 - 0.5)} = 0.5$. So the “worst case” for m with a sample size of n is

$$m = \frac{0.5 z_{\alpha/2}}{\sqrt{n}}.$$

If we choose a margin of error m , we need a sample size of

$$n = \left(\frac{z_{\alpha/2}}{2m} \right)^2$$

A standard setting is $m = 0.05$, which with a 95% confidence needs $n \approx 400$. Note that if we use a 95% confidence level then $z_{\alpha/2} \simeq 2$ and

$$n \simeq \left(\frac{1}{m}\right)^2$$

“Minimum Detectable Effect” – another way to choose a sample size is to ask what deviation from a given value would you like to be able to “reliably detect”. If the default “null hypothesis” is $p = p^0$, we might want to be able to say that if we obtain a point estimate of $p_n = p'$ then it will be “significantly different from p^0 . Assuming a null of $p = p^0$, we would “reject the null” with an estimate of p' at the α level of significance (under a 2-tailed test) if

$$\frac{p' - p^0}{s_n / \sqrt{n}} > z_{\alpha/2}.$$

Again, using $s_n = 0.5$ as the “worst case” scenario, the sample size we need satisfies:

$$\sqrt{n} > \frac{z_{\alpha/2}}{2(p' - p^0)}.$$

For $\alpha = 0.05$, $z_{\alpha/2} = 1.96$, and we need (approximately):

$$n > \frac{1}{(p' - p^0)^2}.$$

If for example you want $p' - p^0 = 0.05$, you'll need $n = 400$. Notice that this is the same thing as having a 0.05 margin of error.

Lecture 2

Overview of the three parts of 142:

- a) descriptive modeling (weeks 1-5)
- b) causal modeling (weeks 6-10)
- c) prediction (weeks 11-15)

a) descriptive modeling

Often we are interested in trying to *summarize the relationship* between some “outcome” y and some other variables $x = (x_1, x_2 \dots x_J)$.

- we **aren't** necessarily trying to measure the causal effect of x_j on y
- we are trying to take account of the fact that y may be strongly related to some x'_j 's and only weakly related to others.
- e.g.: what is the relationship between earnings (y), gender (x_1), and other characteristics, like education (x_2)?
- our benchmark: $E[y|x]$

- benchmark: $E[y|x]$
- we are going to approximate this with a “linear regression function”
- we’ll consider 2 regression functions:
 - the “population” regression: the function we could estimate with ∞ sample
 - the “sample” regression: the one we can actually estimate with a given sample

b) causal modeling

Many debates in economics amount to disputes over the question:

does X “cause” Y ?

Examples

- a. if a student were to attend Stanford instead of Berkeley, would she earn more?
- b. if an unemployed worker has higher UI benefits, will he/she have longer spell of joblessness?

There is a huge metaphysical literature on causality. Among the issues: how to think about causality in an uncertain world.

A very empirical notion of causality:

X causes Y if, in an idealized experiment, we could manipulate X , leaving other factors constant, and observe that the mean of the distribution of outcomes of Y has changed

How can we know if the distribution of outcomes has changed?
We need to be able to see (or at least estimate) two things:

- the distribution of Y when X is manipulated (or “treated”)
- the distribution in the absence of manipulation – the *counterfactual*

Problem: we can't see both the outcome under treatment and the counterfactual outcome in the absence of treatment

How can we resolve the observability problem?

We need a way to infer the counterfactual for the units (people) that are treated

Possible ideas:

1. (“observational design”): calculate mean outcomes for people who are treated and those who are not
2. (“pre-post design”): compare outcomes for people who are treated with their outcomes prior to treatment (i.e., the mean change in outcome after treatment)
3. RCT - randomly assign treatment, calculate mean outcomes for T's and C's
4. other designs.... (stay tuned)

Lecture 3

1. quick review of vector notation (see handout)
2. conditional expectation function (CEF)
 - joint densities
 - properties of $E[y_i|x_i]$
3. the “infeasible” (population) regression function
 - relation to CEF
4. the feasible regression - OLS

We need some vector notation. Suppose we are interested in the linear equation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

where $i = 1\dots N$ indexes elements of a sample. Here (x_{1i}, x_{2i}, y_i) are observed values of two *covariates* (x_1, x_2) and our *outcome of interest* (y) for unit i . We can define the 3-row vector

$$x_i = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \end{pmatrix}$$

and the conformable 3-row vector

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

Using these vectors we can write the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

in vector notation:

$$y_i = x'_i \beta + u_i$$

Suppose we have k equations in k unknowns of the form:

$$a_{11}b_1 + a_{12}b_2 + \dots + a_{1k}b_k = c_1$$

$$a_{21}b_1 + a_{22}b_2 + \dots + a_{2k}b_k = c_2$$

...

$$a_{k1}b_1 + a_{k2}b_2 + \dots + a_{kk}b_k = c_k$$

This system can be represented as the matrix equation $Ab = c$, where

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1k} \\ a_{21} & a_{22} & \dots & a_{2k} \\ \dots & & & \dots \\ a_{lk1} & a_{k2} & \dots & a_{kk} \end{pmatrix}, \quad c = \begin{pmatrix} c_1 \\ c_2 \\ \dots \\ c_k \end{pmatrix}$$

A unique solution for b will exist if A has “full rank”: then A is “invertible” and $b = A^{-1}c$.

Review of probability

x, y are two r.v.'s, joint p.d.f $f(x, y)$

marginal densities $f(x) = \int_y f(x, y) dy, f(y) = \int_x f(x, y) dx$

conditional density

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

Note that this means $f(x, y) = f(y|x)f(x)$.

$$E[y] = \int_y y f(y) dy$$

$$E[y|x] = \int_y y f(y|x) dy$$

Law of interated expectations (LIE):

$$E[E[y|x]] = E[y]$$

Proof:

$$\begin{aligned} E[E[y|x]] &= \int_x E[y|x] f(x) dx \\ &= \int_x \int_y y f(y|x) dy f(x) dx \\ &= \int_y \int_x y f(y|x) f(x) dx dy \\ &= \int_y y \left(\int_x f(x,y) dx \right) dy \\ &= \int_y y f(y) dy \end{aligned}$$

Two excellent properties of CEF $E[y_i|x_i]$

1. We can always write $y_i = E[y_i|x_i] + \varepsilon_i$ where $E[\varepsilon_i|x_i] = 0$

and $E[\varepsilon_i h(x_i)] = 0$ for any function of x .

Proof: we first show $E[\varepsilon_i|x_i] = 0$:

$$\begin{aligned} E[\varepsilon_i|x_i] &= E[(y_i - E[y_i|x_i])|x_i] \\ &= E[y_i|x_i] - E[E[y_i|x_i]|x_i] = 0 \end{aligned}$$

Next, using LIE:

$$\begin{aligned} E[\varepsilon_i h(x_i)] &= E[E[\varepsilon_i h(x_i)|x_i]] \\ &= E[h(x_i)E[\varepsilon_i|x_i]] = 0 \end{aligned}$$

2. the function $m(x_i) = E[y_i|x_i]$ minimizes $E[(y_i - m(x_i))^2]$

Proof:

$$\begin{aligned} y_i - m(x_i) &= y_i - E[y_i|x_i] + E[y_i|x_i] - m(x_i) \\ \Rightarrow (y_i - m(x_i))^2 &= \varepsilon_i^2 + (E[y_i|x_i] - m(x_i))^2 \\ &\quad + 2\varepsilon_i(E[y_i|x_i] - m(x_i)) \\ \Rightarrow E[(y_i - m(x_i))^2] &= E[\varepsilon_i^2] + E[(E[y_i|x_i] - m(x_i))^2] \\ &\quad + 2E[\varepsilon_i(E[y_i|x_i] - m(x_i))] \end{aligned}$$

But the last term is 0, so the minimizing choice is $m(x_i) = E[y_i|x_i]!$

OLS and the CEF

OLS minimizes

$$\frac{1}{N} \sum_{i=1}^N (y_i - x_i' \beta)^2 \rightarrow E[(y_i - x_i' \beta)^2]$$

Consider the “infeasible” (or population) OLS estimator:

$$\min_{\beta} E[(y_i - x_i' \beta)^2]$$

What are the FOC? Consider the derivative w.r.t. j th element of β :

$$\begin{aligned}\frac{\partial x_i' \beta}{\partial \beta_j} &= x_{ji} \\ \Rightarrow \frac{\partial (y_i - x_i' \beta)^2}{\partial \beta_j} &= -2(y_i - x_i' \beta)x_{ji}\end{aligned}$$

So: the foc for the optimal choice β^* that solves:

$$\min_{\beta} E[(y_i - x_i' \beta)^2]$$

are:

$$\begin{aligned} E[-2x_i(y_i - x_i' \beta^*)] &= 0. \\ \Rightarrow E[x_i(y_i - x_i' \beta^*)] &= 0 \end{aligned}$$

How does $x_i' \beta^*$ relate to $E[y_i | x_i]$?

Property #1: If $E[y_i|x_i] = x'_i\beta^e$ then $\beta^* = \beta^e$.

Why? Recall that if we define the CEF error $\varepsilon_i = y_i - E[y_i|x_i]$,

$$E[x_i\varepsilon_i] = 0 \Rightarrow E[x_i(y_i - x'_i\beta^e)] = 0$$

Which implies that β^e satisfies the FOC for infeasible OLS.

This means that *if the true CEF is linear*, then the infeasible OLS represents the CEF.

This happens when x 's are dummies since $E[y_i|x_i]$ is $E[y_i|i]$ in group k]

Property #2: $x_i' \beta^*$ is the “best” linear approx. to $E[y_i|x_i]$ (best as in *minimum-MSE*)

Proof:

$$\begin{aligned}
 y_i - x_i' \beta &= y_i - E[y_i|x_i] + E[y_i|x_i] - x_i' \beta \\
 \Rightarrow (y_i - x_i' \beta)^2 &= \varepsilon_i^2 + (E[y_i|x_i] - x_i' \beta)^2 \\
 &\quad + 2\varepsilon_i(E[y_i|x_i] - x_i' \beta) \\
 \Rightarrow E[(y_i - x_i' \beta)^2] &= E[\varepsilon_i^2] + E[(E[y_i|x_i] - x_i' \beta)^2] \\
 &\quad + 2E[\varepsilon_i(E[y_i|x_i] - x_i' \beta)]
 \end{aligned}$$

And as before, $E[\varepsilon_i(E[y_i|x_i] - x_i' \beta)] = 0$. So the infeasible OLS minimand is

$$E[(y_i - x_i' \beta)^2] = E[\varepsilon_i^2] + E[(E[y_i|x_i] - x_i' \beta)^2]$$

So what is β^* ? Recall that the objective

$$\min_{\beta} E[(y_i - x'_i \beta)^2]$$

has foc that imply:

$$\begin{aligned} E[x_i(y_i - x'_i \beta^*)] &= 0 \\ \Rightarrow E[x_i x'_i] \beta^* &= E[x_i y_i] \\ \Rightarrow \beta^* &= [E[x_i x'_i]]^{-1} E[x_i y_i] \end{aligned}$$

We can think of the “population regression” as:

$$y_i = x'_i \beta^* + u_i$$

Notice that $u_i = \varepsilon_i + \{E[y_i|x_i] - x'_i \beta^*\}$, and $E[x_i u_i] = 0$. (why?)

The feasible regression (OLS) minimizes

$$SSR = \sum_{i=1}^N (y_i - x'_i \beta)^2$$

The foc (in vector form) are:

$$\sum_{i=1}^N -2x_i(y_i - x'_i \beta) = 0$$

which implies that

$$\begin{aligned} \sum_{i=1}^N x_i x'_i \beta &= \sum_{i=1}^N x_i y_i \\ \Rightarrow \hat{\beta} &= [\sum_{i=1}^N x_i x'_i]^{-1} \sum_{i=1}^N x_i y_i \end{aligned}$$

$$\hat{\beta} = \left[\frac{1}{N} \sum_{i=1}^N x_i x_i' \right]^{-1} \frac{1}{N} \left(\sum_{i=1}^N x_i y_i \right)$$

Now using: $y_i = x_i' \beta^* + u_i$

$$\begin{aligned}\hat{\beta} &= \left[\frac{1}{N} \sum_{i=1}^N x_i x_i' \right]^{-1} \frac{1}{N} \left(\sum_{i=1}^N x_i (x_i' \beta^* + u_i) \right) \\ &= \left[\frac{1}{N} \sum_{i=1}^N x_i x_i' \right]^{-1} \frac{1}{N} \sum_{i=1}^N x_i x_i' \beta^* + \left[\frac{1}{N} \sum_{i=1}^N x_i x_i' \right]^{-1} \frac{1}{N} \sum_{i=1}^N x_i u_i \\ &= \beta^* + \left[\frac{1}{N} \sum_{i=1}^N x_i x_i' \right]^{-1} \frac{1}{N} \sum_{i=1}^N x_i u_i\end{aligned}$$

The deviation depends on a term that should be small

Today's agenda

1. recap ideas from Lecture 3:

$E[y_i|x_i]$, population regression $x'_i\beta^*$, and OLS

2. an example where $E[y_i|x_i] \neq x'_i\beta^*$
3. properties of the population regression (F-W)
4. parallel properties of OLS

Recap

- vector notation: $y_i = x'_i \beta + u_i$ $x'_i = (1, x_{1i}, x_{2i} \dots x_{Ji})$
- CEF $\equiv E[y_i|x_i]$, a (possibly messy) function of x
- forecast error $\epsilon_i = y_i - E[y_i|x_i]$

$$E[\epsilon_i|x_i] = 0 \text{ and } E[\epsilon_i h(x_i)] = 0 \text{ for any } h(x_i)$$

Aside: what if $x_i = 1$ (only constant) $\Rightarrow E[y_i|1] = E[y_i]$

- showed CEF minimizes $E[(y_i - m(x_i))^2]$ among all possible $m(\cdot)$ functions

Next: the *population regression function (PRF)*

- for a particular set of x 's, a *regression function* is just a linear combination $x'_i\beta$
- PRF: $\beta^* = \operatorname{argmin}_{\beta} E[(y_i - x'_i\beta)^2]$
- FOC: $E[x_{ji}(y_i - x'_i\beta^*)] = 0$, one row for each covariate $j = 1 \dots J$.
- re-write as $E[x_{ji} x'_i\beta^*] = E[x_{ji} y_i]$
- e.g., 3-covariate case:

$$\begin{aligned} E[x_{1i}x_{1i}\beta_1^* + x_{1i}x_{2i}\beta_2^* + x_{1i}x_{3i}\beta_3^*] &= E[x_{1i}y_i] \\ E[x_{2i}x_{1i}\beta_1^* + x_{2i}x_{2i}\beta_2^* + x_{2i}x_{3i}\beta_3^*] &= E[x_{2i}y_i] \\ E[x_{3i}x_{1i}\beta_1^* + x_{3i}x_{2i}\beta_2^* + x_{3i}x_{3i}\beta_3^*] &= E[x_{2i}y_i] \end{aligned}$$

$$\begin{aligned}
 E[x_1 i x_{1i} \beta_1^* + x_1 i x_{2i} \beta_2^* + x_1 i x_{3i} \beta_3^*] &= E[x_{1i} y_i] \\
 E[x_2 i x_{1i} \beta_1^* + x_2 i x_{2i} \beta_2^* + x_2 i x_{3i} \beta_3^*] &= E[x_{2i} y_i] \\
 E[x_3 i x_{1i} \beta_1^* + x_3 i x_{2i} \beta_2^* + x_3 i x_{3i} \beta_3^*] &= E[x_{2i} y_i]
 \end{aligned}$$

Now use matrix notation:

$$E \begin{pmatrix} x_{1i} x_{1i} & x_{1i} x_{2i} & x_{1i} x_{3i} \\ x_{2i} x_{1i} & x_{2i} x_{2i} & x_{2i} x_{3i} \\ x_{3i} x_{1i} & x_{3i} x_{2i} & x_{3i} x_{3i} \end{pmatrix} \begin{pmatrix} \beta_1^* \\ \beta_2^* \\ \beta_3^* \end{pmatrix} = E \begin{pmatrix} x_{1i} y_i \\ x_{2i} y_i \\ x_{3i} y_i \end{pmatrix}$$

Or

$$\begin{aligned}
 E[x_i x'_i] \beta^* &= E[x_i y_i] \\
 \Rightarrow \beta^* &= E[x_i x'_i]^{-1} E[x_i y_i]
 \end{aligned}$$

Good properties of PRF

1. if $E[y_i|x_i] = x'_i\beta^e$ then $\beta^* = \beta^e$ (i.e., PRF = CEF)
2. $x'_i\beta^*$ is the *best linear approximation* to $E[y_i|x_i]$

How did we prove that? Define $\epsilon_i \equiv y_i - E[y_i|x_i]$.

$$\begin{aligned}y_i - x_i\beta &= y_i - E[y_i|x_i] + E[y_i|x_i] - x_i\beta \\&= \epsilon_i + E[y_i|x_i] - x_i\beta \\ \Rightarrow E[(y_i - x_i\beta)^2] &= E[\epsilon_i^2] + E[(E[y_i|x_i] - x_i\beta)^2]\end{aligned}$$

$$\min \text{ LHS} \Leftrightarrow \min \text{ RHS} \Leftrightarrow \min E[(E[y_i|x_i] - x_i\beta)^2]$$

A direct approach:

$$E[(y_i - x_i\beta)^2] = E[\epsilon_i^2] + E[(E[y_i|x_i] - x_i\beta)^2]$$

To minimize RHS we have FOC:

$$\begin{aligned} E[x_i(E[y_i|x_i] - x_i'\beta)] &= 0 \\ \Rightarrow E[x_i x_i' \beta] &= E[x_i E[y_i|x_i]] \\ &= E[E[x_i y_i | x_i]] \\ &= E[x_i y_i] \end{aligned}$$

which is the FOC for the PRF!

Special case: groups 0, 1, 2; indicators D_{1i}, D_{2i} ; $x'_i = (1, D_{1i}, D_{2i})$;

$E[y_i | i \in \text{group } g] = \mu_g$. Then:

$$E[y_i | x_i] = \mu_0 + D_{1i}(\mu_1 - \mu_0) + D_{2i}(\mu_2 - \mu_0)$$

Thus:

$$\beta^* = \begin{pmatrix} \mu_0 \\ \mu_1 - \mu_0 \\ \mu_2 - \mu_0 \end{pmatrix}$$

If $E[y_i|x_i]$ is not truly linear, we end up with an approximation error at each x_i .

Write:

$$\begin{aligned}y_i &= E[y_i|x_i] + \epsilon_i \\E[y_i|x_i] &= x_i' \beta^* + v_i \\\Rightarrow y_i &= x_i' \beta^* + v_i + \epsilon_i \\&= x_i' \beta^* + u_i\end{aligned}$$

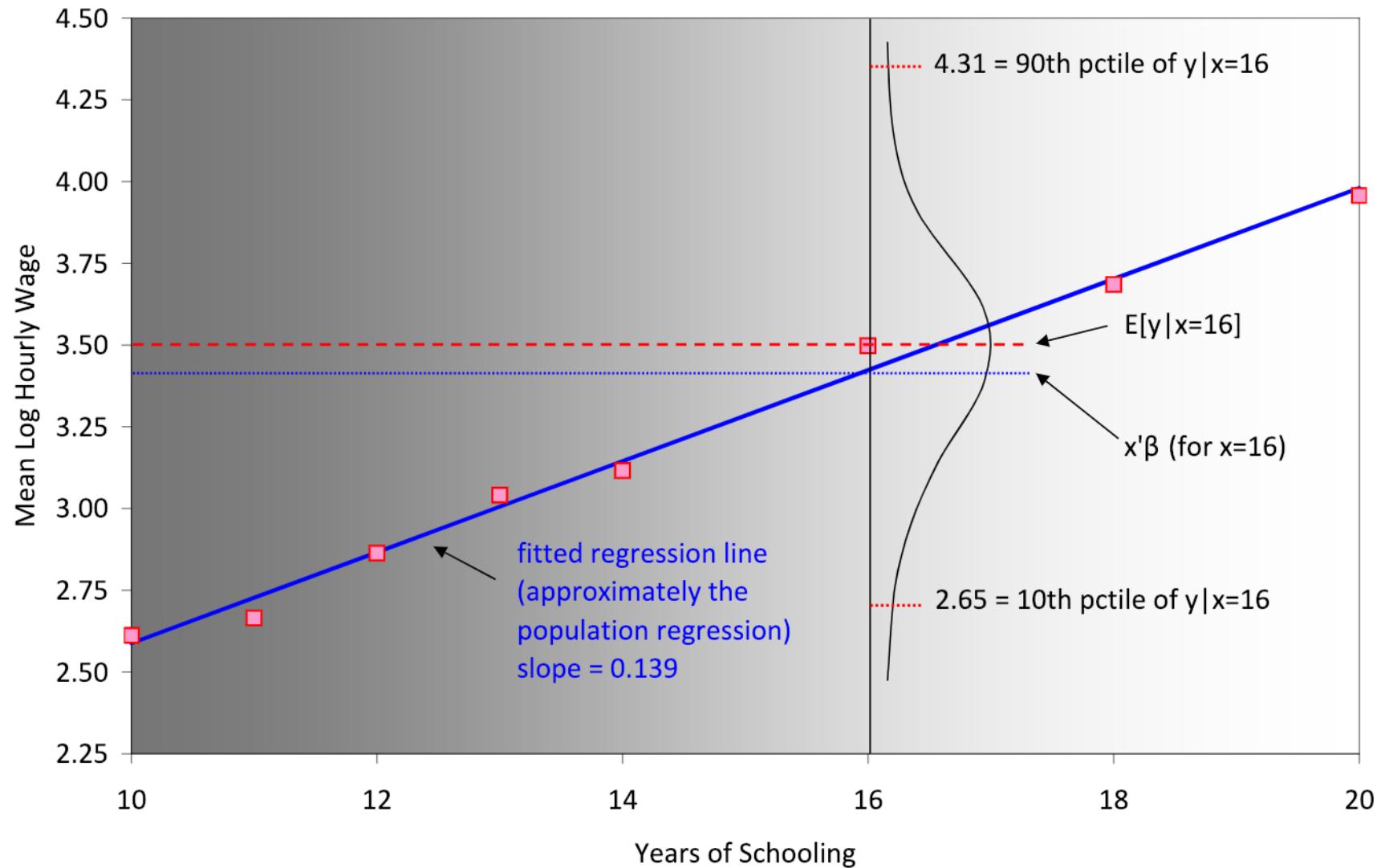
So the error in the population regression is $v_i + \epsilon_i$, where v_i is function of x_i . Note that $E[x_i v_i] = 0$ but $E[v_i|x_i] \neq 0$ (in general).

Example: data on education and earnings from Am. Community Survey

- 1.5 million obs per year; we use 2011/2012 to get “big sample”
- native born men; 25-30 years since they should have finished school
 - i.e., $\text{age-education-8} \in [25, 30]$
- worked >20 weeks last year; 20+ hours week, earned > \$2800
- pretend we can ignore estimation errors

Characteristics of Native Male Workers, 20-25 Yr Experience, 2011-2012 ACS

Years Education	Number of Observations	Weighted Population	Mean Log Wage	Ethnicity (Percent)	
				Black	Hispanic
(1)	(2)	(3)	(4)	(5)	(6)
0	435	18,948	2.436	16.3%	24.0%
1	16	579	2.339	0.0%	33.7%
2	10	859	2.869	0.0%	59.1%
3	21	1,708	2.682	3.4%	74.0%
4	20	1,413	2.458	9.3%	41.7%
5	32	2,376	2.783	22.6%	44.9%
6	161	10,902	2.507	8.5%	61.0%
7	134	8,572	2.52	9.7%	27.5%
8	599	32,592	2.592	7.8%	18.3%
9	1,058	58,181	2.644	8.4%	23.0%
10	1,863	101,253	2.612	9.6%	14.1%
11	4,375	235,055	2.665	15.9%	15.7%
12	53,764	2,795,826	2.864	13.7%	8.4%
13	23,417	1,310,996	3.041	14.1%	8.3%
14	13,729	704,964	3.116	10.1%	7.0%
16	33,989	1,751,522	3.498	6.7%	4.3%
18	13,620	664,961	3.685	5.6%	3.9%
20	7,303	344,503	3.957	3.6%	3.9%



Implications of $E[x_i u_i] = 0$ (defining property of PRF)

a) If x_i includes a constant, then $E[u_i] = 0$.

why? $E[1 \cdot u_i] = E[u_i] = 0$.

A nice implication is that when x includes a constant:

$$\begin{aligned}E[y_i] &= E[x'_i \beta^* + u_i] \\&= E[x'_i \beta^*] + E[u_i] \\&= E[x'_i \beta^*].\end{aligned}$$

The mean of the PRF is the mean of y !

b) If x_i includes a dummy for membership in subgroup g then $E[u_i|i \in g] = 0$.

why? Let $D_i = 1$ if $i \in g$, and 0 otherwise; let $\mu_g = E[y_i|i \in g]$.

Now:

$$\begin{aligned} E[u_i D_i] &= E[u_i D_i | D_i = 1] \times P(D_i = 1) + E[u_i D_i | D_i = 0] \times P(D_i = 0) \\ &= E[u_i | D_i = 1] \times P(D_i = 1) \end{aligned}$$

If the dummy D_i is included in x_i , we know $E[u_i D_i] = 0$, so

$$E[u_i | D_i = 1] = E[u_i | i \in g] = 0.$$

This is useful because it means:

$$E[u_i | i \in g] = E[y_i - x_i' \beta^* | i \in g] = \mu_g - E[x_i | i \in g]' \beta^* = 0$$

So the population regression *fits the mean* of group g exactly.

c) the “Frisch-Waugh” theorem

The j^{th} row of β^* is:

$$\beta_j^* = E[\xi_i^2]^{-1} E[\xi_i y_i]$$

where ξ_i is the residual from a population regression of x_{ji} on all the other x' s:

$$x_{ji} = x'_{(\sim j)i} \pi + \xi_i.$$

Note that $E[\xi_i^2]^{-1} E[\xi_i y_i]$ is the formula for the population regression of y_i on ξ_i : So FW says that you can think of β_j^* as the coefficient from a univariate regression of y_i on x_{ji} , after “partialling out” all the other x' s.

Proof: $x'_i = (x_{1i}, x_{2i} \dots x_{ji} \dots x_{Ki})$ has K elements.

Let $x_{(\sim j)i}$ be x_i after removing row j .

Now write the “auxilliary” regression of x_{ji} on $x_{(\sim j)i}$:

$$x_{ji} = x'_{(\sim j)i}\pi + \xi_i.$$

As usual, the FOC for π require $E[x_{(\sim j)i}\xi_i] = 0$.

Finally, since $y_i = x'_i\beta^* + u_i$ we can write:

$$\begin{aligned} E[\xi_i y_i] &= E[\xi_i(\beta_1^* x_{1i} + \beta_2^* x_{2i} + \dots + \beta_j^* x_{ji} + \dots + \beta_K^* x_{Ki} + u_i)] \\ &= \beta_1^* E[\xi_i x_{1i}] + \beta_2^* E[\xi_i x_{2i}] + \dots + \beta_j^* E[\xi_i x_{ji}] + \dots + \beta_K^* E[\xi_i x_{Ki}] \\ &\quad + E[\xi_i u_i] \end{aligned}$$

Now notice that from the FOC for π , $E[\xi_i x_{mi}] = 0$ unless $m = j$.

$$\begin{aligned} E[\xi_i y_i] &= \beta_1^* E[\xi_i x_{1i}] + \beta_2^* E[\xi_i x_{2i}] + \dots + \beta_j^* E[\xi_i x_{ji}] + \dots + \beta_K^* E[\xi_i x_{Ki}] \\ &\quad + E[\xi_i u_i] \end{aligned}$$

So $E[\xi_i x_{mi}] = 0$ unless $m = j$

Also: $E[\xi_i u_i] = E[(x_{ji} - x'_{(\sim j)i}\pi)u_i] = 0$ because u_i is orthogonal to all the x' s. So the *only nonzero term* on the r.h.s. is $\beta_j^* E[\xi_i x_{ji}] \Rightarrow$

$$E[\xi_i y_i] = \beta_j^* E[\xi_i x_{ji}]$$

Finally: $E[\xi_i x_{ji}] = E[\xi_i(x'_{(\sim j)i}\pi + \xi_i)] = E[\xi_i^2]$ using the FOC for π (again). So

$$E[\xi_i y_i] = \beta_j^* E[\xi_i^2] \Rightarrow \beta_j^* = E[\xi_i^2]^{-1} E[\xi_i y_i]$$

One extremely useful version of FW: Suppose we have a constant and one other x variable: $x'_i = (1, x_{2i})$. Consider the population regression:

$$y_i = \beta_1^* + \beta_2^* x_{2i} + u_i$$

Then

$$\begin{aligned}\beta_2^* &= E[(x_i - E[x_i])^2]^{-1} E[(x_i - E[x_i])y_i] \\ &= Var[x_i]^{-1} Cov[x_i, y_i]\end{aligned}$$

Why? From FW, we can get β_2^* from a '2 step' approach: first regress x_{2i} on the other regressor (i.e., a constant), then regress y_i on the residual from the first regression. But what is the auxilliary regression of x_{i2} on a constant? This is:

$$x_{i2} = \pi + \xi_i$$

And $\pi = E[x_{i2}]$ is the solution. So in this case, $\xi_i = x_{i2} - E[x_{i2}]$.

In fact, there is a slightly more general version of FW. Suppose we are interested in a subset of regressors, e.g., (x_{1i}, x_{2i}) . Then the coefficients (β_1^*, β_2^*) can be expressed as the outcome of a two-step process: first consider the population regression of (x_{1i}, x_{2i}) on all the other regressors, then consider the population regression of y_i on the pair of residuals.

A version of this result: suppose that $x'_i = (1, x_{2i}, x_{3i}, \dots x_{Ki})$. Then we can get the coefficients on the non-constant regressors by considering the population regression of y on the set of variables $(x_{2i} - E[x_{2i}], x_{3i} - E[x_{3i}], \dots)$. But this is just:

$$\begin{pmatrix} \beta_2^* \\ \beta_3^* \\ \dots \\ \beta_K^* \end{pmatrix} = \text{Var}[x_{2i}, x_{3i}, \dots x_{Ki}]^{-1} \text{Cov}[(x_{2i}, x_{3i}, \dots x_{Ki})', y_i]$$

People often express the pop. regression in terms of variances and covariances, but this is a little sloppy unless y_i and *all the elements* of x_i have mean 0. In that case, you can write:

$$y_i = x'_i \beta^* + u_i$$

$$\beta^* = Var[x_i]^{-1} Cov[x_i, y_i]$$

which is certainly very nice looking!

Now let's move from the population regression to the OLS regression. Recall the objective is

$$\min_{\beta} \sum_{i=1}^N (y_i - x_i' \beta)^2$$

The FOC is:

$$\begin{aligned} \sum_{i=1}^N x_i(y_i - x_i' \hat{\beta}) &= 0 \quad \Rightarrow \frac{1}{N} \sum_{i=1}^N x_i(y_i - x_i' \hat{\beta}) \\ \Rightarrow \frac{1}{N} \sum_{i=1}^N x_i y_i &= \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right) \hat{\beta} \\ \Rightarrow \hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i \end{aligned}$$

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i$$

c/w population regression:

$$\beta^* = E[x_i x_i']^{-1} E[x_i y_i]$$

So we are “matching moments”:

We replace $E[x_i x_i']$ with $S_{xx} = \frac{1}{N} \sum_{i=1}^N x_i x_i'$.

We replace $E[x_i y_i]$ with $S_{xy} = \frac{1}{N} \sum_{i=1}^N x_i y_i$.

Computer programs compute S_{xx}, S_{xy} and invert S_{xx} very efficiently

The 3 properties of the (infeasible) population regression are also true of the OLS regression. For the pop. regression, these come from FOC: $E[x_i(y_i - x'_i\beta^*)] = 0$.

For the OLS regression, these come from FOC:

$$\sum_{i=1}^N x_i(y_i - x'_i\hat{\beta}) = 0$$

- a. if x_i contains a constant, then $\bar{y} = \bar{x}'\hat{\beta}$: the regression model “fits the mean of y ”
- b. if x_i contains a dummy variable for membership in group g then $\bar{y}_g = \bar{x}'_g\hat{\beta}$: the regression model “fits the mean of y for subgroup g ”

c. Frisch-Waugh (FW): The j^{th} row of $\hat{\beta}$ is:

$$\hat{\beta}_j = E[\hat{\xi}_i^2]^{-1} E[\hat{\xi}_i y_i]$$

where $\hat{\xi}_i$ is the *estimated residual* from an OLS regression of x_{ji} on all the other x' s:

$$x_{ji} = x'_{(\sim j)i} \hat{\pi} + \hat{\xi}_i.$$

How are we going to prove FW for OLS?

- (i) OLS: get $\hat{\beta}$, define $\hat{u}_i = y_i - x'_i \hat{\beta}$. We know $\frac{1}{N} \sum_{i=1}^N x_i \hat{u}_i = 0$
- (ii) OLS for auxilliary model: $\hat{\xi}_i = x_{ji} - x'_{(\sim j)i} \hat{\pi}$. We know $\frac{1}{N} \sum_{i=1}^N x_{(\sim j)i} \hat{\xi}_i = 0$
- (iii) write: $y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_j x_{ji} + \dots + \hat{\beta}_K x_{Ki} + \hat{u}_i$

Now form

$$\frac{1}{N} \sum_{i=1}^N \hat{\xi}_i y_i = \frac{1}{N} \sum_{i=1}^N \hat{\xi}_i (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_j x_{ji} + \dots + \hat{\beta}_K x_{Ki} + \hat{u}_i)$$

What terms are equal to 0 from the 2 FOC?

Today's agenda building on Lecture 4:

1. FW for the population regression
2. FW for OLS
3. Omitted variable formula

Frisch-Waugh theorem

Population regression: $y_i = x_i\beta^* + u_i$

Auxilliary regression of x_{ji} on all the other $x's$:

$$x_{ji} = x'_{(\sim j)i}\pi + \xi_i.$$

The j^{th} row of β^* is:

$$\beta_j^* = E[\xi_i^2]^{-1}E[\xi_i y_i]$$

β_j^* = coefficient from univariate regression of y_i on x_{ji} , after “partialling out” other $x's$.

How does the proof work? Use FOC for β^* and π !

$$\begin{aligned} E[\xi_i y_i] &= E[\xi_i(\beta_1^* x_{1i} + \beta_2^* x_{2i} + \dots + \beta_j^* x_{ji} + \dots + \beta_K^* x_{Ki} + u_i)] \\ &= \beta_1^* E[\xi_i x_{1i}] + \beta_2^* E[\xi_i x_{2i}] + \dots + \beta_j^* E[\xi_i x_{ji}] + \dots + \beta_K^* E[\xi_i x_{Ki}] \\ &\quad + E[\xi_i u_i] \end{aligned}$$

FOC for $\pi \Rightarrow E[x_{(\sim j)i} \xi_i] = 0 \Rightarrow E[\xi_i x_{ni}] = 0$ unless $n = j$

FOC for $\beta^* \Rightarrow E[x_i u_i] = 0 \Rightarrow E[\xi_i u_i] = E[(x_{ji} - x'_{(\sim j)i} \pi) u_i] = 0$

And: $E[\xi_i x_{ji}] = E[\xi_i(x'_{(\sim j)i} \pi + \xi_i)] = E[\xi_i^2]$ using the FOC for π (again). So

$$E[\xi_i y_i] = \beta_j^* E[\xi_i^2] \Rightarrow \beta_j^* = E[\xi_i^2]^{-1} E[\xi_i y_i]$$

Suppose we have a constant and one other x variable: $x'_i = (1, x_{2i})$:

$$y_i = \beta_1^* + \beta_2^* x_{2i} + u_i$$

In this case, auxilliary regression is

$$x_{i2} = 1 \bullet \pi + \xi_i$$

And we know $\pi = E[x_{i2}]$ (population regression = CEF). So in this case, $\xi_i = x_{i2} - E[x_{i2}]$. Therefore:

$$\begin{aligned}\beta_2^* &= E[(x_i - E[x_i])^2]^{-1} E[(x_i - E[x_i])y_i] \\ &= Var[x_i]^{-1} Cov[x_i, y_i]\end{aligned}$$

A more general version of FW. Suppose we are interested in a subset of regressors, e.g., (x_{1i}, x_{2i}) . Then the coefficients (β_1^*, β_2^*) can be expressed as the outcome of a two-step process: first consider the population regression of (x_{1i}, x_{2i}) on all the other regressors, then consider the population regression of y_i on the *pair of residuals*.

Application: suppose $x'_i = (1, x_{2i}, x_{3i}, \dots x_{Ki})$. Then we can get the coefficients on the non-constant regressors by considering the population regression of y on the set of variables $(x_{2i} - E[x_{2i}], x_{3i} - E[x_{3i}], \dots)$. But this is just:

$$\begin{pmatrix} \beta_2^* \\ \beta_3^* \\ \dots \\ \beta_K^* \end{pmatrix} = Var[x_{2i}, x_{3i}, \dots x_{Ki}]^{-1} Cov[(x_{2i}, x_{3i}, \dots x_{Ki})', y_i]$$

Now let's move from the population regression to the OLS regression. Recall the OLS objective is

$$\min_{\beta} \sum_{i=1}^N (y_i - x_i' \beta)^2$$

The FOC is:

$$\begin{aligned} \sum_{i=1}^N x_i(y_i - x_i' \hat{\beta}) &= 0 \quad \Rightarrow \frac{1}{N} \sum_{i=1}^N x_i(y_i - x_i' \hat{\beta}) \\ \Rightarrow \frac{1}{N} \sum_{i=1}^N x_i y_i &= \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right) \hat{\beta} \\ \Rightarrow \hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i \end{aligned}$$

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i$$

c/w population regression:

$$\beta^* = E[x_i x_i']^{-1} E[x_i y_i]$$

We replace $E[x_i x_i']$ with $S_{xx} = \frac{1}{N} \sum_{i=1}^N x_i x_i'$ and $\text{plim } S_{xx} = E[x_i x_i']$

We replace $E[x_i y_i]$ with $S_{xy} = \frac{1}{N} \sum_{i=1}^N x_i y_i$ and $\text{plim } S_{xy} = E[x_i y_i]$

So

$$\text{plim } S_{xx}^{-1} S_{xy} = \text{plim } S_{xx}^{-1} \text{plim } S_{xy} = [\text{plim } S_{xx}]^{-1} \text{plim } S_{xy}$$

The 3 properties of the (infeasible) population regression are also true of the OLS regression. For the pop. regression, these come from FOC: $E[x_i(y_i - x'_i\beta^*)] = 0$.

For the OLS regression, these come from FOC:

$$\sum_{i=1}^N x_i(y_i - x'_i\hat{\beta}) = 0$$

- a. if x_i contains a constant, then $\bar{y} = \bar{x}'\hat{\beta}$: the regression model “fits the mean of y ”
- b. if x_i contains a dummy variable for membership in group g then $\bar{y}_g = \bar{x}'_g\hat{\beta}$: the regression model “fits the mean of y for subgroup g ”

c. Frisch-Waugh (FW) for OLS: The j^{th} row of $\hat{\beta}$ is:

$$\hat{\beta}_j = \left[\frac{1}{N} \sum_{i=1}^N \hat{\xi}_i^2 \right]^{-1} \left[\frac{1}{N} \sum_{i=1}^N \hat{\xi}_i y_i \right]$$

where $\hat{\xi}_i$ is the *estimated residual* from an OLS regression of x_{ji} on all the other x' s:

$$x_{ji} = x'_{(\sim j)i} \hat{\pi} + \hat{\xi}_i.$$

How are we going to prove FW for OLS?

- (i) OLS: get $\hat{\beta}$, define $\hat{u}_i = y_i - x'_i \hat{\beta}$. We know $\frac{1}{N} \sum_{i=1}^N x_i \hat{u}_i = 0$
- (ii) OLS for auxilliary model: $\hat{\xi}_i = x_{ji} - x'_{(\sim j)i} \hat{\pi}$. We know $\frac{1}{N} \sum_{i=1}^N x_{(\sim j)i} \hat{\xi}_i = 0$
- (iii) write: $y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_j x_{ji} + \dots + \hat{\beta}_K x_{Ki} + \hat{u}_i$

and form

$$\frac{1}{N} \sum_{i=1}^N \hat{\xi}_i y_i = \frac{1}{N} \sum_{i=1}^N \hat{\xi}_i (\hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_j x_{ji} + \dots + \hat{\beta}_K x_{Ki} + \hat{u}_i)$$

What terms are equal to 0 from the 2 FOC?

Omitted Variables Formula

FW: if you add a regressor, the coefficient is “as if” you added only the part of that regressor that is *unexplained by the other regressors*: $x_{ji} - x'_{(\sim j)i}\pi$ (or $\hat{\pi}$)

What about the opposite direction? What happens if you forget a regressor?

$$y_i = \beta_1^*x_{1i} + \beta_2^*x_{2i} + \beta_3^*x_{ji} + u_i$$

Suppose we don't include x_{3i} ?

a. *Population version*

Aux. model for the omitted variable: $x_{ji} = \pi_1 x_{1i} + \pi_2 x_{2i} + \xi_i$

Then:

$$\begin{aligned}y_i &= \beta_1^* x_{1i} + \beta_2^* x_{2i} + \beta_3^*(\pi_1 x_{1i} + \pi_2 x_{2i} + \xi_i) + u_i \\&= (\beta_1^* + \beta_3^* \pi_1) x_{1i} + (\beta_2^* + \beta_3^* \pi_2) x_{2i} + \beta_3^* \xi_i + u_i \\&= \beta_1^0 x_{1i} + \beta_2^0 x_{2i} + \eta_i\end{aligned}$$

Notice that $E[(x_{1i}, x_{2i})' \eta_i] = E[(x_{1i}, x_{2i})' (\beta_3^* \xi_i + u_i)] = (0, 0)'$.

So (β_1^0, β_2^0) satisfy FOC for population regression of y_i on (x_{1i}, x_{2i})

Conclusion: If

$$y_i = \beta_1^* x_{1i} + \beta_2^* x_{2i} + \beta_3^* x_{ji} + u_i$$

and we don't include x_{3i} , the coefficient on x_{2i} is:

$$\beta_2^0 = \beta_2^* + \beta_3^* \pi_2$$

where π_2 is the coefficient on x_{2i} from the regression of the omitted variable on the remaining x' s:

$$x_{ji} = \pi_1 x_{1i} + \pi_2 x_{2i} + \xi_i$$

Intuition: you forgot x_{3i} so the house elf is doing the best he can to predict y given what he has to work with. The best he can do is use the other x' s to predict x_{3i} .

b. OLS (sample) version

Aux. model for the omitted variable: $x_{ji} = \hat{\pi}_1 x_{1i} + \hat{\pi}_2 x_{2i} + \hat{\xi}_i$

Then:

$$\begin{aligned}y_i &= \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3(\hat{\pi}_1 x_{1i} + \hat{\pi}_2 x_{2i} + \hat{\xi}_i) + \hat{u}_i \\&= (\hat{\beta}_1 + \hat{\beta}_3 \hat{\pi}_1) x_{1i} + (\hat{\beta}_2 + \hat{\beta}_3 \hat{\pi}_2) x_{2i} + \hat{\beta}_3 \hat{\xi}_i + \hat{u}_i \\&= \hat{\beta}_1^0 x_{1i} + \hat{\beta}_2^0 x_{2i} + \hat{\eta}_i\end{aligned}$$

Notice $\frac{1}{N} \sum_{i=1}^N (x_{1i}, x_{2i})' \hat{\eta}_i = \frac{1}{N} \sum_{i=1}^N (x_{1i}, x_{2i})' (\hat{\beta}_3 \hat{\xi}_i + \hat{u}_i) = (0, 0)'$.

So $(\hat{\beta}_1^0, \hat{\beta}_2^0)$ satisfy FOC for OLS regression of y_i on (x_{1i}, x_{2i})

Summary of OLS version: OLS if you used x_{3i} :

$$y_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i} + \hat{u}_i$$

If we don't include x_{3i} , the OLS coefficient on x_{2i} is:

$$\hat{\beta}_2^0 = \hat{\beta}_2 + \hat{\beta}_3 \hat{\pi}_2$$

where $\hat{\pi}_2$ is the coefficient on x_{2i} from the OLS regression of the omitted variable on the remaining x' s:

$$x_{ji} = \hat{\pi}_1 x_{1i} + \hat{\pi}_2 x_{2i} + \hat{\xi}_i$$

Same intuition as for population version.

An example: Suppose we are interested in comparing wages of immigrants and natives. We will use a sample of women age 35-44 who were surveyed in the March 2012 CPS about earnings last year. We consider two models:

$$\log(wage) = \beta_1 + \beta_2 Immigrant + \beta_3 Education \quad (1)$$

and a simpler model:

$$\log(wage) = \beta_1^0 + \beta_2^0 Immigrant \quad (2)$$

Using the OVF, we can relate β_2^0 to β_2 , β_3 , and the coefficient from an auxilliary regression:

$$Education = \pi_1 + \pi_2 Immigrant$$

We know that

$$\beta_2^0 = \beta_2 + \beta_3 \pi_2$$

This holds both for the population regression and for the OLS estimates. What is π_2 ? In this regression we are including a dummy for immigrant status. So in the population version

$$\pi_2 = E[\text{education}|\text{immigrant}] - E[\text{education}|\text{native}]$$

and in the sample version π_2 will equal the difference in mean education between immigrants and natives. This will be a pretty big negative number! And since β_3 is a number like 0.11 we can conclude that if you “leave out” education, you will tend to find that immigrants earn less.

Table 1: Relationships Between Wages, Education and Immigrant Status
for Working Women Age 35-44 in March 2012 Current Population Survey

	Log Wage (1)	Log Wage (2)	Immigrant Status (3)	Years of Education (4)	Log Wage (5)
Immigrant Status	-0.1800 (0.0165)	--	--	-1.4920 (0.0674)	-0.0101 (0.0129)
Years of Education	--	0.1141 (0.0021)	-0.0297 (0.0013)	--	0.1139 (0.0021)
R-squared of model	0.0111	0.2239	0.0442	0.0442	0.2239

Notes: Each column reports separate regression of dependent variable in column heading on regressors shown in rows, plus a constant (coefficient not reported). Sample is females age 35-44 in March 2012 CPS who reported earnings for the last year. "Wage" refers to average hourly earnings last year, n=10,601. Means and standard deviations of dependent variables are: for log wage 2.8527 and 0.6677; for immigrant status 0.1872 and 0.3901; for education 14.1724 and 2.7676.

$$\text{NOTE: } -0.0101 + 0.1139 \times (-1.4920) = -0.1800$$

$$0.1139 - 0.0101 \times (-0.0297) = 0.1141$$

The real importance of the OVF is that we can often think about how the omission of a variable affects the estimated coefficient of variables we include, even if we can't estimate the auxilliary regression. Classic example: suppose the true model is:

$$\log(wage) = \beta_1 + \beta_2 Education + \beta_3 Ability$$

But we don't know "ability", and estimate the simpler model:

$$\log(wage) = \beta_1^0 + \beta_2^0 Education$$

We know that

$$\beta_2^0 = \beta_2 + \beta_3 \pi_2$$

where in this case, π_2 is the coefficient from a regression of ability on education. Many people (especially those with high education) think that $\beta_3 > 0$ and $\pi_2 > 0$, which leads them to believe that a model that does not control for ability "overstates" the effect of education.

Today's agenda – more on OLS regressions

1. goodness of fit measures
2. standard errors of the OLS regression coefficients
3. some examples

We have a regression model for an outcome y with explanatory variables x . Write the population model as:

$$y_i = x'_i \beta^* + u_i.$$

In a sample (size N) the FOC for the OLS estimator $\hat{\beta}$ is:

$$\begin{aligned} \sum_{i=1}^N x_i(y_i - x'_i \hat{\beta}) &= 0 \quad \Rightarrow \frac{1}{N} \sum_{i=1}^N x_i(y_i - x'_i \hat{\beta}) \\ \Rightarrow \frac{1}{N} \sum_{i=1}^N x_i y_i &= \left(\frac{1}{N} \sum_{i=1}^N x_i x'_i \right) \hat{\beta} \\ \Rightarrow \hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N x_i x'_i \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i \end{aligned}$$

How well does the model fit? A “null” model is one with just a constant. If we fit that model, we know the estimated coefficient will be the mean of y_i . The “sum of squares” of y_i is:

$$SS = \sum_{i=1}^N (y_i - \bar{y})^2$$

which is the sum of the squared prediction errors from a model with only a constant. From our OLS model, the corresponding “sum of squared residuals” is

$$SSR = \sum_{i=1}^N (y_i - x_i \hat{\beta})^2 = \sum_{i=1}^N \hat{u}_i^2.$$

We define:

$$R^2 = 1 - (SSR/SS)$$

$$\begin{aligned}
 R^2 &= 1 - (SSR/SS) \\
 &= 1 - \frac{\frac{1}{N} \sum_{i=1}^N \hat{u}_i^2}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}
 \end{aligned}$$

This can range from 0 (model is no better than a constant) to 1 (model explains y perfectly).

One problem with R^2 is that it can only get better as you add more x 's – there is no penalty for adding extra variables even if they don't add much. Adjusted R^2 , denoted by \bar{R}^2 , corrects for the use of extra explanatory variables. The idea is to replace the two terms in R^2 with “degrees-of-freedom-corrected” terms:

$$R^2 = 1 - \frac{\frac{1}{N} \sum_{i=1}^N \hat{u}_i^2}{\frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2}$$

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K} \sum_{i=1}^N \hat{u}_i^2}{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2}$$

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-K} \sum_{i=1}^N \hat{u}_i^2}{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2}$$

The term $\frac{1}{N-K} \sum_{i=1}^N \hat{u}_i^2$ is usually called the “mean squared error” (MSE) of the model. The intuition for dividing by $N - K$ is this: if you have K explanatory variables, and you had a sample of size $N = K$, then you could perfectly predict all the observations, and $SSR = 0$. So the “degrees of freedom” in the SSR is $N - K$. It is also conventional to estimate the variance of the regression residual by the MSE:

$$\widehat{Var}[u_i] = \frac{1}{N - K} \sum_{i=1}^N \hat{u}_i^2$$

Inference

After we estimate the coefficients in a model, we want to know how likely it is that our estimates are “close” to the truth (where “truth” means the population regression coefficients). This will depend on the sample size, how much variability there is in y , and on how easy it is to separate out the effects of the various elements of x . We begin with the “classic” case where the X ’s are “fixed” and we have normally distributed residuals.

We assume the population model is:

$$y_i = x'_i \beta^* + u_i.$$

We also assume:

1. we have an independent sample of size N
2. the elements of x_i are not linearly dependent
3. given the x 's each of the u_i are iid normals: $u_i \sim N(0, \sigma_u^2)$

Under the previous assumptions, the vector of OLS coefficients is distributed normally conditional on the sample x 's:

$$\hat{\beta} - \beta^* \sim N(0, V)$$

$$V = \frac{\sigma_u^2}{N} S_{xx}^{-1}, \quad \text{and} \quad S_{xx} = \frac{1}{N} \sum_{i=1}^N x_i x_i'.$$

To prove this we write:

$$\begin{aligned}
 \hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i \\
 &= S_{xx}^{-1} \frac{1}{N} \sum_{i=1}^N x_i (x_i' \beta^* + u_i) \\
 &= S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right) \beta^* + S_{xx}^{-1} \frac{1}{N} \sum_{i=1}^N x_i u_i \\
 &= \beta^* + \sum_{i=1}^N a_i(X) u_i
 \end{aligned}$$

where $a_i(X) = \frac{1}{N} S_{xx}^{-1} x_i$ is a $K \times 1$ vector that is a function of all the x_i' s (X).

$$\hat{\beta} - \beta^* = \sum_{i=1}^N a_i(X) u_i$$

So conditional on X , each row of $\hat{\beta} - \beta^*$ is just a weighted sum of the u'_i s. Thus, conditional on X , $\hat{\beta} - \beta^*$ is normally distributed! We have:

$$E[\hat{\beta} - \beta^*|X] = E\left[\sum_{i=1}^N a_i(X) u_i|X\right] = \sum_{i=1}^N a_i(X) E[u_i|X] = 0$$

What about the variance of $\hat{\beta} - \beta^*$?

To calculate the variance of $\hat{\beta}$, note:

$$\hat{\beta} - \beta^* = \sum_{i=1}^N a_i u_i$$

There are K rows of this expression:

$$\begin{array}{ll} \hat{\beta}_1 - \beta_1^* & \sum_{i=1}^N a_{1i} u_i \\ \hat{\beta}_2 - \beta_2^* & \sum_{i=1}^N a_{2i} u_i \\ \dots & \dots \\ \hat{\beta}_K - \beta_K^* & \sum_{i=1}^N a_{Ki} u_i \end{array}$$

So:

$$\begin{aligned}Var[\hat{\beta}_j - \beta_j^* | X] &= \sigma_u^2 \sum_{i=1}^N a_{ji}^2 \\Cov[\hat{\beta}_j - \beta_j^*, \hat{\beta}_k - \beta_k^* | X] &= \sigma_u^2 \sum_{i=1}^N a_{ji} a_{ki}\end{aligned}$$

since the u_i' s are independent of each other. In matrix form:

$$Var[\hat{\beta} - \beta^* | X] = \sigma_u^2 \sum_{i=1}^N a_i a_i'$$

Thus

$$\begin{aligned}Var[\hat{\beta} - \beta^* | X] &= \sigma_u^2 \sum_{i=1}^N a_i a'_i \\&= \sigma_u^2 \sum_{i=1}^N \frac{1}{N^2} S_{xx}^{-1} x_i x'_i S_{xx}^{-1} \\&= \frac{1}{N} \sigma_u^2 S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i x'_i \right) S_{xx}^{-1} \\&= \frac{1}{N} \sigma_u^2 S_{xx}^{-1}\end{aligned}$$

A very important example: $x'_i = (1, x_{2i})$ – we have a constant and 1 other regressor.

$$\begin{aligned}
 S_{xx} &= \frac{1}{N} \sum_{i=1}^N \begin{pmatrix} 1 \\ x_{2i} \end{pmatrix} (1 \quad x_{2i}) \\
 &= \frac{1}{N} \sum_{i=1}^N \begin{pmatrix} 1 & x_{2i} \\ x_{2i} & x_{2i}^2 \end{pmatrix} \\
 &= \begin{pmatrix} 1 & \bar{x}_2 \\ \bar{x}_2 & \frac{1}{N} \sum_{i=1}^N x_{2i}^2 \end{pmatrix} \\
 S_{xx}^{-1} &= \frac{1}{\frac{1}{N} \sum_{i=1}^N x_{2i}^2 - \bar{x}_2^2} \begin{pmatrix} \frac{1}{N} \sum_{i=1}^N x_{2i}^2 & -\bar{x}_2 \\ -\bar{x}_2 & 1 \end{pmatrix}
 \end{aligned}$$

$$S_{xx}^{-1} = \frac{1}{\frac{1}{N} \sum_{i=1}^N x_{2i}^2 - \bar{x}_2^2} \begin{pmatrix} \frac{1}{N} \sum_{i=1}^N x_{2i}^2 & -\bar{x}_2 \\ -\bar{x}_2 & 1 \end{pmatrix}$$

The (2,2) element is:

$$\begin{aligned} (S_{xx}^{-1})_{2,2} &= \frac{1}{\frac{1}{N} \sum_{i=1}^N x_{2i}^2 - \bar{x}_2^2} \\ &= \frac{1}{\frac{1}{N} \sum_{i=1}^N (x_{2i} - \bar{x}_2)^2} \\ &= \frac{1}{\hat{\sigma}^2(x_2)} \end{aligned}$$

So, conditional on X ,

$$\hat{\beta}_2 - \beta_2^* \sim N\left(0, \frac{1}{N} \frac{\sigma_u^2}{\hat{\sigma}^2(x_2)}\right)$$

which is a formula you should recall from 140/141. Notice that we don't know σ_u^2 : we estimate it using the MSE:

$$\hat{\sigma}_u^2 = \frac{1}{N-2} \sum_{i=1}^N \hat{u}_i^2$$

which makes sense because we are assuming each u_i has the same variance.

In a standard regression package, the estimated “sampling errors” of the OLS coefficients are calculated “as if” the classic normal model were true. In particular, the reported variance/covariance matrix of the estimated OLS coefficients is:

$$\frac{1}{N} \hat{\sigma}_u^2 S_{xx}^{-1}$$

where $S_{xx} = \frac{1}{N} \sum_i x_i x_i'$. Notice the three terms:

- the $1/N$ term reflects the sample size
- the $\hat{\sigma}_u^2$ term reflects our estimate of the variability in y_i conditional on x_i
- the S_{xx}^{-1} term reflects the difficulty of pulling apart the contributions of the different rows of x_i

So we have $\hat{\beta} - \beta^* \sim N(0, V)$, and an estimate of V , $\hat{V} = \frac{1}{N} \hat{\sigma}_u^2 S_{xx}^{-1}$. How do we conduct inference?

1. To test the null $\beta_k^* = 0$: we form the ratio: $\hat{\beta}_k / \hat{V}_{kk}$ and compare this to a $N(0, 1)$

If $|\hat{\beta}_k / \hat{V}_{kk}| > 1.96$ then we “reject the null” under a 2-sided test at 95% confidence

2. How do we test $b\beta_k^* - c\beta_j^* = 0$?

Note $Var[b\beta_k^* - c\beta_j^*] = b^2 Var[\beta_k^*] + c^2 Var[\beta_j^*] - 2bc Cov[\beta_k^*, \beta_j^*]$

So we pull out the corresponding elements of \hat{V} , and form the ratio:

$$\frac{b\hat{\beta}_k - c\hat{\beta}_j}{b^2 \hat{V}_{kk} + c^2 \hat{V}_{jj} - 2bc \hat{V}_{kj}}$$

There is a potentially serious limitation of the “classic normal” model. In lots of cases, we think that $E[y_i|x_i]$ is not really linear. The classic case says that

$$E[y_i|x_i] = x_i' \beta^*$$

When that is not true, we know that the “true error” for the i th observation is

$$u_i = y_i - E[y_i|x_i] + E[y_i|x_i] - x_i' \beta^*$$

In this case the error at each value of x_i can have a non-zero mean!

So we’ll have to “tech up” our sampling errors!

Sampling errors for OLS regressions

1. review of “classic normal” case from Lecture 6
2. more general case (somewhat advanced)

Recap: The population regression is:

$$y_i = x'_i \beta^* + u_i.$$

We assume:

1. independent sample of size N
2. no linear dependency in x_i

Classic normal case:

given the x 's each of the u_i are iid normals: $u_i \sim N(0, \sigma_u^2)$

Write:

$$\begin{aligned}
 \hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i \\
 &= S_{xx}^{-1} \frac{1}{N} \sum_{i=1}^N x_i (x_i' \beta^* + u_i) \\
 &= S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right) \beta^* + S_{xx}^{-1} \frac{1}{N} \sum_{i=1}^N x_i u_i \\
 &= \beta^* + \sum_{i=1}^N a_i(X) u_i
 \end{aligned}$$

where $a_i(X) = \frac{1}{N} S_{xx}^{-1} x_i$. $\hat{\beta} - \beta^*$ is a weighted sum of the u_i' s \Rightarrow normal distribution, given X .

$$\hat{\beta} - \beta^* = \sum_{i=1}^N a_i(X) u_i$$

$$E[\hat{\beta} - \beta^* | X] = E\left[\sum_{i=1}^N a_i(X) u_i | X\right] = \sum_{i=1}^N a_i(X) E[u_i | X] = 0$$

so we know

$$E[\hat{\beta} | X] = \beta^*$$

Now we want to get the *matrix of variances and covariances*, $V_\beta \equiv \text{Var}[\hat{\beta} - \beta^*|X]$. This is a $K \times K$ matrix.

The (r, r) element is:

$$\text{Var}[\hat{\beta}_r - \beta_r^*|X] = E[(\hat{\beta}_r - \beta_r^*)^2|X]$$

The (r, s) element is:

$$\text{Cov}[\hat{\beta}_r - \beta_r^*, \hat{\beta}_s - \beta_s^*|X] = E[(\hat{\beta}_r - \beta_r^*)(\hat{\beta}_s - \beta_s^*)|X]$$

We know $\hat{\beta} - \beta^* = \sum_{i=1}^N a_i u_i \Rightarrow \hat{\beta}_r - \beta_r^* = \sum_{i=1}^N a_{ri} u_i$

AND all the u'_i s are uncorrelated. So:

$$\begin{aligned} Var[(\hat{\beta}_r - \beta_r^*)|X] &= \sum_{i=1}^N a_{ri}^2 \sigma_u^2 = \sigma_u^2 \sum_{i=1}^N a_{ri}^2 \\ Cov[\hat{\beta}_r - \beta_r^*, \hat{\beta}_s - \beta_s^*|X] &= \sum_{i=1}^N a_{ri} a_{si} \sigma_u^2 = \sigma_u^2 \sum_{i=1}^N a_{ri} a_{si} \\ \Rightarrow Var[\hat{\beta} - \beta^*|X] &= \sigma_u^2 \sum_{i=1}^N a_i a'_i \end{aligned}$$

So we've shown

$$Var[\hat{\beta}_r - \beta_r^*|X] = \sigma_u^2 \sum_{i=1}^N a_i a_i'$$

and using $a_i = \frac{1}{N} S_{xx}^{-1} x_i$, $a_i' = \frac{1}{N} x_i' S_{xx}^{-1}$ (since S_{xx} and S_{xx}^{-1} are symm.)

$$\begin{aligned} Var[\hat{\beta}_r - \beta_r^*|X] &= \sigma_u^2 \sum_{i=1}^N \frac{1}{N^2} S_{xx}^{-1} x_i x_i' S_{xx}^{-1} \\ &= \frac{1}{N} \sigma_u^2 S_{xx}^{-1} \end{aligned}$$

In standard regression packages, the reported matrix of samp. errors is:

$$\hat{V}_\beta = \frac{1}{N} \hat{\sigma}_u^2 S_{xx}^{-1}$$

There are 2 limitations of the “classic normal” model.

1. The classic case says that $E[y_i|x_i] = x_i'\beta^*$. When that is not true, the error for the i^{th} observation is

$$\begin{aligned} u_i &= y_i - E[y_i|x_i] + E[y_i|x_i] - x_i'\beta^* \\ &= \varepsilon_i + v_i \end{aligned}$$

and $v_i \equiv E[y_i|x_i] - x_i'\beta^*$ (the specification error) depends on x_i .

2. In lots of cases $u_i \sim Normal$ is not appropriate (e.g., $y_i = 1[z_i > 0]$ discrete)

So in this lecture we “tech up” our sampling errors!

We need 4 key results from statistics. Suppose we have an iid sample of size N ; A_N is $K \times K$ matrix of sample statistics with property $\text{plim } A_N = A$

1. If A is invertible then $\text{plim } (A_N)^{-1} = (\text{plim } A_N)^{-1} = A^{-1}$
2. If B_N is $K \times 1$ vector with $\text{plim } B_N = 0$, then $\text{plim } A_N B_N = 0$.
3. (vector CLT). If z_i i.i.d $(K \times 1)$ vector with $E[z_i] = 0$, $\text{Var}[z_i] = V$ then:

$$\sqrt{N}\bar{z}_N = \frac{1}{\sqrt{N}} \sum_{i=1}^N z_i \xrightarrow{a} N(0, V)$$

4. If $\sqrt{N}B_N \xrightarrow{a} N(0, V)$ then $\sqrt{N}A_N B_N \xrightarrow{a} N(0, AVA')$

Using these results:

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^N x_i x'_i \right)^{-1} \frac{1}{N} \sum_{i=1}^N x_i y_i \\&= S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i x'_i \right) \beta^* + S_{xx}^{-1} \frac{1}{N} \sum_{i=1}^N x_i u_i \\&= \beta^* + S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i u_i \right) \\&\Rightarrow \hat{\beta} - \beta^* = S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i u_i \right)\end{aligned}$$

$$\Rightarrow \hat{\beta} - \beta^* = S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i u_i \right)$$

Let $z_i = x_i u_i$. We know $E[z_i] = E[x_i u_i] = 0$ (FOC for β^*).

Let $E[(z_i - E[z_i])(z_i - E[z_i])'] = V$ (the variance of the r.v. z_i)

We know from *w.l.l.n* that $\text{plim } \frac{1}{N} \sum_{i=1}^N z_i = E[z_i] = 0$

We know from vector CLT that $\left(\frac{1}{\sqrt{N}} \sum_{i=1}^N z_i \right) \xrightarrow{a} N(0, V)$

Assume: $\text{plim } S_{xx} = \text{plim } \left(\frac{1}{N} \sum_{i=1}^N x_i x'_i \right) = S_{xx}^*$ exists and is invertible

and $\text{Var}[x_i u_i] = V$

Then

$$\text{plim } \hat{\beta} - \beta^* = 0 \quad (1)$$

$$\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{a} N(0, [S_{xx}^*]^{-1} V [S_{xx}^*]^{-1}) \quad (2)$$

How to prove (1): $\text{plim } \hat{\beta} - \beta^* = 0$?

$$\begin{aligned}\text{plim } \hat{\beta} - \beta^* &= \text{plim } S_{xx}^{-1} \left(\frac{1}{N} \sum_{i=1}^N x_i u_i \right) \\ &= [S_{xx}^*]^{-1} \text{plim } \frac{1}{N} \sum_{i=1}^N x_i u_i = 0\end{aligned}$$

How to prove (2): $\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{a} N(0, [S_{xx}^*]^{-1} V [S_{xx}^*]^{-1})$?

$$\sqrt{N}(\hat{\beta} - \beta^*) = S_{xx}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^N x_i u_i$$

Now $\text{plim } S_{xx}^{-1} = [S_{xx}^*]^{-1}$. And by vecCLT, $\frac{1}{\sqrt{N}} \sum_{i=1}^N x_i u_i \xrightarrow{a} N(0, V)$.
So using “result 4” we’re done!

So in the “general case”, we have

$$(\hat{\beta} - \beta^*) \approx N(0, \frac{1}{N} [S_{xx}^*]^{-1} V [S_{xx}^*]^{-1})$$

where $V = \text{Var}[x_i u_i]$

In the classic normal case, we have (conditional on X)

$$(\hat{\beta} - \beta^*) \sim N(0, \frac{1}{N} [S_{xx}]^{-1} \sigma_u^2)$$

where $\sigma_u^2 = \text{Var}[u_i]$.

What's the difference?

-approx. vs exact distribution

- we aren't conditioning on X
- we don't assume $E[u_i|x_i] = 0$, only $E[x_i u_i] = 0$

$$N(0, \frac{1}{N} [S_{xx}^*]^{-1} V [S_{xx}^*]^{-1}) \quad vs. \quad N(0, \frac{1}{N} [S_{xx}]^{-1} \sigma_u^2)$$

How do we actually estimate the var-cov, \hat{V}_β in the general case?

a) Approximate $S_{xx}^* = S_{xx}$. so estimate of $[S_{xx}^*]^{-1}$ is $[S_{xx}]^{-1}$

b) estimate

$$\hat{V} = \frac{1}{N} \sum_i (x_i \hat{u}_i)(x_i \hat{u}_i)' = \frac{1}{N} \sum \hat{u}_i^2 (x_i x_i')$$

c) form $\hat{V}_\beta = \frac{1}{N} S_{xx}^{-1} \hat{V} S_{xx}^{-1}$

We are estimating: $\hat{V} = \frac{1}{N} \sum_i \hat{u}_i^2 (x_i x_i')$. In the 3×3 case:

$$\hat{V} = \frac{1}{N} \sum_i \begin{pmatrix} \hat{u}_i^2 & \hat{u}_i^2 x_{2i} & \hat{u}_i^2 x_{3i} \\ \hat{u}_i^2 x_{2i} & \hat{u}_i^2 x_{2i}^2 & \hat{u}_i^2 x_{3i} x_{2i} \\ \hat{u}_i^2 x_{3i} & \hat{u}_i^2 x_{3i} x_{2i} & \hat{u}_i^2 x_{3i}^2 \end{pmatrix}$$

What happens if \hat{u}_i^2 and $x_i x_i'$ are uncorrelated across obs?

USEFUL FACTOID:

If two random variables a_i and b_i are uncorrelated in a sample then

$$\frac{1}{N} \sum_i a_i b_i = \frac{1}{N} \sum_i a_i \frac{1}{N} \sum_i b_i = \bar{a}\bar{b}$$

Proof: write: $a_i = \bar{a} + a_i - \bar{a}$; $b_i = \bar{b} + b_i - \bar{b}$. Finish as an exercise.

So if \hat{u}_i^2 and $x_i x'_i$ uncorrelated in the sample,

$$\hat{V} = \frac{1}{N} \sum_i \hat{u}_i^2 (x_i x'_i) = \frac{1}{N} \sum_i \hat{u}_i^2 \frac{1}{N} \sum_i x_i x'_i$$

In the uncorrelated case

$$\hat{V}_\beta = \frac{1}{N} S_{xx}^{-1} \hat{V} S_{xx}^{-1} = \frac{1}{N} \left(\frac{1}{N} \sum_i \hat{u}_i^2 \right) S_{xx}^{-1}$$

$$\Rightarrow \hat{V}_\beta = \frac{1}{N} \tilde{\sigma}_u^2 S_{xx}^{-1}$$

which is the same as the classic normal except $\tilde{\sigma}_u^2$ does not correct for d.f.!

We have $\hat{\beta} - \beta^* \sim N(0, V_\beta)$, and an estimate of $\hat{V}_\beta = \frac{1}{N} \hat{\sigma}_u^2 S_{xx}^{-1}$ or $\frac{1}{N} S_{xx}^{-1} \hat{V} S_{xx}^{-1}$

How do we conduct inference?

1. To test the null $\beta_k^* = 0$: we form the ratio: $\hat{\beta}_k / \sqrt{[\hat{V}_\beta]_{k,k}}$ and compare this to a $N(0, 1)$

If $|\hat{\beta}_k / \sqrt{[\hat{V}_\beta]_{k,k}}| > 1.96$ then we “reject the null” under a 2-sided test at 95% confidence

2. How do we test $b\beta_k^* - c\beta_j^* = 0$?

Note $Var[b\beta_k^* - c\beta_j^*] = b^2Var[\beta_k^*] + c^2Var[\beta_j^*] - 2bcCov[\beta_k^*, \beta_j^*]$

So we pull out the corresponding elements of \hat{V} , and form the ratio:

$$\frac{b\hat{\beta}_k - c\hat{\beta}_j}{\sqrt{b^2 [\hat{V}_\beta]_{k,k} + c^2 [\hat{V}_\beta]_{j,j} - 2bc [\hat{V}_\beta]_{k,j}}}$$

Today's topic:

using regression models to “decompose” differences in means

Very widely used in applied studies e.g.

- differences in wages between men and women
- differences in college completion rates between 2 groups

Originally invented by Ron Oaxaca – “Oaxaca decomposition”

Assume to start that there is a “population model”:

$$y_i = x'_i \beta^* + u_i.$$

In our example for today, y_i is log wage of person i , x_i is i 's education (and possibly other characteristics).

We are going to be working with the OLS estimator $\hat{\beta}$. We recall the critical FOC:

$$\sum_{i=1}^N x_i(y_i - x'_i \hat{\beta}) = 0$$

We will assume $x'_i = (1, x_{2i}, \dots x_{Ki})$ – the first regressor is a constant.

We will assume there are two groups, a and b . Let \bar{y}^a represent the mean of y_i for group a , and let \bar{x}^a represent the mean of the vector of x 's for group a . Suppose we fit a model

$$y_i = \beta_1 + \sum_{j=2}^K x_{ji}\beta_j + u_i$$

just for group a (with sample size N^a). Then we know:

$$\bar{y}^a = \hat{\beta}_1 + \sum_{j=2}^K \bar{x}_j^a \hat{\beta}_j = \bar{x}^a \hat{\beta}$$

Why?

Refresher. The FOC require:

$$\sum_{i=1}^N x_i(y_i - x'_i \hat{\beta}) = 0$$

But one of the elements of x_i is 1 so (dividing by N^a) :

$$\frac{1}{N^a} \sum_{i=1}^N 1(y_i - x'_i \hat{\beta}) = 0$$

$$\Rightarrow \bar{y}^a = \frac{1}{N^a} \sum_{i=1}^N y_i = \frac{1}{N^a} \sum_{i=1}^N x'_i \hat{\beta} = \bar{x}^a \hat{\beta}$$

Intuitively, the constant can always be selected so that, given the other $\hat{\beta}'s$ the predictions from the regression model fit the mean – so that is what the regression does!

Now let's add the second group to the sample, and add a dummy variable indicating observations from group b : $x_{iK+1} = D_i = 1[i \in b]$. For this new model (with total sample size $N = N^a + N^b$) it will be true that:

$$\begin{aligned}\bar{y}^a &= \bar{x}^{a\prime} \hat{\beta} \\ \bar{y}^b &= \bar{x}^{b\prime} \hat{\beta}\end{aligned}$$

Why? From the row of the FOC corresponding to β_{K+1} :

$$\sum_{i=1}^N D_i(y_i - x_i' \hat{\beta}) = 0$$

But this requires that

$$\sum_{i \in b} (y_i - x_i' \hat{\beta}) = 0$$

FOC dummy for group b requires:

$$\sum_{i \in b} (y_i - x'_i \hat{\beta}) = 0$$

$$\Rightarrow \bar{y}^b = \frac{1}{N^b} \sum_{i \in b} y_i = \frac{1}{N^b} \sum_{i \in b} x'_i \hat{\beta} = (\bar{x}^b)' \hat{\beta}$$

And the FOC for the constant terms requires

$$\sum_i (y_i - x'_i \hat{\beta}) = \sum_{i \in a} (y_i - x'_i \hat{\beta}) + \sum_{i \in b} (y_i - x'_i \hat{\beta})$$

$$\begin{aligned}\Rightarrow \sum_{i \in a} (y_i - x'_i \hat{\beta}) &= 0 \\ \Rightarrow \bar{y}^a - (\bar{x}^a)' \hat{\beta} &= 0\end{aligned}$$

As a result, in the pooled model with a constant and a dummy for group b we know:

$$\begin{aligned}\bar{y}^a &= \hat{\beta}_1 + \sum_{j=2}^K \bar{x}_j^a \hat{\beta}_j \\ \bar{y}^b &= \hat{\beta}_1 + \sum_{j=2}^K \bar{x}_j^b \hat{\beta}_j + \hat{\beta}_{K+1}\end{aligned}$$

Thus:

$$\bar{y}^b - \bar{y}^a = \sum_{j=2}^K (\bar{x}_j^b - \bar{x}_j^a) \hat{\beta}_j + \hat{\beta}_{K+1}$$

We can use this expression to “decompose” the difference in means. For example, if regressor number 2 is “education”, and $\hat{\beta}_2 = 0.10$, $\bar{x}_2^b = 12$ and $\bar{x}_2^a = 14$ then differences in education explain a gap of $(\bar{x}_2^b - \bar{x}_2^a) \hat{\beta}_2 = (12 - 14) \times 0.10 = -0.20$

We have shown that

$$\bar{y}^b - \bar{y}^a = \sum_{j=2}^K (\bar{x}_j^b - \bar{x}_j^a) \hat{\beta}_j + \hat{\beta}_{K+1}$$

Let's think of the special case where our model only has 2 explanatory variables: a constant, and a dummy for group b . In this case,

$$\begin{aligned}\bar{y}^a &= \hat{\beta}_1 \\ \bar{y}^b &= \hat{\beta}_1 + \hat{\beta}_{K+1}\end{aligned}$$

which implies that the coefficient on the dummy for group b is

$$\hat{\beta}_{K+1} = \bar{y}^b - \bar{y}^a$$

which we knew from Lecture 3! When we add other explanatory variables, however, the estimate will (in general) change.

Example: let's look at our 2012 sample from the CPS. Here we will focus on men, age 30-35, and consider group a = natives and group b = immigrants. Some relevant information:

Natives:

mean log wage = 3.0129

mean education = 14.092 years

Immigrants:

mean log wage = 2.7660

mean education = 12.409 years

Pooled Model: Fit to Natives and Immigrants			
	(1)	(2)	
Constant	3.013 (0.006)	1.546 (0.025)	Difference in mean wages
Immigrant	-0.247 (0.013)	-0.072 (0.013)	Difference in mean wages after "controlling" for education
Education (yrs)	--	0.104 (0.002)	
MSE	0.757	0.695	
Adj. R-sq	0.018	0.173	
Sample Size	19,092	19,092	

Notes: Fit to data for males age 30-45 in March 2012 CPS.

Dependent variable is log average hourly wage. Mean and standard deviation are 2.959 (0.764). Standard errors in parentheses.

Let's perform the decomposition. We have $K = 2$, with the second variable being education.

$$\bar{y}^b - \bar{y}^a = 2.766 - 3.013 = -0.247$$

From the model in column 2 of the table, we have that

$$\bar{y}^b - \bar{y}^a = \sum_{j=2}^K (\bar{x}_j^b - \bar{x}_j^a) \hat{\beta}_j + \hat{\beta}_{K+1}$$

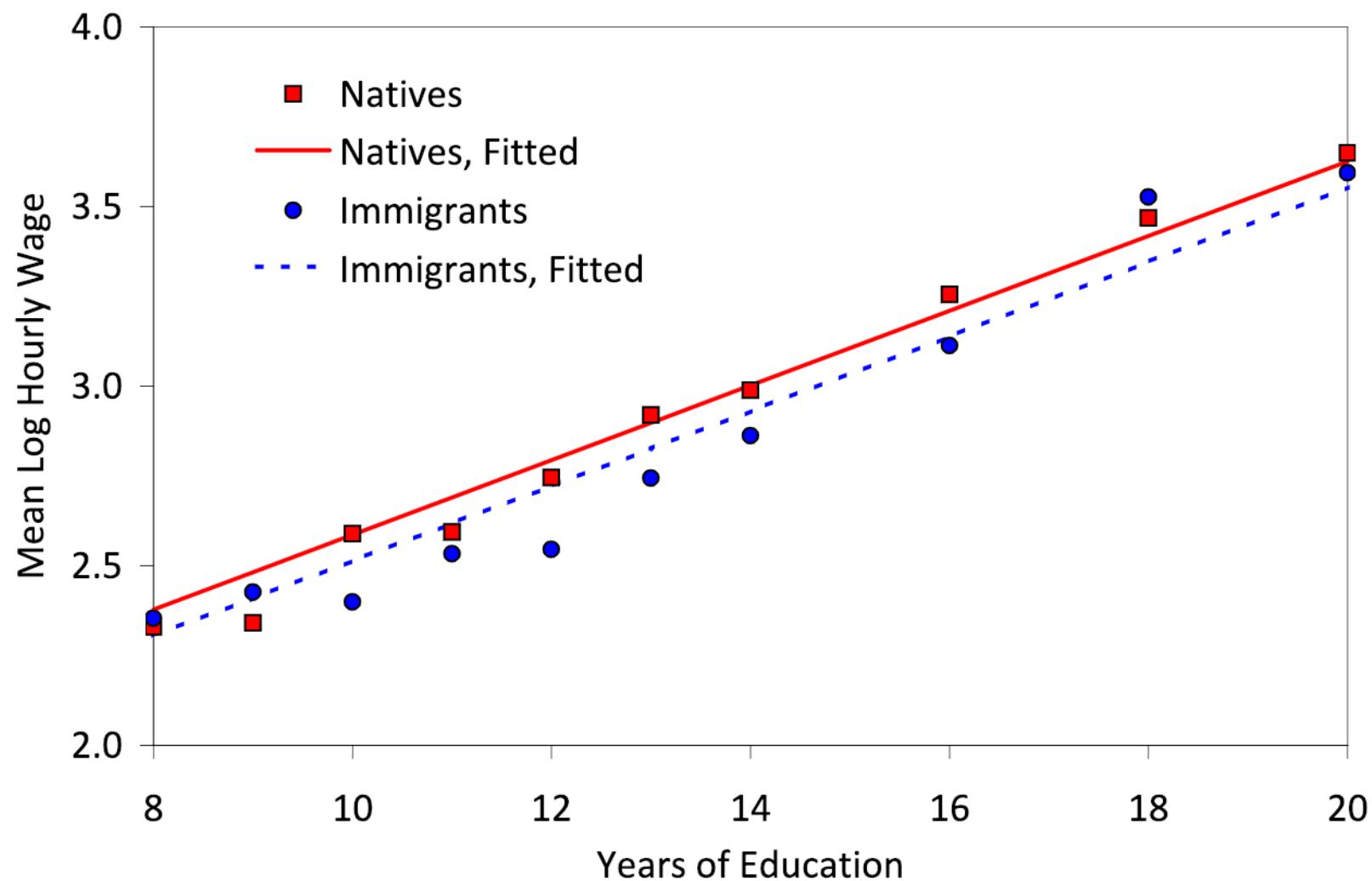
$$-0.247 = (12.409 - 14.092) \times 0.1041 - 0.0718$$

So the “effect of education” is $-1.683 \times 0.1041 = -0.175$ which is 70.9% of the wage gap. The remainder (29.1%) is “unexplained”

	Pooled Model: Fit to Natives and Immigrants		Model for Natives	Model for Immigrants
	(1)	(2)	(3)	(4)
Constant	3.013 (0.006)	1.546 (0.025)	1.365 (0.033)	1.676 (0.035)
Immigrant	-0.247 (0.013)	-0.072 (0.013)	--	--
Education (yrs)	--	0.104 (0.002)	0.117 (0.002)	0.088 (0.002)
MSE	0.757	0.695	0.689	0.707
Adj. R-sq	0.018	0.173	0.146	0.208
Sample Size	19,092	19,092	14,921	4,141

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are: for overall sample, 2.959 (0.764); for natives 3.013 (0.746); for immigrants 2.766 (0.795). Standard errors in parentheses.

Wages by Education -- Males Age 30-45



A common problem that arises in applying this idea is that the coefficients of the explanatory variables are not necessarily the same in the two samples. This gives rise to a more general version, which is “the” Oaxaca decomposition. Suppose we fit our model separately in the two subsamples. For group a we obtain OLS estimates $\hat{\beta}^a$, and (since we have a constant in the model) we know

$$\bar{y}^a = (\bar{x}^a)' \hat{\beta}^a.$$

For group b we obtain OLS estimates $\hat{\beta}^b$, and we know

$$\bar{y}^b = (\bar{x}^b)' \hat{\beta}^b.$$

So we can construct

$$\bar{y}^b - \bar{y}^a = (\bar{x}^b)' \hat{\beta}^b - (\bar{x}^a)' \hat{\beta}^a$$

Now let's manipulate this expression:

$$\begin{aligned}\bar{y}^b - \bar{y}^a &= (\bar{x}^b)' \hat{\beta}^b - (\bar{x}^a)' \hat{\beta}^a \\&= (\bar{x}^b - \bar{x}^a)' \hat{\beta}^a + (\bar{x}^b)' (\hat{\beta}^b - \hat{\beta}^a) \\&= (\bar{x}^b - \bar{x}^a)' \hat{\beta}^b + (\bar{x}^a)' (\hat{\beta}^b - \hat{\beta}^a)\end{aligned}$$

These are both algebraically true. The first says that the difference is the difference in mean x 's, weighted by the estimated coefficients from group a , plus the difference in the coefficients, weighted by the mean from group b . The second reverses the groups. Note that if $\hat{\beta}^a = \hat{\beta}^b = \hat{\beta}$ we get our previous method, taking account of the fact that our previous model had a dummy for group b included as one of the x 's.

Let's apply this to our example. Here we have

$$\hat{\beta}_2^a = 0.117$$

$$\hat{\beta}_2^b = 0.088$$

$$(\bar{x}_2^b - \bar{x}_2^a) = 12.409 - 14.092 = 1.683$$

And we know $\bar{y}^b - \bar{y}^a = -0.247$. So if we use the coefficient for natives we have:

$$(\bar{x}_2^b - \bar{x}_2^a)\hat{\beta}_2^a = -0.197$$

$$\bar{x}_2^b(\hat{\beta}_2^b - \hat{\beta}_2^a) = -0.360$$

Whereas if we use the coefficient for immigrants we have

$$(\bar{x}_2^b - \bar{x}_2^a)\hat{\beta}_2^b = -0.148$$

$$\bar{x}_2^a(\hat{\beta}_2^b - \hat{\beta}_2^a) = -0.409$$

This shows a couple of important things. First, we have 2 estimates of the contribution of the difference in mean education:

$$\begin{aligned}(\bar{x}_2^b - \bar{x}_2^a)\hat{\beta}_2^a &= -0.197 \\(\bar{x}_2^b - \bar{x}_2^a)\hat{\beta}_2^b &= -0.148\end{aligned}$$

Usually people interpret this as meaning that the effect is somewhere between -0.15 and -0.20 out of the total -0.247 wage gap. But what do we make out of the other term?

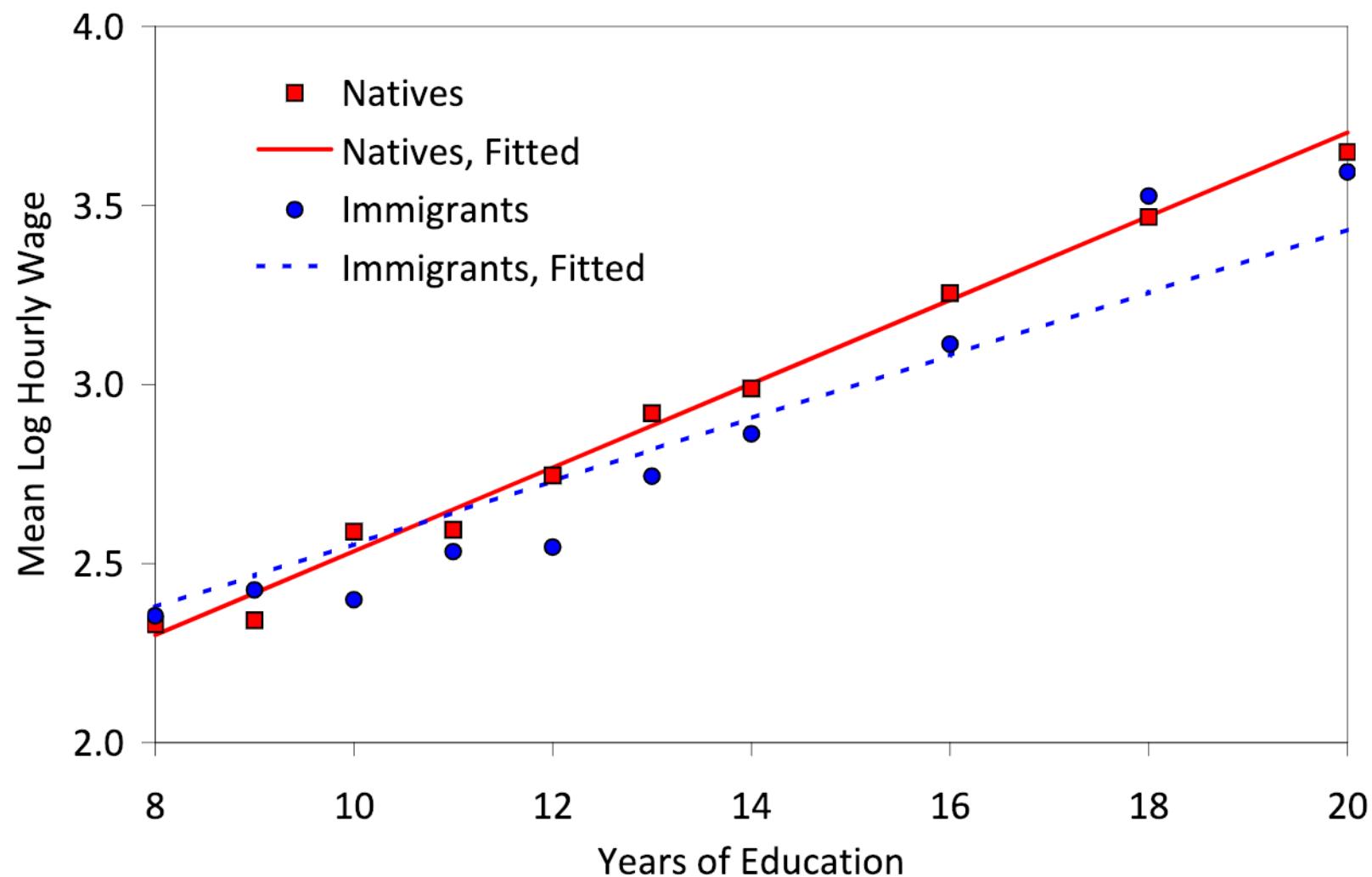
$$\begin{aligned}\bar{x}_2^b(\hat{\beta}_2^b - \hat{\beta}_2^a) &= -0.360 \\\bar{x}_2^a(\hat{\beta}_2^b - \hat{\beta}_2^a) &= -0.409\end{aligned}$$

In either case we are “over-explaining” the wage gap (by a lot). If you look back at the fitted models you can see what is happening

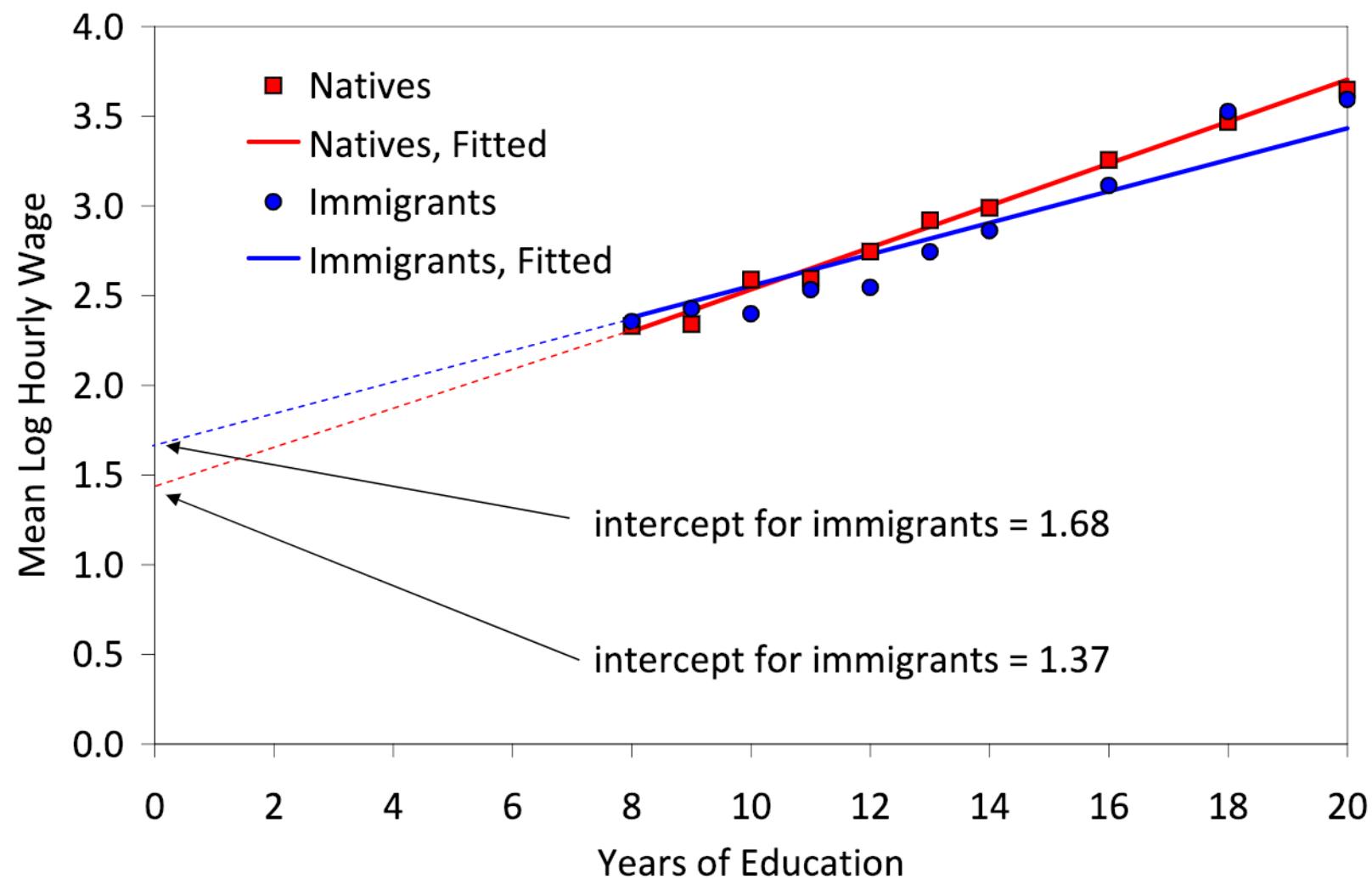
	Pooled Model: Fit to Natives and Immigrants		Model for Natives	Model for Immigrants
	(1)	(2)	(3)	(4)
Constant	3.013 (0.006)	1.546 (0.025)	1.365 (0.033)	1.676 (0.035)
Immigrant	-0.247 (0.013)	-0.072 (0.013)	--	--
Education (yrs)	--	0.104 (0.002)	0.117 (0.002)	0.088 (0.002)
MSE	0.757	0.695	0.689	0.707
Adj. R-sq	0.018	0.173	0.146	0.208
Sample Size	19,092	19,092	14,921	4,141

Notes: Fit to data for males age 30-45 in March 2012 CPS. Dependent variable is log average hourly wage. Mean and standard deviation are: for overall sample, 2.959 (0.764); for natives 3.013 (0.746); for immigrants 2.766 (0.795). Standard errors in parentheses.

Wages by Education -- Males Age 30-45



Wages by Education -- Males Age 30-45



The decomposition is multiplying the difference in estimated “returns to education” – which is $0.088 - 0.117 = -0.029$ by numbers like 12 or 14, which “explains” a quite large difference in wages. The estimated constants are offsetting this so the total explained difference is always exactly -0.247 .

We can see from this example that the part of the Oaxaca decomposition attributed to the difference in coefficients has to be interpreted carefully.

Let's probe this a little more. Suppose instead of measuring education in "years," we measured in "years of high school or more" i.e., we subtracted 8 from all measures of education.

$$\begin{aligned}
 \bar{y}^a &= \hat{\beta}_1^a + \hat{\beta}_2^a \bar{x}_2^a \\
 &= \hat{\beta}_1^a + \hat{\beta}_2^a (\bar{x}_2^a - 8) + 8\hat{\beta}_2^a \\
 &= (\hat{\beta}_1^a + 8\hat{\beta}_2^a) + \hat{\beta}_2^a (\bar{x}_2^a - 8)
 \end{aligned}$$

If we were to measure education as years of high school or more, we would get *exactly the same coefficient* on education, but the constant would be bigger (by exactly $8\hat{\beta}_2^a$). Likewise for group b :

$$\bar{y}^b = \hat{\beta}_1^b + \hat{\beta}_2^b \bar{x}_2^b = (\hat{\beta}_1^b + 8\hat{\beta}_2^b) + \hat{\beta}_2^b (\bar{x}_2^b - 8)$$

If we examined the “difference in x ’s” part of the Oaxaca decomposition, we would compare differences in renormalized education:

$$(\bar{x}_2^b - 8) - (\bar{x}_2^a - 8) = \bar{x}_2^b - \bar{x}_2^a$$

multiplying by $\hat{\beta}_2^a$ or $\hat{\beta}_2^b$ – so we would get the same answer as before. But for the “difference in coefficients” part of the decomposition, we would look at

$$(\hat{\beta}_2^b - \hat{\beta}_2^a) \times (\bar{x}_2^b - 8)$$

or

$$(\hat{\beta}_2^b - \hat{\beta}_2^a) \times (\bar{x}_2^a - 8)$$

Returning to our example:

$$\bar{x}_2^a = 14.09$$

$$\bar{x}_2^b = 12.41$$

$$\hat{\beta}_2^a = 0.117$$

$$\hat{\beta}_2^b = 0.088$$

So if we use the renormalized mean for immigrants we have:

$$(\bar{x}_2^b - 8)(\hat{\beta}_2^b - \hat{\beta}_2^a) = 4.41 \times -0.029 = -0.128$$

Whereas if we use renormalized mean for natives we have:

$$(\bar{x}_2^a - 8)(\hat{\beta}_2^b - \hat{\beta}_2^a) = 6.09 \times -0.029 = -0.177$$

Which still “over-explains” the immigrant-native wage gap!

Bottom line:

1. we can always use a pooled model to evaluate the effect of differences in mean x 's:

$$\bar{y}^b - \bar{y}^a = \sum_{j=2}^K (\bar{x}_j^b - \bar{x}_j^a) \hat{\beta}_j + \hat{\beta}_{K+1}$$

2. when the coefficients of the x variables are different for the two groups, we can evaluate two alternative terms:

$$\begin{aligned} & \sum_{j=2}^K (\bar{x}_j^b - \bar{x}_j^a) \hat{\beta}_j^a \\ & \sum_{j=2}^K (\bar{x}_j^b - \bar{x}_j^a) \hat{\beta}_j^b \end{aligned}$$

3. when the coefficients are different there is also a “difference in coefficients” component

The two estimates of the difference in coefficients component are:

$$\begin{aligned}\sum_{j=2}^K \bar{x}_j^b (\hat{\beta}_j^b - \hat{\beta}_j^a) \\ \sum_{j=2}^K \bar{x}_j^a (\hat{\beta}_j^b - \hat{\beta}_j^a)\end{aligned}$$

And can be evaluated. BUT – we have to be careful, because we can re-normalize the x variable and get different answers!!

Table 5: Models For Probability of Registering in STEM-Related University Program

	Dependent Variable = 1 if Register in STEM-related Program		
	(1)	(2)	(3)
Female Indicator ($\times 100$)	-5.0 (0.2)	-5.8 (0.2)	-1.7 (0.2)
Within-cohort Rank of Top 3 Grade 12 STEM courses	--	--	0.73 (0.01)
Within-cohort Rank of Top 6 Grade 12 Course	--	--	-0.52 (0.01)
Age, Year and High School Effects?	no	yes	yes
Gifted/special needs)?	no	yes	yes

Note: standard errors in parentheses. Table reports linear probability model coefficients for event of registering in STEM-related program. Sample is 170,288 STEM-ready students.

Models in columns 2-4 include dummies for age, graduating cohort, student's main language, foreign-born status, and high school. See Table 2 for sample information.

Mean Ranks in Top 3 Grade 12 STEM courses: females 0.50, males 0.51

Mean Ranks in Top 6 Grade 12 courses: females 0.61, males 0.55

Today:

more on decomposition methods

"reweighting" methods

Two building blocks we will need:

weighted regression

linear probability model

As usual we will posit a “population model”:

$$y_i = x'_i \beta^* + u_i.$$

Let's consider the special case where

$$x_i = \begin{pmatrix} D_{1i} \\ D_{2i} \\ \dots \\ D_{Gi} \end{pmatrix}$$

and $D_{gi} = 1$ if $i \in g$, where $\{g = 1 \dots G\}$ is a set of mutually exclusive “categories” – like education levels, or occupations. In this case

$$\frac{1}{N} \sum_i x_i x'_i = \begin{pmatrix} \frac{1}{N} \sum D_{1i}^2 & \frac{1}{N} \sum D_{1i} D_{2i} & \dots & \frac{1}{N} \sum D_{1i} D_{Gi} \\ \frac{1}{N} \sum D_{2i} D_{1i} & \frac{1}{N} \sum D_{2i}^2 & \dots & \\ \dots & & & \\ & & & \frac{1}{N} \sum D_{Gi}^2 \end{pmatrix}$$

So

$$\frac{1}{N} \sum_i x_i x'_i = \begin{pmatrix} \bar{p}_1 & 0 & \dots & 0 \\ 0 & \bar{p}_2 & \dots & 0 \\ \dots & & & \\ 0 & & & \bar{p}_G \end{pmatrix}$$

and

$$\frac{1}{N} \sum_i x_i y_i = \begin{pmatrix} \frac{1}{N} \sum D_{1i} y_i \\ \frac{1}{N} \sum D_{2i} y_i \\ \dots \\ \frac{1}{N} \sum D_{Gi} y_i \end{pmatrix} = \begin{pmatrix} \bar{p}_1 \bar{y}_1 \\ \bar{p}_2 \bar{y}_2 \\ \dots \\ \bar{p}_G \bar{y}_G \end{pmatrix}$$

So clearly

$$\hat{\beta} = \left(\frac{1}{N} \sum_i x_i x'_i \right)^{-1} \left(\frac{1}{N} \sum_i x_i y_i \right) = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \dots \\ \bar{y}_G \end{pmatrix}$$

Also:

$$\bar{x} = \frac{1}{N} \sum_i x_i = \begin{pmatrix} \frac{1}{N} \sum D_{1i} \\ \frac{1}{N} \sum D_{2i} \\ \dots \\ \frac{1}{N} \sum D_{Gi} \end{pmatrix} = \begin{pmatrix} \bar{p}_1 \\ \bar{p}_2 \\ \dots \\ \bar{p}_G \end{pmatrix}$$

So obviously

$$\bar{y} = \bar{x}' \hat{\beta} = \sum_g \bar{p}_g \bar{y}_g.$$

So the OLS regression is just a way to get the category means.

Now consider what happens with 2 groups, a and b . Let's denote:

$$\bar{y}^a = \frac{1}{N^a} \sum_{i \in a} y_i = \text{mean of outcome for group } a$$

$$\bar{y}_g^a = \frac{1}{N_g^a} \sum_{i \in g, a} y_i = \text{mean of outcome for group } a, \text{ category } g$$

$N^a = \#\text{obs}$ in group a , $N_g^a = \#\text{obs}$ in group a , category g

$\bar{y}_g^a = \frac{1}{N_g^a} \sum_{i \in g, a} y_i = \text{mean of outcome for group } a, \text{ category } g$

$\bar{p}_g^a = \frac{1}{N^a} \sum_{i \in a} D_{gi} = \text{fraction of group } a \text{ in category } g.$

Clearly:

$$\bar{y}^a = \sum_g \bar{p}_g^a \bar{y}_g^a = (\bar{x}^a)' \hat{\beta}^a$$

and for the other group b :

$$\bar{y}^b = \sum_g \bar{p}_g^b \bar{y}_g^b = (\bar{x}^b)' \hat{\beta}^b$$

So we can think about an Oaxaca decomposition of $\bar{y}^b - \bar{y}^a$.

An application: the “public-private pay gap”.

There is a lot of recent interest in the idea that government workers are “overpaid”. Is that true? If so, by how much?

To test this out, we’ll use our March 2012 sample. We’ll focus on female workers age 30-50 with 12 or more years of education. The data are summarized in the next table. Note that government workers are paid more, but also have more education. So our categories will be schooling categories.

Differences in Public and Private Sector Workers:

Women Age 30-50 in 2012 CPS

	Private Sector Workers (1)	Government Workers (2)	Public-Private Gap (2)
Mean Log Wage	2.762	2.976	0.214
Mean Education	14.320	15.460	1.140
Fraction < BA	0.616	0.398	-0.218
Fraction with BA	0.264	0.294	0.030
Fraction with MA	0.089	0.264	0.175
Fraction with PhD	0.031	0.044	0.013
Sample Size	17,354	4,315	

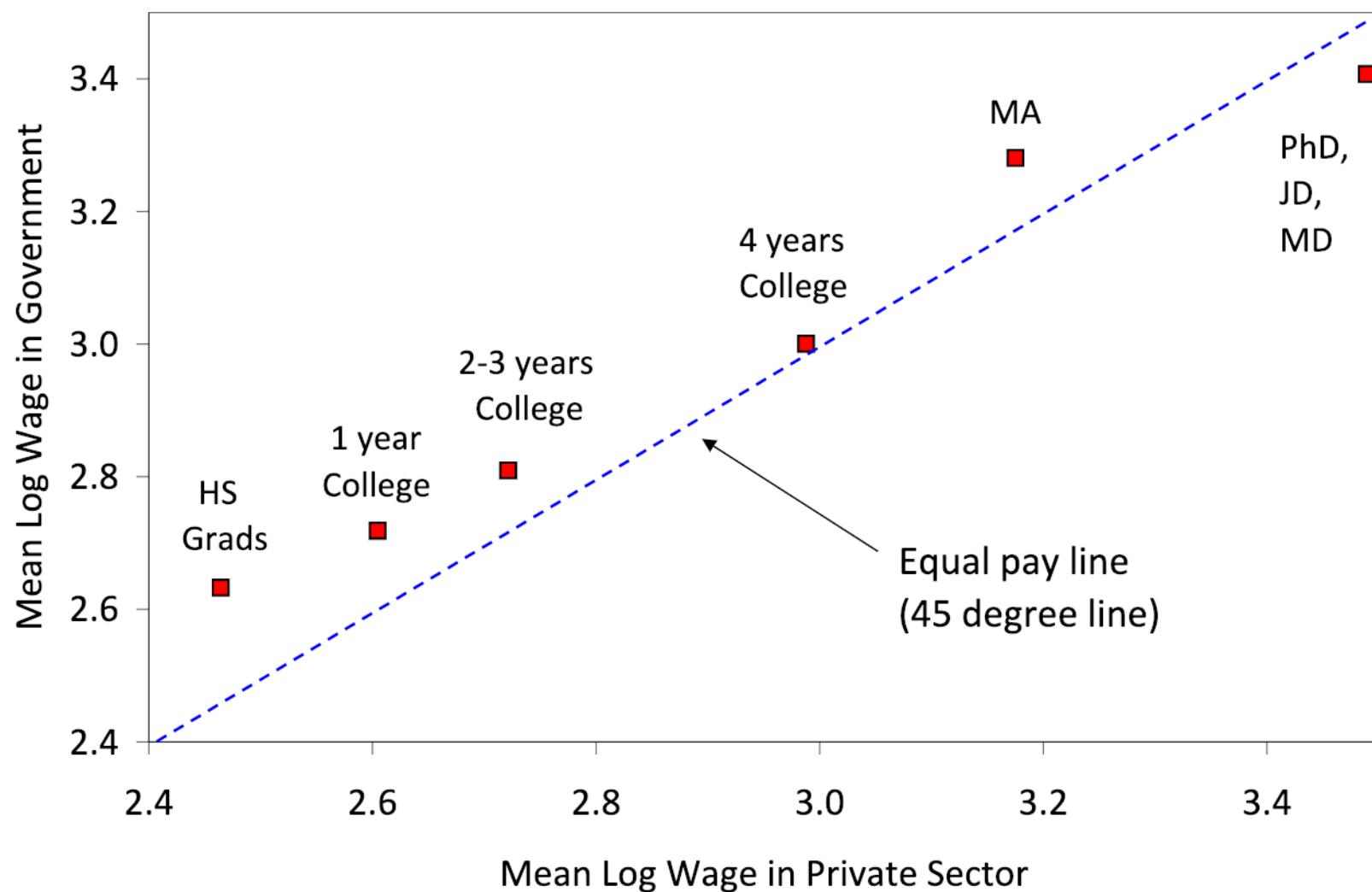
Note: Sample includes females age 30-50 in March 2012 CPS with 12 or more years of education and earnings in the last year.

Public-Private Wage Differentials for Women Age 30-50 in 2012 CPS

Education	Private Sector Workers		Government Workers		Govt Wage Premium
	Mean Log Wage	Fraction in Group	Mean Log Wage	Fraction in Group	
	(1)	(2)	(3)	(4)	(5)
12	2.464	0.275	2.633	0.153	0.168
13	2.605	0.193	2.719	0.140	0.114
14	2.721	0.148	2.810	0.107	0.088
16	2.988	0.264	3.001	0.293	0.013
18	3.175	0.089	3.281	0.264	0.105
20	3.489	0.031	3.407	0.044	-0.082
All	2.762	1.000	2.976	1.000	0.214

Note: Sample includes females age 30-50 in March 2012 CPS with 12 or more years of education and earnings in the last year. Sample size is 17,354 private sector workers and 4,315 government workers.

Wages in Government vs Private Sector: Women Age 30-50



When we look at the “distribution table” we can think of many ways to explore the nature of the overall pay gap. One way is to imagine a “counterfactual world” where group b (the gov’t workers) had the same distribution across categories as group a (the private sector workers) but within each category, the mean for group b remains unchanged. Under this counterfactual, the mean for group b would be:

$$\bar{y}_{counterf}^b = \sum_g \bar{p}_g^a \bar{y}_g^b = (\bar{x}^a)' \hat{\beta}^b$$

This “reweighted counterfactual” allows us to think about think about how much of the gap between \bar{y}^b and \bar{y}^a is due to the distribution across categories (rather than to differences within categories).

Under this counterfactual, the gap between b and a would be:

$$\begin{aligned}\bar{y}_{counterf}^b - \bar{y}^a &= \sum_g \bar{p}_g^a \bar{y}_g^b - \sum_g \bar{p}_g^a \bar{y}_g^a \\ &= \sum_g \bar{p}_g^a (\bar{y}_g^b - \bar{y}_g^a) \\ &= (\bar{x}^a)'(\hat{\beta}^b - \hat{\beta}^a)\end{aligned}$$

Notice that this is a “group a - weighted” average of the pay gaps within each category.

Public-Private Wage Differentials for Women Age 30-50 in 2012 CPS

Education	Private Sector Workers		Government Workers		Govt Wage Premium	Col 2	Col 4
	Mean Log Wage	Fraction in Group	Mean Log Wage	Fraction in Group			
	(1)	(2)	(3)	(4)	(5)	x Col 5	x Col 5
12	2.464	0.275	2.633	0.153	0.168	0.046	0.026
13	2.605	0.193	2.719	0.140	0.114	0.022	0.016
14	2.721	0.148	2.810	0.107	0.088	0.013	0.009
16	2.988	0.264	3.001	0.293	0.013	0.003	0.004
18	3.175	0.089	3.281	0.264	0.105	0.009	0.028
20	3.489	0.031	3.407	0.044	-0.082	-0.003	-0.004
All	2.762	1.000	2.976	1.000	0.214	0.092	0.079

Note: Sample includes females age 30-50 in March 2012 CPS with 12 or more years of education and earnings in the last year. Sample size is 17,354 private sector workers and 4,315 government workers.

In general we can write:

$$\bar{y}^b - \bar{y}^a = (\bar{y}_{counterf}^b - \bar{y}^a) + (\bar{y}^b - \bar{y}_{counterf}^b)$$

The second term is how much of the gap is “explained” by the deviation of the actual mean for b from the “reweighted as a counterfactual”:

$$\begin{aligned}\bar{y}^b - \bar{y}_{counterf}^b &= \sum_g \bar{p}_g^b \bar{y}_g^b - \sum_g \bar{p}_g^a \bar{y}_g^b \\ &= \sum_g (\bar{p}_g^b - \bar{p}_g^a) \bar{y}_g^b \\ &= (\bar{x}^a - \bar{x}^b)' \hat{\beta}^b\end{aligned}$$

This is a weighted average of the difference in shares across the categories, using the aveage pay of group b (the gov't workers) as weights.

We could also think of a different re-weighting counterfactual. Imagine that group a (the private sector workers) had the same distribution across categories as group b (the gov't workers) but within each category, the mean for group a remains unchanged. Under this counterfactual, the mean for group a would be:

$$\bar{y}_{counterf}^a = \sum_g p_g^b \bar{y}_g^a = (\bar{x}^b)' \hat{\beta}^a$$

We can write:

$$\begin{aligned}\bar{y}^b - \bar{y}^a &= (\bar{y}^b - \bar{y}_{counterf}^a) + (\bar{y}_{counterf}^a - \bar{y}^a) \\ &= (\bar{x}^b)' \hat{\beta}^b - (\bar{x}^b)' \hat{\beta}^a + (\bar{x}^b)' \hat{\beta}^a - (\bar{x}^a)' \hat{\beta}^a \\ &= (\bar{x}^b)' (\hat{\beta}^b - \hat{\beta}^a) + (\bar{x}^b - \bar{x}^a)' \hat{\beta}^a\end{aligned}$$

which is the alternative version of the Oaxaca decomp.

Public-Private Wage Differentials for Women Age 30-50 in 2012 CPS

Education	Private Sector Workers		Government Workers		Govt Wage Premium	Col 2	Col 4
	Mean Log Wage	Fraction in Group	Mean Log Wage	Fraction in Group			
	(1)	(2)	(3)	(4)	(5)	x Col 5	x Col 5
12	2.464	0.275	2.633	0.153	0.168	0.046	0.026
13	2.605	0.193	2.719	0.140	0.114	0.022	0.016
14	2.721	0.148	2.810	0.107	0.088	0.013	0.009
16	2.988	0.264	3.001	0.293	0.013	0.003	0.004
18	3.175	0.089	3.281	0.264	0.105	0.009	0.028
20	3.489	0.031	3.407	0.044	-0.082	-0.003	-0.004
All	2.762	1.000	2.976	1.000	0.214	0.092	0.079

Note: Sample includes females age 30-50 in March 2012 CPS with 12 or more years of education and earnings in the last year. Sample size is 17,354 private sector workers and 4,315 government workers.

A nice thing about the “reweighted counterfactual” gap is that we can construct it very quickly using a *weighted regression*.

Weights arise in applied problems where we have observations on y_i and x_i for a sample of individuals, and a set of “weights” $w_i \geq 0$ for each observation. The “weighted average” of y is

$$\bar{y}(w) = \frac{\sum_i w_i y_i}{\sum_i w_i}.$$

A lot of surveys have weights. For example, the March CPS sample has weights, which you can use to make the sample “nationally representative” of the US population. The sample has too many observations from small states and from certain cities to be representative without the weights.

Many samples (including the CPS) are drawn from “sampling strata”, which are pre-determined sub-groups (or sub-areas) with known total populations. Observations from different strata are sampled with different probabilities - in this case the weight is just $w_i = 1/p_{s(i)}$, where $s(i)$ is the strata that i is drawn from, and $p_{s(i)}$ is the sampling probability in that strata (i.e., the number of sample units to be selected from that strata, divided by the known population in that strata). In this sample design, each sample member from strata s “represents” $1/p_s$ people.

The weighted OLS regression coefficient is:

$$\hat{\beta}(w) = \left(\sum_i w_i x_i x_i' \right)^{-1} \sum_i w_i x_i y_i$$

which can be computed in standard regression packages.

Back to the counterfactuals...

We now show that we can construct the reweighted counterfactual mean for group b using a certain weighted average. For any observation from group b in category g , consider the weight:

$$w_g = \frac{N_g^a}{N_g^b}$$

Then

$$\bar{y}_{counterf}^b = \sum_g \bar{p}_g^a \bar{y}_g^b = \frac{\sum_{i \in b} w_g y_i}{\sum_{i \in b} w_g}$$

We'll prove this in two steps. First we look at the denominator. Then the numerator.

Denominator:

$$\begin{aligned}\sum_{i \in b} w_g &= \sum_{i \in b} \frac{N_g^a}{N_g^b} \\&= \sum_g \sum_{i \in g, b} \frac{N_g^a}{N_g^b} \\&= \sum_g \frac{N_g^a}{N_g^b} \sum_{i \in g, b} 1 \\&= \sum_g N_g^a = N^a\end{aligned}$$

Numerator:

$$\begin{aligned}\sum_{i \in b} w_g y_i &= \sum_{i \in b} \left(\frac{N_g^a}{N_g^b} \right) y_i \\&= \sum_g \left(\frac{N_g^a}{N_g^b} \right) \sum_{i \in g, b} y_i \\&= \sum_g N_g^a \bar{y}_g^b \\&= N^a \sum_g \bar{p}_g^a \bar{y}_g^b\end{aligned}$$

So dividing by the denominator gets us

$$\frac{\sum_{i \in b} w_g y_i}{\sum_{i \in b} w_g} = \sum_g \bar{p}_g^a \bar{y}_g^b$$

So we can get the counterfactual mean for group b using a weighted mean. Note the weights “rebalance” the observations across categories. So if there are lots of a 's in category g , and not many b 's, then we give more weight to the observations from the b group that we see in this category.

We can also easily construct the difference $\bar{y}_{counterf}^b - \bar{y}^a$. The first term is the weighted average of outcomes for group b . The second term is mean for group a . So we can pool the two groups and fit a weighted regression on a constant and a dummy for group b . The weight for each observation in group b , category g is $w_g = \frac{N_g^a}{N_g^b}$. The weight for all observations in group a is 1. The weighted OLS coefficient on the dummy for group b will equal the difference in the weighted means.

That then leaves us with the question: is there a “fast way” to get the weights? The answer is yes. We will use a very important “trick” invented by John DiNardo, Thomas Lemieux, and Nicole Fortin – known by applied economists as the “DFL” method.

The trick is this: if you considered a “crazy” regression on the combined sample of a and b :

$$m_i = x'_i \theta + \eta$$

where the dependent variable is $m_i = 1[i \in a]$ and $x'_i = (D_{1i}, D_{2i} \dots D_{Gi})$ what would you get for estimates of θ ? From the first part of the lecture, we know that the OLS estimate of θ_g will be:

$$\hat{\theta}_g = \frac{1}{N_g} \sum_{i \in g} m_i = \frac{N_g^a}{N_g^a + N_g^b}$$

So if we fit the “crazy” model predicting membership in group a , and took the predictions from this model, we’ll have:

$$\hat{m}_i = x_i' \hat{\theta} = \frac{N_{g(i)}^a}{N_{g(i)}^a + N_{g(i)}^b}$$

where $g(i)$ is the category that observation i fall into. Now form the weight

$$w_{g(i)} = \frac{\hat{m}_i}{1 - \hat{m}_i} = \frac{N_{g(i)}^a}{N_{g(i)}^b}$$

This is the weight we want to form our counterfactual mean (and run the counterfactual gap regression).

Since the object we are trying to predict in the “crazy regression” is the probability of membership in category g , most times people use a logit model rather than a linear probability model.

How does this work in our example?

Mean wage of private sector workers = 2.7621

Mean wage of government workers = 2.9764

Counterf. mean of gov't workers = 2.85364 (same ed as pri-vates)

Unadjusted wage gap = 0.2143 (0.012)

Counterfactual wage gap = 0.09152 (0.009)

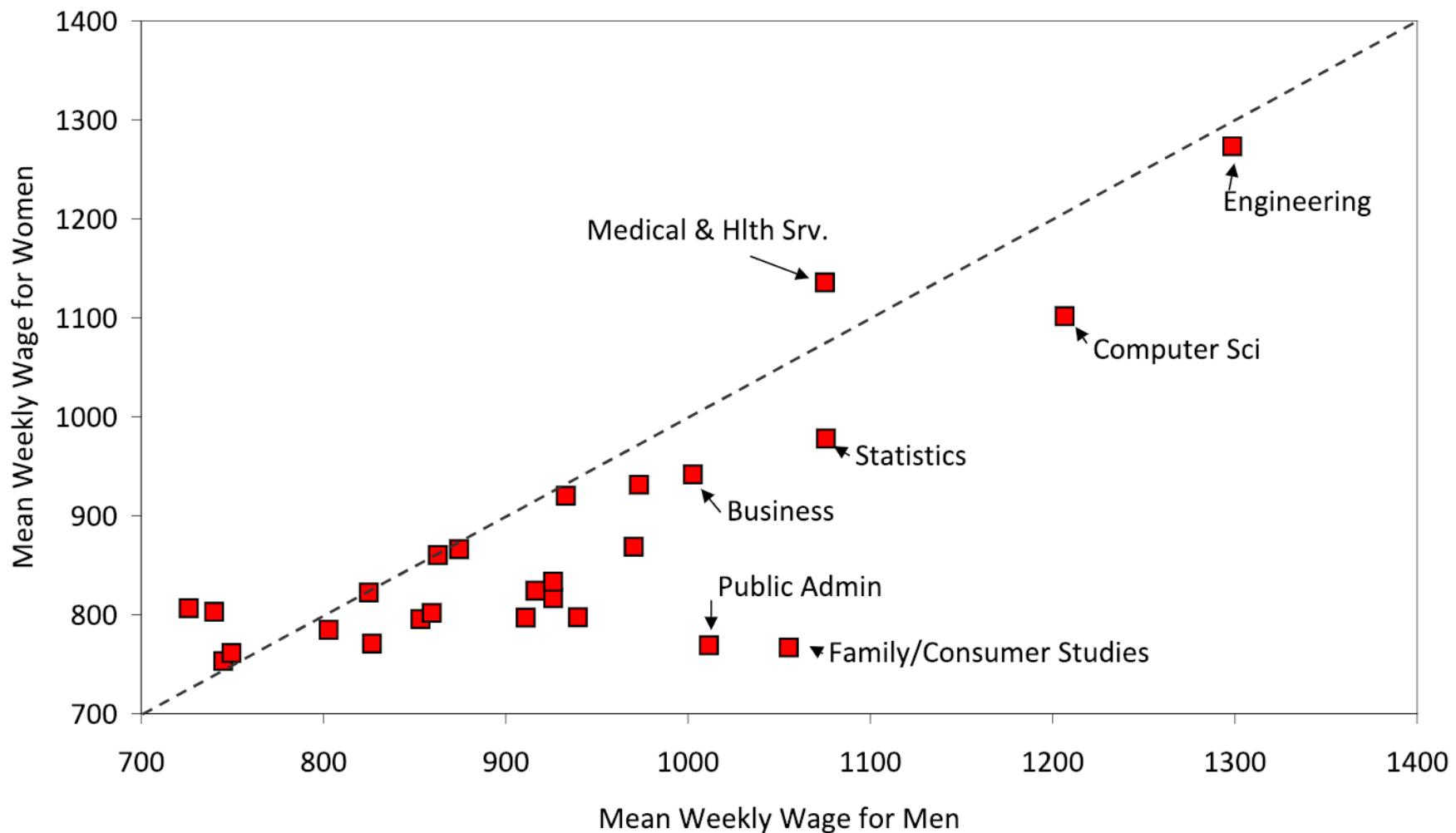
We can also do the alternative counterfactual mean for the pri-vates, giving them the same education distribution as gov't work-ers.

Contribution of Field of Study to Gender Wage Gap: Workers Age 24-28 With BA Degree

Field of BA:	Male Workers		Female Workers		Male Wage	Male Shrs	Female Shrs
	Mean	Pct.	Mean	Pct.	Premium	× Premium	× Premium
Agriculture	6.61	1.8%	6.62	0.9%	-0.011	0.000	0.000
Environ Sci	6.85	0.8%	6.68	0.4%	0.164	0.001	0.001
Architecture	6.88	0.9%	6.84	0.5%	0.044	0.000	0.000
Ethnic Studies	6.83	0.2%	6.71	0.5%	0.126	0.000	0.001
Communications/journalism	6.76	5.5%	6.76	9.1%	0.003	0.000	0.000
Computer Science	7.10	7.1%	7.00	1.3%	0.091	0.007	0.001
Education	6.62	3.8%	6.64	11.2%	-0.016	-0.001	-0.002
Engineering (General)	7.17	12.4%	7.15	2.4%	0.020	0.002	0.000
Foreign languages	6.59	0.5%	6.69	1.0%	-0.105	-0.001	-0.001
Family/Consumer Science	6.92	0.2%	6.65	1.5%	0.274	0.001	0.004
Literature	6.69	1.7%	6.67	2.9%	0.023	0.000	0.001
Humanities	6.61	0.7%	6.69	1.2%	-0.082	-0.001	-0.001
Biology	6.77	2.7%	6.76	3.8%	0.009	0.000	0.000
Statistics	6.98	1.2%	6.89	0.8%	0.095	0.001	0.001
Interdisciplinary	6.82	0.5%	6.71	1.1%	0.106	0.001	0.001
Physical Fitness/Leisure	6.72	1.8%	6.71	1.3%	0.003	0.000	0.000
Physical Sciences	6.84	1.8%	6.82	1.4%	0.014	0.000	0.000
Psychology	6.75	2.2%	6.68	5.8%	0.070	0.002	0.004
Criminal justice/fire protection	6.81	3.3%	6.68	2.3%	0.134	0.004	0.003
Public Admin/Policy	6.96	0.3%	6.64	1.5%	0.319	0.001	0.005
Social Sciences	6.88	8.5%	6.77	7.0%	0.111	0.009	0.008
Arts	6.72	3.9%	6.65	5.9%	0.070	0.003	0.004
Medical and Health Services	6.98	1.5%	7.04	9.7%	-0.055	-0.001	-0.005
Business	6.91	28.5%	6.85	23.4%	0.063	0.018	0.015
History	6.76	2.2%	6.69	1.3%	0.069	0.002	0.001
Miscellaneous	6.83	5.6%	6.73	1.6%	0.105	0.006	0.002
Row Average/Total	6.89	100.0%	6.78	100.0%	0.108	0.056	0.043

Note: Source: Am. Comm. Survey. Fields with less than 50 female workers are merged into the "Miscellaneous" field.

Weekly Wages by Field of Study: Females vs. Males
Age 24-28 With BA Degree, 2010 Am. Comm. Survey



These are the codes in the 2010 American Community Survey for Field of Degree (first entry).

FOD1P

Recoded field of degree - first entry

bbbb .N/A (less than bachelor's degree)
1100 .GENERAL AGRICULTURE
1101 .AGRICULTURE PRODUCTION AND MANAGEMENT
1102 .AGRICULTURAL ECONOMICS
1103 .ANIMAL SCIENCES
1104 .FOOD SCIENCE
1105 .PLANT SCIENCE AND AGRONOMY
1106 .SOIL SCIENCE52
1199 .MISCELLANEOUS AGRICULTURE
1301 .ENVIRONMENTAL SCIENCE
1302 .FORESTRY
1303 .NATURAL RESOURCES MANAGEMENT
1401 .ARCHITECTURE
1501 .AREA ETHNIC AND CIVILIZATION STUDIES
1901 .COMMUNICATIONS
1902 .JOURNALISM
1903 .MASS MEDIA
1904 .ADVERTISING AND PUBLIC RELATIONS
2001 .COMMUNICATION TECHNOLOGIES
2100 .COMPUTER AND INFORMATION SYSTEMS
2101 .COMPUTER PROGRAMMING AND DATA PROCESSING
2102 .COMPUTER SCIENCE
2105 .INFORMATION SCIENCES
2106 .COMPUTER ADMINISTRATION MANAGEMENT AND SECURITY
2107 .COMPUTER NETWORKING AND TELECOMMUNICATIONS
2201 .COSMETOLOGY SERVICES AND CULINARY ARTS
2300 .GENERAL EDUCATION
2301 .EDUCATIONAL ADMINISTRATION AND SUPERVISION
2303 .SCHOOL STUDENT COUNSELING
2304 .ELEMENTARY EDUCATION
2305 .MATHEMATICS TEACHER EDUCATION
2306 .PHYSICAL AND HEALTH EDUCATION TEACHING
2307 .EARLY CHILDHOOD EDUCATION
2308 .SCIENCE AND COMPUTER TEACHER EDUCATION
2309 .SECONDARY TEACHER EDUCATION
2310 .SPECIAL NEEDS EDUCATION
2311 .SOCIAL SCIENCE OR HISTORY TEACHER EDUCATION
2312 .TEACHER EDUCATION: MULTIPLE LEVELS
2313 .LANGUAGE AND DRAMA EDUCATION
2314 .ART AND MUSIC EDUCATION
2399 .MISCELLANEOUS EDUCATION
2400 .GENERAL ENGINEERING

2410 .ENVIRONMENTAL ENGINEERING
2411 .GEOLOGICAL AND GEOPHYSICAL ENGINEERING
2412 .INDUSTRIAL AND MANUFACTURING ENGINEERING
2413 .MATERIALS ENGINEERING AND MATERIALS SCIENCE
2414 .MECHANICAL ENGINEERING
2415 .METALLURGICAL ENGINEERING
2416 .MINING AND MINERAL ENGINEERING
2417 .NAVAL ARCHITECTURE AND MARINE ENGINEERING
2418 .NUCLEAR ENGINEERING
2419 .PETROLEUM ENGINEERING
2499 .MISCELLANEOUS ENGINEERING
2413 .MATERIALS ENGINEERING AND MATERIALS SCIENCE
2414 .MECHANICAL ENGINEERING
2415 .METALLURGICAL ENGINEERING
2416 .MINING AND MINERAL ENGINEERING
2417 .NAVAL ARCHITECTURE AND MARINE ENGINEERING
2418 .NUCLEAR ENGINEERING
2419 .PETROLEUM ENGINEERING
2499 .MISCELLANEOUS ENGINEERING
2500 .ENGINEERING TECHNOLOGIES
2501 .ENGINEERING AND INDUSTRIAL MANAGEMENT
2502 .ELECTRICAL ENGINEERING TECHNOLOGY
2503 .INDUSTRIAL PRODUCTION TECHNOLOGIES
2504 .MECHANICAL ENGINEERING RELATED TECHNOLOGIES
2599 .MISCELLANEOUS ENGINEERING TECHNOLOGIES
2601 .LINGUISTICS AND COMPARATIVE LANGUAGE AND LITERATURE
2602 .FRENCH GERMAN LATIN AND OTHER COMMON FOREIGN LANGUAGE STUDIES
2603 .OTHER FOREIGN LANGUAGES
2901 .FAMILY AND CONSUMER SCIENCES
3201 .COURT REPORTING
3202 .PRE-LAW AND LEGAL STUDIES
3301 .ENGLISH LANGUAGE AND LITERATURE
3302 .COMPOSITION AND RHETORIC
3401 .LIBERAL ARTS
3402 .HUMANITIES
3501 .LIBRARY SCIENCE
3600 .BIOLOGY
3601 .BIOCHEMICAL SCIENCES
3602 .BOTANY
3603 .MOLECULAR BIOLOGY
3604 .ECOLOGY
3605 .GENETICS
3606 .MICROBIOLOGY
3607 .PHARMACOLOGY
3608 .PHYSIOLOGY
3609 .ZOOLOGY
3611 .NEUROSCIENCE
3699 .MISCELLANEOUS BIOLOGY

4007 .INTERDISCIPLINARY SOCIAL SCIENCES
4101 .PHYSICAL FITNESS PARKS RECREATION AND LEISURE
4801 .PHILOSOPHY AND RELIGIOUS STUDIES
4901 .THEOLOGY AND RELIGIOUS VOCATIONS
5000 .PHYSICAL SCIENCES
5001 .ASTRONOMY AND ASTROPHYSICS
5002 .ATMOSPHERIC SCIENCES AND METEOROLOGY
5003 .CHEMISTRY
5004 .GEOLOGY AND EARTH SCIENCE
5005 .GEOSCIENCES
5006 .OCEANOGRAPHY
5007 .PHYSICS
5008 .MATERIALS SCIENCE
5098 .MULTI-DISCIPLINARY OR GENERAL SCIENCE
5102 .NUCLEAR, INDUSTRIAL RADILOGY, AND BIOLOGICAL TECHNOLOGIES
5200 .PSYCHOLOGY
5201 .EDUCATIONAL PSYCHOLOGY
5202 .CLINICAL PSYCHOLOGY
5203 .COUNSELING PSYCHOLOGY
5205 .INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY
5206 .SOCIAL PSYCHOLOGY
5299 .MISCELLANEOUS PSYCHOLOGY
5301 .CRIMINAL JUSTICE AND FIRE PROTECTION
5401 .PUBLIC ADMINISTRATION
5402 .PUBLIC POLICY
5403 .HUMAN SERVICES AND COMMUNITY ORGANIZATION
5404 .SOCIAL WORK
5500 .GENERAL SOCIAL SCIENCES
5501 .ECONOMICS
5502 .ANTHROPOLOGY AND ARCHEOLOGY
5503 .CRIMINOLOGY
5504 .GEOGRAPHY
5505 .INTERNATIONAL RELATIONS
5506 .POLITICAL SCIENCE AND GOVERNMENT
5507 .SOCIOLOGY
5599 .MISCELLANEOUS SOCIAL SCIENCES
5601 .CONSTRUCTION SERVICES
5701 .ELECTRICAL, MECHANICAL, AND PRECISION TECHNOLOGIES AND PRODUCTION
5901 .TRANSPORTATION SCIENCES AND TECHNOLOGIES
6000 .FINE ARTS
6001 .DRAMA AND THEATER ARTS
6002 .MUSIC
6003 .VISUAL AND PERFORMING ARTS
6004 .COMMERCIAL ART AND GRAPHIC DESIGN
6005 .FILM VIDEO AND PHOTOGRAPHIC ARTS
6006 .ART HISTORY AND CRITICISM
6007 .STUDIO ARTS
6099 .MISCELLANEOUS FINE ARTS

6110 .COMMUNITY AND PUBLIC HEALTH
6199 .MISCELLANEOUS HEALTH MEDICAL PROFESSIONS
6200 .GENERAL BUSINESS
6201 .ACCOUNTING
6202 .ACTUARIAL SCIENCE
6203 .BUSINESS MANAGEMENT AND ADMINISTRATION
6204 .OPERATIONS LOGISTICS AND E-COMMERCE
6205 .BUSINESS ECONOMICS
6206 .MARKETING AND MARKETING RESEARCH
6207 .FINANCE
6209 .HUMAN RESOURCES AND PERSONNEL MANAGEMENT
6210 .INTERNATIONAL BUSINESS
6211 .HOSPITALITY MANAGEMENT
6212 .MANAGEMENT INFORMATION SYSTEMS AND STATISTICS
6299 .MISCELLANEOUS BUSINESS & MEDICAL ADMINISTRATION
6402 .HISTORY
6403 .UNITED STATES HISTORY

Lecture 10:

reweighting methods and the “propensity score”

New building block we will need:

logistic regression model (or “logit”)

Recap - re-weighted counterfactuals

Two groups, a and b . Group a is the “reference group”

\bar{y}^a = mean of outcome for group a

\bar{y}_g^a = mean of outcome for group a , category g

N^a = #obs in group a , N_g^a = #obs in group a , category g

$\bar{p}_g^a = \frac{N_g^a}{N^a}$ = fraction of group a in category g .

We know:

$$\bar{y}^a = \sum_g \bar{p}_g^a \bar{y}_g^a, \quad \bar{y}^b = \sum_g \bar{p}_g^b \bar{y}_g^b$$

We considered a regression-approach: we fit OLS models to each group, with group dummies (and no constant). So $x'_i = (D_{1i}, D_{2i}, \dots, D_{Gi})$.

Clearly $(\bar{x}^a)' = (\bar{p}_1^a, \dots, \bar{p}_G^a)$; and we showed that $\hat{\beta}^a = (\bar{y}_1^a, \dots, \bar{y}_G^a)$.

So we can write:

$$\bar{y}^a = \sum_g \bar{p}_g^a \bar{y}_g^a = (\bar{x}^a)' \hat{\beta}^a$$

We also defined the counterfactual for group b if this group had the same distribution across categories as group a :

$$\bar{y}_{counterf}^b = \sum_g \bar{p}_g^a \bar{y}_g^b = (\bar{x}^a)' \hat{\beta}^b$$

Next we showed that we can write the counterfactual mean as a “reweighted mean”, where the weight for a person i who is category g is $w_i = w_{g(i)} = N_g^a/N_g^b$:

$$\bar{y}_{counterf}^b = \sum_g \bar{p}_g^a \bar{y}_g^b = \frac{\sum_{i \in b} w_{g(i)} y_i}{\sum_{i \in b} w_{g(i)}}$$

Finally, we considered the “crazy” regression for the combined sample:

$$m_i = x_i' \theta + \eta$$

where $m_i = 1[i \in a]$ and $x_i' = (D_{1i}, D_{2i} \dots D_{Gi})$. We showed that

$$w_{g(i)} = \frac{\hat{m}_i}{1 - \hat{m}_i} = \frac{N_{g(i)}^a}{N_{g(i)}^b}$$

When you stack up the observations from 2 groups, and fit a model to predict who belongs to group a given the covariates – i.e., $P(i \in a|x_i)$ – the predicted probability is called the “propensity score” or “p-score” for membership in a . So the “crazy regression” is just a model for the propensity score.

Summary of reweighted counterfactual method:

1. combine groups and fit a model for $p_i = p(m_i = 1|x_i)$
2. form the weight $w_i = \hat{p}_i/(1 - \hat{p}_i)$
3. form $\bar{y}_{counterf}^b = \frac{\sum_{i \in b} w_i y_i}{\sum_{i \in b} w_i}$
4. Interpret $\bar{y}_{counterf}^b - \bar{y}^a$ as the difference between group b and group a (the reference group) after “adjusting” group b to have the same x 's as a .

This method can be used even the x 's are not really discrete categories. Suppose for example that we have information on education and age. We would not necessarily want to divide the data into “buckets” with only one year of age and each possible value of education, because some of the buckets will be empty. Instead, we might want to “smooth” across categories in some way to perform the reweighting.

It turns out that using a *flexible estimated propensity score* is the right approach. So we might fit a model for the propensity score that has dummies for each education, linear and quadratic terms in age, and interactions of the linear and quadratic age terms with the education dummies.

We then fit the propensity score model, and we can use the $\hat{p}'s$ to reweight group b . The interpretation is that we are approximating what we could do if we had a very large data set and used very small buckets.

There is a potential problem. If we fit a simple linear model for m_i :

$$m_i = x'_i \theta + \eta$$

where one of the elements of x_i is age-squared – we might end up with a predicted value bigger than 1. In that case the weight $\hat{p}_i/(1-\hat{p}_i)$ will be negative. The solution is to move from a “linear probability” model to a logit model.

A linear regression model for a 0/1 variable like m_i is known as a linear probability model. Why? Suppose we assume that

$$E[m_i|x_i] = P(m_i = 1|x_i) = x'_i \theta$$

Then we know that we should estimate a linear regression and obtain an estimate $\hat{\theta}$ of θ . But when x_i has a large range, a model like this is less attractive. Instead, it would make sense to assume that

$$E[m_i|x_i] = P(m_i = 1|x_i) = G(x'_i \theta)$$

for some function G that maps from $(-\infty, \infty)$ to $(0, 1)$. But this is exactly what distribution functions do: so people normally pick a d.f. they like. A very convenient d.f. is the logistic:

$$G(z) = \frac{e^z}{1 + e^z} = \Lambda(z).$$

The Logistic Model

we can motivate the logistic as follows. Suppose that there is a “latent index”

$$m_i^* = x_i' \theta + \epsilon_i$$

where ϵ_i is a r.v. with a logistic distribution, and suppose that

$$m_i = 1 \iff m_i^* > 0$$

In this setting m_i^* represents the “tendency” of person i to have $m_i = 1$. This tendency depends on x_i and on the random factor ϵ_i . What is the probability of $m_i = 1$?

$$\begin{aligned} P(m_i = 1 | x_i) &= P(m_i^* > 0 | x_i) \\ &= P(x_i' \theta + \epsilon_i > 0) \\ &= P(\epsilon_i > -x_i' \theta) = 1 - G(-x_i' \theta) \\ &= G(x_i' \theta) = \frac{e^{x_i' \theta}}{1 + e^{x_i' \theta}}. \end{aligned}$$

Suppose that we posit $P(m_i = 1|x_i) = E[m_i|x_i] = G(x'_i\theta)$. We could choose $\hat{\theta}$ to:

$$\min_{\theta} \frac{1}{N} \sum_i (m_i - G(x'_i\theta))^2$$

But it is more conventional to choose $\hat{\theta}$ to maximize:

$$\frac{1}{N} \sum_i \log P(m_i|x_i, \theta)$$

where $P(m_i|x_i, \theta)$ is the probability of observing the actual value of m_i that occurs for the i^{th} observation, given x_i and the assumed model $G(x'_i\theta)$. This procedure is called “maximum likelihood estimation” or MLE, invented by R.A. Fisher in the 1930s. As you may recall from previous classes, MLE has many good properties.

Let's look at the FOC for MLE. Notice that

$$P(m_i|x_i, \theta) = G(x_i'\theta)^{m_i}(1 - G(x_i'\theta))^{1-m_i}$$

So taking logs:

$$\log P(m_i|x_i, \theta) = m_i \log(G(x_i'\theta)) + (1 - m_i) \log(1 - G(x_i'\theta))$$

and our objective is to maximize

$$L = \frac{1}{N} \sum_i m_i \log(G(x_i'\theta)) + (1 - m_i) \log(1 - G(x_i'\theta))$$

$$\begin{aligned}
L &= \frac{1}{N} \sum_i m_i \log(G(x'_i \theta)) + (1 - m_i) \log(1 - G(x'_i \theta)) \\
\Rightarrow \frac{\partial L}{\partial \theta_k} &= \frac{1}{N} \sum_i m_i \frac{g(x'_i \theta)}{G(x'_i \theta)} x_{ki} - (1 - m_i) \frac{g(x'_i \theta)}{1 - G(x'_i \theta)} x_{ki} \\
\Rightarrow \frac{\partial L}{\partial \theta} &= \frac{1}{N} \sum_i m_i \frac{g(x'_i \theta)}{G(x'_i \theta)} x_i - (1 - m_i) \frac{g(x'_i \theta)}{1 - G(x'_i \theta)} x_i
\end{aligned}$$

where $g(z) = G'(z)$ is the density function (since G is a d.f.).

Now let's simplify:

$$\begin{aligned} \left(m_i \frac{g(.)}{G(.)} - (1 - m_i) \frac{g(.)}{1 - G(.)} \right) x_i &= \frac{m_i g(.) (1 - G(.)) - (1 - m_i) g(.) G(.)}{G(.) (1 - G(.))} x_i \\ &= \frac{g(.) (m_i - G(.))}{G(.) (1 - G(.))} x_i \end{aligned}$$

And using the logistic properties

$$\begin{aligned} G(z) &= \frac{e^z}{1 + e^z} \\ \Rightarrow g(z) &= \frac{e^z}{(1 + e^z)^2} \\ &= G(z)(1 - G(z)) \end{aligned}$$

So:

$$\begin{aligned}
 \frac{\partial L}{\partial \theta} &= \frac{1}{N} \sum_i m_i \frac{g(x'_i \theta)}{G(x'_i \theta)} x_i - (1 - m_i) \frac{g(x'_i \theta)}{1 - G(x'_i \theta)} x_i \\
 &= \frac{1}{N} \sum_i \frac{g(x'_i \theta)(m_i - G(x'_i \theta))}{G(x'_i \theta)(1 - G(x'_i \theta))} x_i \\
 &= \frac{1}{N} \sum_i x_i (m_i - G(x'_i \theta))
 \end{aligned}$$

Notice that when we set the FOC to 0, we get

$$\frac{1}{N} \sum_i x_i (m_i - G(x'_i \hat{\theta})) = 0$$

which is K equations in the K parameters of $\hat{\theta}$. Also notice that $G(x'_i \hat{\theta})$ is the predicted probability that $m_i = 1$. So the parameters are selected so that the “prediction error” $m_i - \hat{p}_i$ is orthogonal to x_i – just like the FOC for OLS!

Standard programs like R or Stata have a “built in” procedure for finding the parameter vector that satisfies the FOC, using an iterative search method. They also report the estimated sampling errors (or standard errors) of the estimated parameters.

One of the main reasons why *people love logits* is the FOC lead to the nice orthogonality condition:

$$\frac{1}{N} \sum_i x_i(m_i - \hat{p}_i) = 0$$

If the model contains a constant (first row of x_i), this implies:

$$\frac{1}{N} \sum_i (m_i - \hat{p}_i) = 0 \Rightarrow \frac{1}{N} \sum_i m_i = \frac{1}{N} \sum_i \hat{p}_i$$

or the mean predicted probability is equal to the sample average of m_i .

Also, if the model contains only categorical dummies, the model will fit the actual averages of m_i within each category perfectly.

How should we interpret the estimated parameters $\hat{\theta}$? Since

$$\begin{aligned}
 P(m_i = 1|x_i) &= \frac{\exp(x_i\theta)}{1 + \exp(x_i\theta)} \\
 \frac{\partial P}{\partial x_{ki}} &= \frac{\theta_k \exp(x_i\theta)(1 + \exp(x_i\theta)) - \theta_k (\exp(x_i\theta))^2}{(1 + \exp(x_i\theta))^2} \\
 &= \frac{\theta_k \exp(x_i\theta)}{(1 + \exp(x_i\theta))^2} \\
 &= p_i(1 - p_i)\theta_k
 \end{aligned}$$

So when p_i is close to 0 or close to 1, the effect of an incremental change in x_{ki} is attenuated.

Summary of reweighted counterfactual method:

1. fit a logit model for $p_i = p(m_i = 1|x_i)$, using dummies for each value of x if x is a categorical variable, or using “flexible” function of x in the more general case – e.g., a fully interacted polynomial of some order.
2. form the weight $w_i = \hat{p}_i/(1 - \hat{p}_i)$
3. form $\bar{y}_{counterf}^b = \frac{\sum_{i \in b} w_i y_i}{\sum_{i \in b} w_i}$
4. Interpret $\bar{y}_{counterf}^b - \bar{y}^a$ as the difference between group b and group a (the reference group) after “adjusting” group b to have the same x 's as a .

To check that the weights are working, compute

$$\bar{x}_{counterf}^b = \frac{\sum_{i \in b} w_i x_i}{\sum_{i \in b} w_i}$$

This should be exactly equal to \bar{x}^a for all categorical dummy variables in the model, and very close for the other x 's.

It is also often informative to compare $\bar{y}_{counterf}^b$ to the simple “OLS regression counterfactual”

$$\bar{y}_{OLS-counter}^b = (\bar{x}^b)' \hat{\beta}^a$$

In the case where the model has only category dummies the two will be exactly the same. Otherwise, they will differ slightly: the OLS counterfactual depends on the specification of the regression model. The reweighted counterfactual will vary a bit depend on the specification of the *p-score*.

Finally, it is also interesting to note that with a reweighting approach, you can estimate not just the *counterfactual mean*, but also other counterfactual statistics such as the median or standard deviation. For example, the counterfactual standard deviation is:

$$\sigma_{counterf}^b = \frac{1}{N - 1} \frac{\sum_{i \in b} w_i (y_i - \bar{y}_{counterf}^b)^2}{\sum_{i \in b} w_i}$$

Economists since DFL have used this approach to estimate the effect of demographic changes on wage inequality

Brief review of key concepts

1. Law of iterated expectations: $E[y] = E[E[y|x]]$

\Leftrightarrow the mean is the weighted average of the mean for subgroups.

2. Population regression function: $\operatorname{argmin} E[(y_i - x'_i \beta)^2]$

\Leftrightarrow f.o.c. $E[x_i(y_i - x'_i \beta)] = 0$

$\Leftrightarrow y_i = x'_i \beta + u_i$ where $E[x_i u_i] = 0$ \leftarrow defining property of PRF

3. OLS regression: sample version of PRF

3. If $E[y_i|x_i] = x'_i\beta$ then PRF $y_i = x'_i\beta + u_i$ has *the same* β
 why? $\beta = E[x_i x'_i]^{-1} E[x_i y_i]$. Write $y_i = E[y_i|x_i] + \epsilon_i\dots$
4. Using the FOC: (i) if x_i has a constant. (ii) if x_i has a dummy
5. F-W theorem: PRF: $y_i = x_{1i}\beta_1 + x_{2i}\beta_2\dots + x_{Ji}\beta_J + u_i$

$$\beta_j = E[\xi_i^2]^{-1} E[\xi_i y_i]$$

where ξ_i is the residual from a PRF of x_{ji} on all the other x' s:

$$x_{ji} = x'_{(\sim j)i}\pi + \xi_i.$$

6. Implications of FW: in baseline case: $y_i = \beta_0 + x_i\beta_1 + u_i$

$$\xi_1 = x_i - E[x_i] \Rightarrow \beta_1 = E[(x_i - E[x_i])^2]^{-1} E[(x_i - E[x_i])y_i]$$

7. FW for OLS (same properties)

8. OVT: break out $x'_i = (x'_{A_i}, z_i)$. The PRF is:

$$y_i = x'_{A_i}\beta_A + z_i\beta_z + u_i$$

If we omit z , PRF is $y_i = x'_{A_i}\delta_A + v_i$, with

$$\delta_A = \beta_A + \beta_z\pi_A$$

$$z_i = x'_{A_i}\pi_a + e_i$$

the second is called the “auxilliary regression”

9. definition of R-sq

10. standard errors

11. Oaxaca decomp

12. weighted OLS, logit, reweighting

Lecture 11: Intro to Part 2 of the course: “causal modeling”

- we know what regression does, and how to use it to describe patterns in the data (best linear approximation, counterfactuals, etc.)
- but in many situations we want to use regression models to answer “causal” questions: how does x influence y , holding constant other factors?
- our approach: carefully consider what is included in the residual, and consider estimation approaches (“designs”) that arguably fix the problem.

Today: panel data

Panel data have a *group structure*

- y_{it} = wage of i in period t : group = person
- R_{ic} = rent paid by family i , city c : group = city

Some examples:

1. two identical twins (i=sibling, j=family)
2. workers observed in different jobs in different years (i=year, j=person)
3. students in a given grade assigned to different teachers (3-way: i=student, j=teacher, k=school)

Why are panel data useful? Consider an example:

Suppose the “true” model generating wages is:

$$y_i = x'_i \beta + \nu_i$$

where x includes the usual variables (age, education, gender) and a dummy for government sector job. This is *not* the way we have been describing regression models up to now. Instead, this is a model-based approach where we say that the “true” coefficients of interest are β .

What happens when we run a regression? OVF says:

$$\beta^o = \beta + \pi$$

where

$$\pi = E[x_i x'_i]^{-1} E[x_i \nu_i]$$

Let's look at the auxilliary regression coefficient vector

$$\pi = E[x_i x_i']^{-1} E[x_i \nu_i]$$

Assume $x_i' = (z_i', D_i)$ where z is a $K - 1$ vector that includes a constant and other controls. From FW, we know the last row of the coefficient vector π is

$$\pi_K = E[\xi_i^2]^{-1} E[\xi_i \nu_i]$$

where ξ_i is the residual from a population regression of D_i on all the other elements of x_i (i.e. z_i) :

$$D_i = z_i' \tau + \xi_i.$$

So if D_i is not related to the other z' s, π_K is just the *difference in all unobserved factors that determine wages* between the government and private sector. Otherwise its the difference, after taking out whatever can be explained by the difference in z 's across sectors.

To recap: we assume a “true model”

$$y_i = x'_i \beta + \nu_i$$

We have $x'_i = (z'_i, D_i)$, and we’re interested in the coefficient on the dummy D_i . The population coefficient for the observed x' s is:

$$\begin{aligned}\beta_j^o &= \beta_j + \pi_j \\ &= \beta_j + E[\xi_{ji}^2]^{-1} E[\xi_{ji} \nu_i]\end{aligned}$$

where ξ_{ji} is the part of x_{ji} we can’t explain from the x ’s.

The regression estimates a potentially biased effect, unless $E[x_i \nu_i] = 0$. But in general this will not be true (except in an RCT)

If we are interested in trying to get a “pure” estimate of β_K – the “wage effect” of working in the government sector, after taking account of ability differences – we need another approach.

Suppose we see the same worker in 2 periods, and sometimes she is working in one sector, and sometimes in the other. Write the model for periods 1 and 2 as:

$$\begin{aligned}y_{i1} &= x'_{i1}\beta + \nu_{i1} \\y_{i2} &= x'_{i2}\beta + \nu_{i2}\end{aligned}$$

Now let's make the *modeling assumption* that:

$$\nu_{it} = \alpha_i + \epsilon_{it}$$

So we are breaking out the unobserved stuff into two parts: a permanent component α_i , and time-varying components ϵ_{it} .

Now the model can be written:

$$\begin{aligned}y_{i1} &= x'_{i1}\beta + \alpha_i + \epsilon_{i1} \\y_{i2} &= x'_{i2}\beta + \alpha_i + \epsilon_{i2}\end{aligned}$$

Suppose that we think all the “problems” that cause a correlation of (x_{it}, ν_{it}) arise because of the permanent stuff. So $E[x_{it}\epsilon_{it}] = 0$.

Then we can take differences:

$$\Delta y_i = y_{i2} - y_{i1} = \Delta x_i\beta_1 + \Delta \epsilon_i$$

and “difference out” the ability term!

The key advantage of panel data is that we can potentially eliminate the effect of any variables that are constant within the “group”. To see this more generally, write the model as

$$y_{it} = x'_{it}\beta + \alpha_i + \epsilon_{it} \quad t = 1, 2, \dots, T; i = 1, 2, \dots, N \quad (*)$$

Then consider the averages: $\bar{y}_i, \bar{x}_i, \bar{\epsilon}_i$ (e.g., $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$). Clearly:

$$\bar{y}_i = \bar{x}_i\beta + \alpha_i + \bar{\epsilon}_i \quad (**)$$

(You should check this by taking the average of both sides of (*).

Subtracting the time averages from each period we get:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)' \beta + \epsilon_{it} - \bar{\epsilon}_i \quad (***)$$

This says that in a panel data setting, we can get the coefficients for the time-varying x 's by deviating the data from group means! Note that if there is some element in x_{it} that does not vary within the group, it “drops out”, along with the α_i term. (This means we don't get an estimate of the constant coefficient).

Deviating the (y_{it}, x_{it}) data for each observation in a group from the group means (\bar{y}_i, \bar{x}_i) is called a “within-group” transformation. The OLS estimator of β after deviating from group means is called the “within” estimator. Notice that you could also get an estimate of β from the group means equation (**). This estimate is called the “between” estimator.

If you stare at (**) and think about it, you will see that although there are T deviations from means for each i , they always sum to 0. So you can drop one of the observations for each group.

The between and within estimators will not necessarily be equal. It is sometimes interesting to compare the three possible OLS estimators:

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{NT} \sum_{i,t} (x_{it} x'_{it}) \right)^{-1} \frac{1}{NT} \sum_{i,t} x_{it} y_{it} \\ \hat{\beta}^B &= \left(\frac{1}{N} \sum_i (\bar{x}_i \bar{x}'_i) \right)^{-1} \frac{1}{N} \sum_i \bar{x}_i \bar{y}_i \\ \hat{\beta}^W &= \left(\frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i) \right)^{-1} \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)\end{aligned}$$

It can be shown that $\hat{\beta}$ is a weighted average of $\hat{\beta}^B$ and $\hat{\beta}^W$. In general we are most concerned about the between estimator, since it will pick up the effect of any fixed unobserved effects in each group most strongly (intuition...).

It can also be shown (pretty easily) that if there are only 2 time periods per group, then the within estimator is the same as the “differenced” estimator discussed above. This will be on the problem set!

It is painful to construct the “deviations from group means” for all the x ’s and for y . So people don’t normally construct $\hat{\beta}^W$ by hand. Instead they rely on the following fact. Consider the expanded model:

$$y_{it} = x'_{it}\beta + (D_{1i}, D_{2i}, \dots, D_{Ni})\theta + \alpha_i + \epsilon_{it}$$

where D_{gi} is a dummy variable for “person g ” ($D_{gi} = 1[g = i]$). This is a model in which we add N dummies, one for each person (and take the constant out of x). Then the OLS estimator from this model is $\hat{\beta}^W$. Why? Mr. Frisch and Mr. Waugh!

Using FW, we know that you get the same coefficients for the x ’s if you first regress the x ’s on the dummies, then take the residuals, and regress y on the residuals. Let’s think what this does...

First we regress x_{kit} on the person dummies for each k . But we know from Lecture 9 that the coefficients for the dummies will just be the person means of x_{kit} . In other words, if we fit the model:

$$x_{kit} = (D_{1i}, D_{2i}, \dots, D_{Ni})\theta^k + \xi_{kit}$$

we'll get that the i^{th} row of $\hat{\theta}^k$ will be \bar{x}_{ki} . So the "residual" will be $\xi_{kit} = x_{kit} - \bar{x}_{ki}$. Now what happens in the 2nd step of FW when we regress y_{it} on the vector of deviations from means? We get an OLS coefficient vector:

$$\left(\frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)^T \right)^{-1} \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i) y_{it}$$

But

$$\begin{aligned}\frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i) y_{it} &= \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i + \bar{y}_i) \\&= \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) + \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)\bar{y}_i \\&= \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i) + \\&\quad + \frac{1}{N} \sum_i \bar{y}_i \frac{1}{T} \sum_t (x_{it} - \bar{x}_i) \\&= \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)\end{aligned}$$

So we get the within estimator. This is also called a “fixed effects” estimator (since we add “effects” for each group) or an analysis of covariance estimator.

Lets look at an example.

Example 1: return to education using twins.

As we've discussed before, if you are looking at the relationship between wages and education, you might be concerned that there is an "omitted ability effect". Twins can help. Here our groups are "families" and we have 2 observations per family: twin 1 and twin 2. We assume:

$$y_{fi} = \beta_1 + \beta_2 S_{fi} + \alpha_f + \epsilon_{fi} \quad i = 1, 2$$

where y = log wage, S = schooling. We can compare the conventional OLS model (stacking up the data for each twin) to a "within family" estimator.

The data are from the “Princeton Twins Survey”, a survey collected by Orley Ashenfelter and colleagues at the “Twinsburg Twins Festival” in the early 1990s. Grad students and faculty (including DC) interviewed sets of twins over multiple years. (Ashenfelter and Rouse, QJE, 1998). For today we’ll work with a set of 1416 twins (from 708 families). Here are the key results:

OLS: $\log \text{wage} = 1.310 + .0823 * \text{Education}$ $R^2 = 0.084$; $MSE = 0.552$

standard error on education coefficient = 0.007

Within: $\log \text{wage} = 0.012 + .0683 * \text{Education}$ $R^2 = 0.037$;
 $MSE = 0.521$

standard error on education coefficient = 0.013

We can see that $\hat{\beta} = 0.082$ (std err=0.007) while $\hat{\beta}^W = 0.068$ (std err=0.013). The “within” estimator is a little smaller. But there is a problem which we will come back to in a few lectures: sometimes twins mis-report their sibling’s education. The survey was designed to measure the degree of error: they asked each twin about his/her own education, and the education of his/her twin. Thus there are 2 estimates of each twin’s schooling. The correlation of the estimates is high but less than 1.

Lecture 12: Panel Data, Difference in Differences Studies

1. recap: model based approaches: correlation vs. causality
2. panel data - controlling for the mean
3. example from twins data
4. difference of differences
5. NJ/PA

Model-based analysis

OLS gives the best linear approximation to $E[y_i|x_i]$. But in many situations we are not really interested in approximating $E[y_i|x_i]$. Instead, we are interested in a “*true model*”:

$$y_i = x'_i \beta + \nu_i$$

in which:

1. we know that y depends on observed x 's and unobserved stuff ν
2. we are interested in $\partial y / \partial x_k$ holding constant ν
3. BUT we are concerned that some of the things in ν are correlated with x

If the true model (“data generating process”) is:

$$y_i = x'_i \beta + \nu_i$$

and $E[x_i u_i] \neq 0$, then the population regression is:

$$\begin{aligned}\beta^{OLS} &= \beta + \pi \\ \pi &= E[x_i x'_i]^{-1} E[x_i \nu_i]\end{aligned}$$

i.e., π is the vector of regression coefficients when we regress ν_i on x_i (which we can't really do)!

Model based approach: make *assumptions* on the nature of ν_i , and how its related to x_i .

Case 1: panel data setting, $i = 1, 2, \dots, N$; $t = 1, 2, \dots, T$:

$$\begin{aligned}y_{it} &= x'_{it}\beta + \nu_{it} \\ \nu_{it} &= \alpha_i + \epsilon_{it} \\ E[x_{it}\nu_{it}] &= E[x_{it}\alpha_i] + E[x_{it}\epsilon_{it}]\end{aligned}$$

We might be able to assume $E[x_{it}\epsilon_{it}] = 0$ – the ϵ_{it} part of the error is “OK”.

Subtract the i -specific averages from each period we get an *estimating model*:

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)' \beta + \epsilon_{it} - \bar{\epsilon}_i$$

which leads to the “within estimator”:

$$\hat{\beta}^W = \left(\frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i) \right)^{-1} \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)$$

Only works for the time-varying x' s.

There is another way to think about the same problem. Assume:

$$y_{it} = x'_{it}\beta + \alpha_i + \epsilon_{it} \quad (1)$$

We want to “control” for α_i . We could do this by introducing a full set of dummies for each person (which leads to the “within estimator”, as shown in Lecture 11).

A more sophisticated approach is to realize that if (1) is correct, then x_{is} does not directly affect y_{it} (for $s \neq t$) – but it may be useful in controlling for α_i . For example, in the $T = 2$ case we have $\{x_{i1}, x_{i2}\}$. We could consider regressing y_{i1} on both x_{i1} and x_{i2} . What will this do?

If (1) is correct, and $E[x_{it}\epsilon_{is}] = 0$, then

$$E[y_{i1}|x_{i1}, x_{i2}] = x'_{i1}\beta + E[\alpha_i|x_{i1}, x_{i2}]$$

Now suppose we believe that x_{i1}, x_{i2} are “equally informative” about α_i . Then:

$$E[\alpha_i|x_{i1}, x_{i2}] = \bar{x}'_i\gamma$$

Relating α_i to the *average* of the x 's makes sense if we think there is nothing “special” about any particular t . In this case

$$\alpha_i = \bar{x}'_i\gamma + \phi_i$$

where ϕ_i has the property that $E[x_{it}\phi_i] = 0$ for any t .

Combining 2 equations we get a different *estimating model*:

$$\begin{aligned}y_{it} &= x'_{it}\beta + \alpha_i + \epsilon_{it} \\ \alpha_i &= \bar{x}'_i\gamma + \phi_i \\ \Rightarrow y_{it} &= x'_{it}\beta + \bar{x}'_i\gamma + \phi_i + \epsilon_{it}\end{aligned}$$

where now the remaining error is $\phi_i + \epsilon_{it}$ and $E[x_i(\phi_i + \epsilon_{it})] = 0$. So we can do OLS, using \bar{x}_i as an extra set of regressors to control for α_i .

Note that if some element x_{kit} does not vary with time then $x_{kit} = \bar{x}_{ki}$ so this regressor appears twice – once as part of x_{it} , once as part of \bar{x}_i . So this approach won't work for the non-varying x' s.

We have a new model to estimate by OLS:

$$y_{it} = x'_{it}\beta + \bar{x}'_i\gamma + \epsilon'_{it}, \quad \epsilon'_{it} = \phi_i + \epsilon_{it}$$

What happens when we include \bar{x}_i as an extra regressor? We need to use FW. We'll prove the case where $x_{it} = x_{1it}$ and $\bar{x}_i = \bar{x}_{1i}$.

From FW, we know

$$\hat{\beta}_1 = \left(\frac{1}{NT} \sum_{i,t} \hat{\xi}_{it}^2 \right)^{-1} \frac{1}{NT} \sum_{i,t} \hat{\xi}_{it} y_{it}$$

where $\hat{\xi}_{it}$ is the estimated residual after we regress x_{1it} on \bar{x}_{1i} .

What happens when we regress x_{1it} on \bar{x}_{1i} ? The coefficient will be 1!

From the FOC, the coefficient will be 1 if:

$$\frac{1}{NT} \sum_{i,t} \bar{x}_{1i} (x_{1it} - \bar{x}_{1i}) = 0$$

But this is true (exercise!). So we end up with:

$$\hat{\beta}_1 = \left(\frac{1}{NT} \sum_{i,t} (x_{1it} - \bar{x}_{1i})^2 \right)^{-1} \frac{1}{NT} \sum_{i,t} (x_{1it} - \bar{x}_{1i}) y_{it}$$

which is just the “within” estimator.

The same algebra works when there are many rows in x_{it} .

So now we've shown that you can obtain the "within" estimator in 3 ways:

1. construct all the deviations from means and estimate

$$\hat{\beta}^W = \left(\frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i) \right)^{-1} \frac{1}{NT} \sum_{i,t} (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)$$

2. estimate a model with dummies for each i – group:

$$y_{it} = x'_{it}\beta + (D_{1i}, D_{2i}, \dots, D_{Ni})\theta + \epsilon''_{it}$$

3. estimate a model that includes x_{it} and \bar{x}_i :

$$y_{it} = x'_{it}\beta + \bar{x}'_i\gamma + \epsilon'_{it}$$

Let's take another look at the twins sample. Recall the model is

$$y_{fi} = \beta_1 + \beta_2 S_{fi} + \alpha_f + \epsilon_{fi} \quad i = 1, 2$$

where y = log wage, S = schooling. The estimating model is

$$y_{fi} = \beta_1 + \beta_2 S_{fi} + \gamma \bar{S}_f + \epsilon'_{fi}$$

In the model for twin 1 we include his/her schooling and average schooling of twin 1 and twin 2 from the same family. We have a sample of 705 twin-pairs.

Note that you can break apart the augmented model:

$$\begin{aligned}y_{fi} &= \beta_1 + \beta_2 S_{fi} + \gamma \bar{S}_f + \epsilon'_{fi} \\&= \beta_1 + \beta_2 S_{fi} + \gamma [(S_{fi} + S_{f\sim i})/2] + \epsilon'_{fi} \\&= \beta_1 + (\beta_2 + \gamma/2) S_{fi} + (\gamma/2) S_{f\sim i} + \epsilon'_{fi}\end{aligned}$$

Which shows that we could just include S_{fi} and $S_{f\sim i}$, then get an estimate of $\gamma/2$ on sibling's education which we then subtract off the coefficient on own education to get a "clean" estimate of β .

Sometimes people don't have information on sibling education so instead they use parent's education. Let's compare that estimate too.

Alternative Estimates of Return to Education Using Twins Sample

	OLS	OLS with Family Dummies	OLS with Mean Educ. Of Pair	OLS with Educ. Of Sibling	OLS with Educ. Of Father
Own Education	0.0802 (0.0071)	0.0678 (0.0131)	0.0678 (0.0197)	0.0749 (0.0106)	0.0843 (0.0075)
Mean Educ. of Pair	--	--	0.0143 (0.0212)	--	--
Sibling Education	--	--	--	0.0072 (0.0106)	--
Father's Education	--	--	--	--	-0.0097 (0.0052)

Difference in Differences

Consider the causal model:

$$y_{it} = x'_{it}\beta + \nu_{it}$$

The assumption in the basic panel data approach is that the “problem” with ν_{it} is confined to the time-invariant component. We decompose $\nu_{it} = \alpha_i + \epsilon_{it}$ and **assume** $E[x_{it}\epsilon_{it}] = 0$. It can be very hard to evaluate this assumption.

There is a case where it is easier. Suppose the component of x_{it} we’re interested in has a specific “design feature”:

$$\begin{aligned}x_{1it} &= 0 & i \in \text{group 0, all } t \\x_{1it} &= 0 & i \in \text{group 1, } t < t^* \\x_{1it} &= 1 & i \in \text{group 1, } t \geq t^*\end{aligned}$$

In this case, x_{1it} is a dummy variable that is 0 for all members of group 0, is also 0 for all members of group 1 in the periods up to $t^* - 1$, and then “turns on” in period t^* for group 1:

$$x_{it} = D_i \times 1[t \geq t^*]$$

In this design we call group 0 the “comparison” group and group 1 the “treatment” group.

If units were *randomly assigned* to the 2 groups we could simply use the data for the post-treatment period and estimate a model like:

$$y_{it} = \beta_0 + D_i\beta_1 + \nu_{it}$$

But what if assignment is non-random?

Without random assignment, the population OLS estimator is:

$$\beta_1 + \pi_t$$

$$\pi_t = \frac{E[(D_i - E[D_i])\nu_{it}]}{E[(D_i - E[D_i])^2]}$$

Suppose we think that $\nu_{it} = \alpha_i + \epsilon_{it}$ and **all the problem with ν_{it} arises through α_i .** In this case:

$$\pi_t = \frac{E[(D_i - E[D_i])\alpha_i]}{E[(D_i - E[D_i])^2]}$$

Let's think about a period before treatment, say period 1. We could run a regression of y_{i1} on D_i even though there is no treatment yet. So the “true” causal model is:

$$y_{i1} = \beta_0 + \nu_{i1}$$

but we estimate a model including a constant and D_i . In this case we expect to recover an OLS coefficient:

$$\pi_1 = \frac{E[(D_i - E[D_i])\nu_{i1}]}{E[(D_i - E[D_i])^2]}$$

But under our assumption this should be the same as π_t **and it should be constant in every pre-treatment period.**

This gives rise to the “difference in differences” idea. The difference in outcomes in the pre-treatment period is an estimate of the bias term. In post-treatment periods the difference is the sum of the treatment effect β_1 and the bias. So we want to eliminate the bias by subtracting the post-treatment difference from the pre-treatment difference: the D-in-D.

Example:

Card-Krueger study of the increase in the minimum wage in New Jersey.

- NJ min. wage rose from \$4.25/hr to \$5.05/hr on April 1, 1992
- surveys of restaurants in NJ and PA before and after

Assumed causal model is:

$$y_{rst} = \beta_0 + \beta_1 High_{rst} + \beta_2 Post_{rst} + v_{rst}$$

where y_{rst} is employment at restaurant r in state s in period $t = 1, 2$, $High_{rst}$ is a dummy for the “high” min. wage, and $Post_{rst}$ is a dummy for the second period (which we need to adjust for seasonal factors). Notice that

$$High_{rst} = NJ_{rst} \times Post_{rst}$$

so we have a D-in-D design. We are concerned that v_{rst} may be correlated with $High$ but we assume that $v_{rst} = \alpha_{rs} + \epsilon_{rst}$ so this correlation can be estimated using the pre-period.

Assumed causal model is:

$$y_{rst} = \beta_0 + \beta_1 High_{rst} + \beta_2 Post_{rst} + \alpha_{rs} + \epsilon_{rst}$$

Estimating model:

$$y_{rst} = \gamma_0 + \gamma_1 NJ_{rst} \times Post_{rst} + \gamma_3 Post_{rst} + \gamma_4 NJ_{rst} + \epsilon_{rst}$$

$$\gamma_0 = \beta_0 + E[\alpha_{rs}|NJ = 0],$$

$$\gamma_4 = E[\alpha_{rs}|NJ = 1] - E[\alpha_{rs}|NJ = 0]$$

Average Full Time Equivalent Employment Per Restaurant, NJ and PA

	PA	NJ	NJ - PA
Mean FTE Employment, Before	23.33 (1.35)	20.44 (1.44)	-2.89 (1.44)
Mean FTE Employment, After	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
Change in Mean FTE Emp.	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Average Full Time Equivalent Employment Per Restaurant, NJ Stores with Different Wages

	Starting Wage Before Rise in Min.			Low Wage - High Wage
	\$4.25	\$4.26-4.99	>\$5.00	
Mean FTE Employment, Before	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)	-2.69 (1.37)
Mean FTE Employment, After	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)	0.67 (1.44)
Change in Mean FTE Emp.	1.32 (0.95)	0.87 (0.84)	-2.04 (1.14)	3.36 (1.48)

Lecture 13: Model-based analysis, part 2

- more general difference-of-differences
- introduction to instrumental variables

Let's consider a panel data model with individuals i in different periods t . We assume there are 2 groups (indexed by D_i), one of which is "treated" in periods $t \geq t^*$ (indicated by the *Post* dummy). Our assumed model is:

$$y_{it} = \delta_t + z'_{it}\beta + \alpha_i + \epsilon_{it} \quad D_i = 0 \text{ (comparison)}$$

$$y_{it} = \delta_t + z'_{it}\beta + Post_t\theta + \alpha_i + \epsilon_{it} \quad D_i = 1 \text{ (program)}$$

- there is a *separate intercept* for each period, δ_t . These are called "time effects": they are included to allow for movements over time the mean of y . Crucially, the time shifts are the same for the two groups.

Observations:

1. if we only have the program group we cannot distinguish between changes in δ_t and the treatment effect.
2. if we only have data for the “post” period, we cannot distinguish between a gap due to *unobserved permanent differences* between the groups:

$$\pi = E[\alpha_i | D_i = 1] - E[\alpha_i | D_i = 0]$$

and the program effect.

3. Let $\bar{\alpha}_1 = E[\alpha_i | D_i = 1]$, $\bar{\alpha}_0 = E[\alpha_i | D_i = 0]$, and define $\phi_i = \alpha_i - E[\alpha_i | D_i]$.

Then we get a combined estimating model:

$$y_{it} = \bar{\alpha}_0 + \delta_t + z'_{it}\beta + D_i\pi + (Post_t \times D_i)\theta + \phi_i + \epsilon_{it}$$

$$y_{it} = \bar{\alpha}_0 + \delta_t + z'_{it}\beta + D_i\pi + (Post_t \times D_i)\theta + \phi_i + \epsilon_{it}$$

- $E[y_{it}|z_{it}, D_i = 0] = \bar{\alpha}_0 + \delta_t + z'_{it}\beta$
- $E[y_{it}|z_{it}, D_i = 1, Post = 0] = \bar{\alpha}_1 + \delta_t + z'_{it}\beta$
- $E[y_{it}|z_{it}, D_i = 1, Post = 1] = \bar{\alpha}_1 + \delta_t + z'_{it}\beta + \theta$

In any pre-period, for obs $z_{it} = z$:

$$E[y_{it}|z_{it} = z, D_i = 1, Post = 0] - E[y_{it}|z_{it} = z, D_i = 0, Post = 0] = \pi$$

In any post-period, for obs $z_{it} = z$:

$$E[y_{it}|z_{it} = z, D_i = 1, Post = 1] - E[y_{it}|z_{it} = z, D_i = 0, Post = 1] = \pi + \theta$$

$$y_{it} = \bar{\alpha}_0 + \delta_t + z'_{it}\beta + D_i\pi + (Post_t \times D_i)\theta + \phi_i + \epsilon_{it}$$

We can estimate the model as follows:

- 1) include dummies for each time period (excluding the first)
- 2) include dummy for the program group
- 3) include z_{it}
- 4) include the variable $D_{it} \times Post$ which is 0 in all periods for the comparison group, 0 for the “pre-treatment” period(s) for the treated group, and 1 in the “post-treatment” period(s) for the treated group

The key assumption for DD is that the two groups would exhibit “parallel trends” in the absence of the program effect. With 2 or more periods in the pre- or post-periods we can evaluate the truth of this assumption. DD allows $E[D_i\epsilon_{it}] \neq 0$ but *assumes* $E[D_i\epsilon_{it}] = 0$. “Selection into treatment” – i.e., $P(D_i = 1)$ – cannot depend on the transitory error components.

This rules out things like:

- management shakeup if a firm is performing worse than average
- enrollment in retraining if a person loses a job
- assignment of kids to different classes based on grades in previous year

Now let's go back to consideration of a *true model*

$$y_i = x'_i \beta + u_i$$

in which:

1. not all the x' s we can see belong in the model
2. we know something about u_i

Possible approaches

- a. use the other x' s to “control for” u_i (or part of u_i)
- b. use some extra variable z to isolate part of x that's unrelated to u

Let's move to the second approach, which is the method of *instrumental variables* (IV) also called two-stage least squares (2sls). If the model is

$$y_i = x'_i \beta + u_i$$

and $E[x_i u_i] \neq 0$, when we estimate by OLS we get a $\hat{\beta}$ that includes the part of u_i that we can predict with x_i :

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N} \sum_i x_i x'_i \right)^{-1} \frac{1}{N} \sum_i x_i y_i \\ &= \left(\frac{1}{N} \sum_i x_i x'_i \right)^{-1} \frac{1}{N} \sum_i x_i (x_i \beta + u_i) \\ \Rightarrow \hat{\beta} - \beta &= \left(\frac{1}{N} \sum_i x_i x'_i \right)^{-1} \frac{1}{N} \sum_i x_i u_i \not\rightarrow 0\end{aligned}$$

In IV we assume there is some set of variables z_i with $\#\text{rows}(z) \geq \#\text{rows}(x)$ and

$$E[z_i u_i] = 0$$

Then we find the part of x_i that is explained by z_i , and use only that in estimating the regression. This is implemented by regressing each row of x on z , then using the predicted fits as explanatory variables in a model for y . Hence the name “two stage” least squares.

Note the key assumption: the z 's are uncorrelated with u .

This is equivalent to assuming that z **does not directly effect** y once you control for x -- so the only reason it is related to y is because it helps to explain x

Let's start with the case of one regressor and a constant: $(1, x_i)$:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

We assume that there is some z_i such that: (a) $E[z_i u_i] = 0$ and (b) $\text{cov}[x_i, z_i] \neq 0$. We first get the part of x that is “clean” by the “first stage” regression:

$$x_i = \pi_0 + \pi_1 z_i + \eta_i$$

When we run this regression we know the FOC will require

$$\frac{1}{N} \sum_i z_i (x_i - \hat{\pi}_0 + \hat{\pi}_1 z_i) = 0$$

Let's call $\text{pred}x_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$. So we know from the FOC that

$$\frac{1}{N} \sum_i z_i(x_i - \text{pred}x_i) = 0$$

Now we consider a regression of y_i on a constant and $\text{pred}x_i$. We know the OLS coefficient from this “second stage” regression will be:

$$\hat{\beta}_1^{iv} = \frac{\frac{1}{N} \sum_i (\text{pred}x_i - \bar{\text{pred}}x) y_i}{\frac{1}{N} \sum_i (\text{pred}x_i - \bar{\text{pred}}x)^2}$$

and now let's write:

$$y_i = \beta_0 + \beta_1 x_i + u_i = \beta_0 + \beta_1 \text{pred}x_i + u_i + \beta_1(x_i - \text{pred}x_i)$$

$$\begin{aligned}
\hat{\beta}_1^{iv} &= \frac{\frac{1}{N} \sum_i (predx_i - \overline{predx}) y_i}{\frac{1}{N} \sum_i (predx_i - \overline{predx})^2} \\
&= \frac{\frac{1}{N} \sum_i (predx_i - \overline{predx})(\beta_0 + \beta_1 predx_i + u_i + \beta_1(x_i - predx_i))}{\frac{1}{N} \sum_i (predx_i - \overline{predx})^2}
\end{aligned}$$

Now lets look at the terms in the numerator.

$$1) \frac{1}{N} \sum_i (predx_i - \overline{predx}) \beta_0 = 0 \text{ easy!}$$

$$2) \frac{1}{N} \sum_i (predx_i - \overline{predx})(x_i - predx_i) = \frac{1}{N} \sum_i predx_i (x_i - predx_i) = 0$$

why? because $predx_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$ and FOC

So that leaves:

$$\hat{\beta}_1^{iv} = \beta_1 + \frac{\frac{1}{N} \sum_i (predx_i - \overline{predx}) u_i}{\frac{1}{N} \sum_i (predx_i - \overline{predx})^2}$$

Note: $predx_i = \hat{\pi}_0 + \hat{\pi}_1 z_i \Rightarrow \overline{predx} = \hat{\pi}_0 + \hat{\pi}_1 \bar{z}$,

$$\Rightarrow \frac{1}{N} \sum_i (predx_i - \overline{predx}) u_i = \frac{1}{N} \sum_i \hat{\pi}_1 (z_i - \bar{z}) u_i$$

and if $E[z_i u_i] = E[u_i] = 0$ the numerator will converge to 0 in probability.

When we “plug in” $\text{pred}x_i$ we get:

$$y_i = \beta_0 + \beta_1 \text{pred}x_i + u_i + \beta_1(x_i - \text{pred}x_i)$$

the combined error is $u'_i = u_i + \beta_1(x_i - \text{pred}x_i)$. The reason why 2sls “works” is that although we’ve got an omitted error component $= \beta_1(x_i - \text{pred}x_i)$, this is orthogonal to the regressor since

$$\frac{1}{N} \sum_i \text{pred}x_i(x_i - \text{pred}x_i) = 0$$

2sls uses an important feature of the 1st stage regression, which is breaks up x_i into 2 parts: $\text{pred}x_i$ and $(x_i - \text{pred}x_i)$, and these 2 parts are orthogonal!

What is the IV estimate?

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{N} \sum_i (predx_i - \overline{predx}) y_i}{\frac{1}{N} \sum_i (predx_i - \overline{predx})^2}$$

But

$$predx_i = \hat{\pi}_0 + \hat{\pi}_1 z_i \Rightarrow predx_i - \overline{predx} = \hat{\pi}_1(z_i - \bar{z})$$

$$\begin{aligned}\hat{\beta}_1^{IV} &= \frac{\frac{1}{N} \sum_i \hat{\pi}_1(z_i - \bar{z}) y_i}{\frac{1}{N} \sum_i (\hat{\pi}_1(z_i - \bar{z}))^2} \\ &= \frac{1}{\hat{\pi}_1} \times \frac{\frac{1}{N} \sum_i (z_i - \bar{z}) y_i}{\frac{1}{N} \sum_i (z_i - \bar{z})^2}\end{aligned}$$

$$\hat{\beta}_1^{IV} = \frac{1}{\hat{\pi}_1} \times \frac{\frac{1}{N} \sum_i (z_i - \bar{z}) y_i}{\frac{1}{N} \sum_i (z_i - \bar{z})^2}$$

Now notice that

$$\frac{\frac{1}{N} \sum_i (z_i - \bar{z}) y_i}{\frac{1}{N} \sum_i (z_i - \bar{z})^2} = \hat{\delta}_1$$

This is the coefficient estimate from a regression of y_i on z_i and a constant. So:

$$\hat{\beta}_1^{IV} = \frac{\hat{\delta}_1}{\hat{\pi}_1}$$

This is the ratio of the coefficients from a regression of y_i on z_i (numerator) and a regression of x_i on z_i (denominator).

Why does this work? We have 2 equations:

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (1)$$

$$x_i = \pi_0 + \pi_1 z_i + \eta_i \quad (2)$$

We can't estimate (1) by OLS because $E[x_i u_i] \neq 0$. BUT we assume $E[z_i u_i] = 0$ (i.e., z_i does not belong in (1) directly). Then we substitute:

$$\begin{aligned} y_i &= \beta_0 + \beta_1(\pi_0 + \pi_1 z_i + \eta_i) + u_i \\ &= \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 z_i + (u_i + \beta_1 \eta_i) \\ &= \delta_0 + \delta_1 z_i + v_i \end{aligned} \quad (3)$$

Let's look at the error term in (3): $v_i \equiv u_i + \beta_1 \eta_i$. This is orthogonal to z_i since we assume $E[z_i u_i] = 0$ and $E[z_i \eta_i] = 0$ since (2) is a population regression. So the regression elves can estimate (3) and with an ∞ sample will return $\delta_1 = \beta_1 \pi_1$. So we need to divide $\delta_1 / \pi_1 = \beta_1$.

Lecture 14: Instrumental Variables II

review

- causal model vs. population regression model
- IV with 1 endogenous variable, no controls

new stuff

- adding controls
- example: IV for RCT with incomplete compliance

population regression model vs. causal model

a) population regression model (PRM)

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i = x'_i \beta + u_i$$

PRM: β is defined by f.o.c.: $E[x_i(y_i - x'_i \beta)] = 0$

OLS: $\hat{\beta}$ is sample analogue of PRM, defined by

$$\frac{1}{N} \sum_{i=1}^N y_i (y_i - x'_i \hat{\beta}) = 0$$

lot's of good things about PRM/OLS - best linear approximation to $E[y_i|x_i]$, Frish-Waugh, omitted variables formula,...

b) causal model

Imagine we are thinking about a manipulation of x_{1i} . We ask “what is the **causal effect** of an exogenous change in x_{1i} , holding constant all other factors?” To address this, we posit:

$$y_i = f(x_{1i}, \dots)$$

where we recognize that many other things affect y . Now for simplicity, assume

$$f(x_{1i}, \dots) = \beta_1 x_{1i} + v_i$$

i.e., the causal effect of x_1 is linear and the other factors are additive. Next, let $\beta_0 = E[v_i]$, and $u_i = v_i - E[v_i]$.

So:

$$\begin{aligned}y_i &= f(x_{1i}, \dots) \\&= \beta_1 x_{1i} + v_i \\&= \beta_0 + \beta_1 x_{1i} + u_i\end{aligned}$$

where $E[u_i] = 0$. This is our prototype “causal model.” The problem with this model is that we cannot be sure that $v_i \perp x_{1i}$. So the same is true of u_i : we may well have:

$$E[x_i u_i] \neq 0$$

Population regression coefficient is $\beta_1 + E[(x_i - E(x_i))u_i]/E(x_i - E(x_i))^2]$

Examples of causal (or “structural”) models:

1. schooling and earnings: $y_i = a + bS_i + u_i$

what's in u ? All other determinants of earnings except schooling.

2. maternal smoking and birthweight

$$BW_i = a + b \text{Smoke}_i + u_i$$

what's in u ? All other determinants of BW except smoking.

Method of IV (or 2SLS) with single expl. variable:

- causal model: $y_i = \beta_0 + \beta_1 x_i + u_i$
- $\exists z_i$ s.t. $E[z_i u_i] = 0$ and $cov[x_i, z_i] \neq 0$.
- fit “first stage” regression:

$$x_{1i} = \pi_0 + \pi_1 z_i + \eta_i$$

- form $predx_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$, fit “second stage” regression:

$$y_i = \beta_0^{iv} + \beta_1^{iv} predx_i + \varphi_i$$

- from causal model:

$$y_i = \beta_0 + \beta_1 predx_i + \beta_1(x_i - predx_i) + u_i$$

$\Rightarrow \varphi_i = \beta_1(x_i - predx_i) + u_i$ AND $E[(predx_i - \overline{predx})(x_i - predx_i)] = 0$ so β_1^{iv} is “good” estimate of β_1

- $(x_i - \text{pred}x_i) \perp \text{pred}x_i \Rightarrow$ the error in the 2nd stage is OK
- reflects a general result: when we fit a regression $x_i = \pi' z_i + \eta_i$
the predicted value of x_i is uncorrelated with the prediction error.
- from FOC, we know

$$\begin{aligned}\frac{1}{N} \sum_i z_{ji}(x_i - \hat{\pi}' z_i) &= 0 \quad \text{for each regressor } j \\ \Rightarrow \frac{1}{N} \sum_i \hat{\pi}_j z_{ji}(x_i - \hat{\pi}' z_i) &= 0 \quad \text{for each } j \\ \Rightarrow \frac{1}{N} \sum_i \hat{\pi}' z_i(x_i - \hat{\pi}' z_i) &= 0\end{aligned}$$

Note that this works for any regression, regardless of the #regressors.

Let's look more closely at the IV formula with 1 regressor:

$$\begin{aligned}
 \hat{\beta}_1^{iv} &= \frac{\frac{1}{N} \sum_i (predx_i - \overline{predx}) y_i}{\frac{1}{N} \sum_i (predx_i - \overline{predx})^2} \\
 predx_i &= \hat{\pi}_0 + \hat{\pi}_1 z_i \\
 \Rightarrow predx_i - \overline{predx} &= \hat{\pi}_1 (z_i - \bar{z}) \\
 \Rightarrow \hat{\beta}_1^{iv} &= \frac{\frac{1}{N} \sum_i \hat{\pi}_1 (z_i - \bar{z}) y_i}{\frac{1}{N} \sum_i (\hat{\pi}_1 (z_i - \bar{z}))^2} \\
 &= \frac{1}{\hat{\pi}_1} \frac{\frac{1}{N} \sum_i (z_i - \bar{z}) y_i}{\frac{1}{N} \sum_i (z_i - \bar{z})^2} = \frac{\hat{\delta}_1}{\hat{\pi}_1}
 \end{aligned}$$

where $\hat{\delta}_1$ is the estimated slope coefficient from the OLS model:

$$y_i = \delta_0 + \delta_1 z_i + \nu_i$$

Where does this formula come from? Note that:

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_i + u_i \quad (\text{structural model}) \\x_i &= \pi_0 + \pi_1 z_i + \eta_i \quad (\text{1st stage}) \\\Rightarrow y_i &= \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 z_i + u_i + \beta_1 \eta_i \\&= \delta_0 + \delta_1 z_i + \nu_i\end{aligned}$$

where

$$\begin{aligned}\delta_0 &= \beta_0 + \beta_1 \pi_0 \\\delta_1 &= \beta_1 \pi_1 \\\nu_i &= u_i + \beta_1 \eta_i\end{aligned}$$

The “structural equation” $y_i = \beta_0 + \beta_1 x_i + u_i$ and the “first stage model” imply a relationship between z and y :

$$\begin{aligned}y_i &= (\beta_0 + \beta_1 \pi_0) + \beta_1 \pi_1 z_i + (u_i + \beta_1 \eta_i) \\&= \delta_0 + \delta_1 z_i + \nu_i\end{aligned}$$

This is called the “reduced form” model. Notice that the reduced form error $\nu_i = u_i + \beta_1 \eta_i$ is “OK”, since $E[z_i \nu_i] = E[z_i u_i] + \beta_1 E[z_i \eta_i] = 0$. So we can estimate the reduced form by OLS and get an estimate of $\hat{\delta}_1 = \widehat{\beta_1 \pi_1}$. And so to estimate β_1 we form:

$$\hat{\beta}_1 = \hat{\delta}_1 / \hat{\pi}_1.$$

We “unpack” β_1 from the reduced form estimate of $\beta_1 \pi_1$ by dividing by our estimate of π_1 .

So far, we have been assuming there is only one x in the structural model. However, the same process works when we have a more complex structural model with other covariates. Consider:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + u_i$$

Let's assume that the "other regressors" $x_{Oi} = (x_{2i}, x_{3i} \dots x_{Ki})'$ are uncorrelated with u_i but regressor 1 is problematic: $E[x_{ji}u_i] = 0$ for $j = 2 \dots K$ but $E[x_{1i}u_i] \neq 0$. Think of an example like

$$y_i = \beta_0 + \beta_1 Schooling_i + \beta_2 age_i + u_i$$

We aren't worried that unobserved ability is correlated with age (unless we are grumpy old guys) but we are worried about schooling.

We follow the same process. We have z_i s.t. (1) $E[z_i u_i] = 0$ and (2) z_i helps predict x_{1i} , conditional on the other regressors.

Step 1. We fit the “first stage” regression:

$$x_i = \pi_0 + \pi_1 z_i + x'_{Oi} \pi_O + \eta_i$$

form $\text{pred}x_{1i} = \hat{\pi}_0 + \hat{\pi}_1 z_i + x'_{Oi} \hat{\pi}_O$

Step 2. We fit the “second stage” regression of y on $\text{pred}x_{1i}$ and x_{Oi} :

$$y_i = \beta_0 + \beta_1 \text{pred}x_{1i} + x'_{Oi} \beta_O + e_i$$

Note that we can think of this as predicting $(x_{1i}, x_{2i}, x_{3i} \dots x_{Ki})$ using $(z_i, x_{2i}, x_{3i} \dots x_{Ki})$ and fitting a second stage model of y on all the predicted regressors.

As in the 1-variable case note that

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_{1i} + x'_{Oi} \beta_O + u_i \\&= \beta_0 + \beta_1 \text{pred}x_{1i} + x'_{Oi} \beta_O + u_i + \beta_1(x_{1i} - \text{pred}x_{1i})\end{aligned}$$

So the error in the second stage model is

$$e_i = u_i + \beta_1(x_{1i} - \text{pred}x_{1i})$$

And as before, $x_{1i} - \text{pred}x_{1i}$ will be uncorrelated with $\text{pred}x_{1i}$ and x_{Oi} . So the second stage model is a legitimate “estimating model” *under our assumptions*.

Using FW,

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{N} \sum_i \widetilde{predx}_{1i} y_i}{\frac{1}{N} \sum_i (\widetilde{predx}_{1i})^2}$$

where \widetilde{predx}_{1i} is the residual after regressing $predx_{1i}$ on $(1, x_{Oi})$.
But

$$\begin{aligned} predx_{1i} &= \hat{\pi}_0 + \hat{\pi}_1 z_i + x'_{Oi} \hat{\pi}_O \\ &= \hat{\pi}_1 z_i + x'_{(\sim 1)i} \hat{\pi}_{(\sim 1)} \end{aligned}$$

So what happens when we regress this on $x_{(\sim 1)i} = (1, x_{Oi})$?

Let's think about a more general problem: if $y_i = a y_i^1 + b y_i^2$ and we regress all 3 y 's on the same x , we get $\hat{\beta} = a \hat{\beta}^1 + b \hat{\beta}^2$. Moreover, $y_i - x'_i \hat{\beta} = a(y_i^1 - x_i \hat{\beta}^1) + b(y_i^2 - x_i \hat{\beta}^2)$.

So we have

$$predx_{1i} = \hat{\pi}_1 z_i + x'_{(\sim 1)i} \hat{\pi}_{(\sim 1)}$$

- when we regress $x'_{(\sim 1)i} \hat{\pi}_{(\sim 1)}$ on $x_{(\sim 1)i}$ the residual is 0
- when we regress $\hat{\pi}_1 z_i$ on $x_{(\sim 1)i}$ the residual is $\widetilde{predx}_{1i} = \hat{\pi}_1 \hat{\xi}_i$ where $\hat{\xi}_i$ is the residual from a regression of z_i on $x_{(\sim 1)i}$. So

$$\hat{\beta}_1^{IV} = \frac{\frac{1}{N} \sum_i \hat{\pi}_1 \hat{\xi}_i y_i}{\frac{1}{N} \sum_i (\hat{\pi}_1 \hat{\xi}_i)^2} = \frac{1}{\hat{\pi}_1} \frac{\frac{1}{N} \sum_i \hat{\xi}_i y_i}{\frac{1}{N} \sum_i \hat{\xi}_i^2} = \frac{\hat{\delta}_1}{\hat{\pi}_1}$$

where $\hat{\delta}_1$ is the coefficient on z_i from the regression of y_i on z_i and $x_{(\sim 1)i}$ (FW).

As in the 1-variable case:

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_{1i} + x'_{Oi} \beta_O + u_i \quad (\text{structural model}) \\x_i &= \pi_0 + \pi_1 z_i + x'_{Oi} \pi_O + \eta_i \quad (\text{1st stage}) \\\Rightarrow y_i &= \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 z_i + x'_{Oi}(\beta_O + \beta_1 \pi_O) + u_i + \beta_1 \eta_i \\&= \delta_0 + \delta_1 z_i + x'_{Oi} \delta_O + \nu_i\end{aligned}$$

So if we estimate the reduced form model relating y to z and x_O , the coefficient on z_i is $\hat{\delta}_1 = \widehat{\beta_1 \pi_1}$. We can unpack the estimate of β_1 by dividing by the coefficient of z_i in the first stage model, $\widehat{\pi}_1$.

Example: “Using geographic variation in proximity to college...”

- causal model: $y_i = \beta_0 + \beta_1 S_i + \beta'_x x_i + u_i$
- 1st stage: $S_i = \pi_0 + \pi_1 Near_i + \pi'_x x_i + \eta_i$
- S = years of education
- y = log wage measured in 1976 (age 26-34)
- $Near_i = 1$ if lived near 4-year college when teenager

First Stage, Reduced Form, and Structural Models for Return to Education

	OLS Model	First Stage Model	Reduced Form	IV Model
Live Near 4-year College as Youth (1966)	--	0.322 (0.083)	0.045 (0.018)	--
Years of Education	0.073 (0.004)	--	--	0.140 (0.055)
Other Controls	YES	YES	YES	YES
Sample Size	3,010	3,010	3,010	3,010

Another example - RCT with incomplete compliance.

Head Start program: provides subsidized pre school for young kids of poor families. Program started in 1960s.

Head Start Impact Study: RCT: randomization (using lottery). Applicants for HS assigned to program group (offered HS spot) or control group (denied). n=4000 kids

Outcome (y): cognitive test after 1 year

Causal Model:

$$y_i = \beta_0 + \beta_1 P_i + u_i$$

where $P_i = 1$ if attended preschool at age 3.

Causal Model:

$$y_i = \beta_0 + \beta_1 P_i + u_i$$

First stage:

$$P_i = \pi_0 + \pi_1 T_i + \eta_i$$

where $T_i = 1$ if assigned to treatment group.

Note that with “full compliance” $\pi_0 = 0$, and $\pi_1 = 1$.

Results from Head Start Impact Study, Age 3 Cohort

Experimental Group	Probability Attend Preschool	Test Score 1 Year Later
Control Group	0.403	-0.302
Treatment Group	0.909	-0.108
Difference: T-C	0.506	0.194
IV Estimate	0.383 (0.051)	

Lecture 15: Instrumental Variables III - LATE

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad (\text{structural model})$$

$$x_i = \pi_0 + \pi_1 z_i + \eta_i \quad (\text{1st stage})$$

$$\begin{aligned} \Rightarrow y_i &= \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 z_i + u_i + \beta_1 \eta_i \\ &= \delta_0 + \delta_1 z_i + \nu_i \end{aligned}$$

$$\hat{\beta}_1^{iv} = \hat{\delta}_1 / \hat{\pi}_1.$$

Example: experiment to study the effect of smoking by mothers on birthweight.

Can't command mothers-to-be to smoke or not. So: conduct an "encouragement design" experiment among pregnant women who smoke and want to quit.

z_i = treatment status = 1 if given "treatment" (counselling)

D_i = dummy for stop smoking (1=stop, 0=keep smoking)

y_i = weight of newly born infant

$$\begin{aligned}y_i &= \beta_0 + \beta_1 D_i + u_i && \text{(structural model)} \\D_i &= \pi_0 + \pi_1 z_i + \eta_i && \text{(1st stage)} \\y_i &= \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 D_i + u_i + \beta_1 \eta_i\end{aligned}$$

This is the “dummy-dummy” case

What are the first stage and reduced form coefficients?

$$\begin{aligned}D_i &= \pi_0 + \pi_1 z_i + \eta_i \\y_i &= \delta_0 + \delta_1 z_i + \nu_i\end{aligned}$$

By the usual (regression elf) logic, we get:

$$\begin{aligned}\hat{\pi}_1 &= \bar{D}^1 - \bar{D}^0 \\\hat{\delta}_1 &= \bar{y}^1 - \bar{y}^0 \\\Rightarrow \hat{\beta}_1^{iv} &= \frac{\bar{y}^1 - \bar{y}^0}{\bar{D}^1 - \bar{D}^0}\end{aligned}$$

So in the “dummy/dummy” case:

$$\hat{\beta}_1^{iv} = \frac{\bar{y}^1 - \bar{y}^0}{\bar{D}^1 - \bar{D}^0}$$

We are dividing the difference in birthweights (between treatment and control groups) by the difference in the fraction of mothers who quit.

Clearly, $\hat{\beta}_1^{iv}$ is driven by the gains in birth outcomes for the moms who quit smoking because of the encouragement. How do we formalize this?

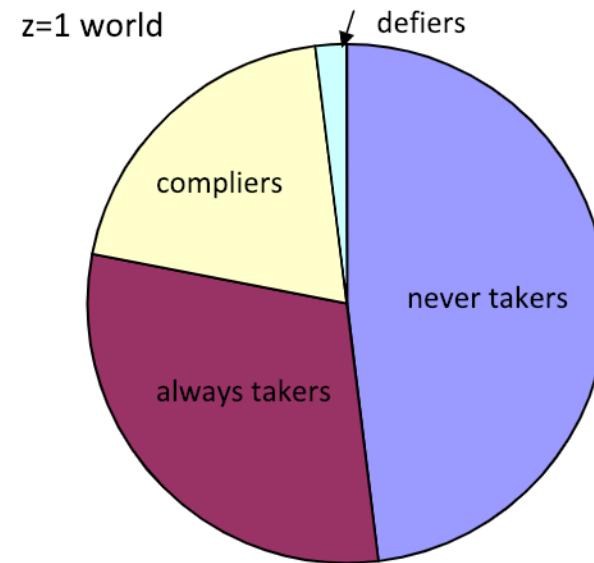
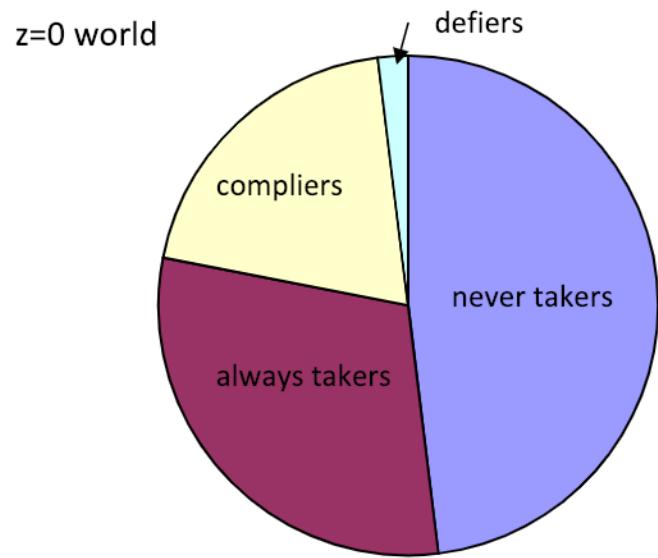
let's think about how people behave and how they can respond to "treatment". Let's introduce 2 new random variables – called "potential outcomes" – that describe types of people and how they react:

- D_{0i} = indicator of quit smoking if person i is assigned to $z = 0$
- D_{1i} = indicator of quit smoking if person i is assigned to $z = 1$

Thinking about the pair of r.v.'s (D_{0i}, D_{1i}) , there are 4 “types” of people:

- $D_{0i} = 0, D_{1i} = 0$ – “never takers” - never quit
- $D_{0i} = 1, D_{1i} = 1$ – “always takers” - always quit
- $D_{0i} = 0, D_{1i} = 1$ – “compliers” - quit if treated
- $D_{0i} = 1, D_{1i} = 0$ – “defiers” - quit if untreated, not if treated.

all 4 groups should be equally likely in T and C groups.



If z is "as good as random" the shares of the 4 groups are the same when $z=0$ and $z=1$

If there are no defiers: $E[D|z=1]-E[D|z=0] = \Pr(\text{complier})$

And: $E[y|z=1]-E[y|z=0] = E[Y(i,1)-Y(i,0)|\text{complier}] \times \Pr(\text{complier})$

4 types based on (D_{0i}, D_{1i}) : $NT, AT, Comp, Def$

Who is still smoking when kid is born?

- in C-group ($z_i = 0$) : smokers = $NT's + Comp's$
- in T-group ($z_i = 1$) : smokers = $NT's$ (if no defiers)

So: if no defiers, difference in fraction of smokers = fraction of $C's$

Formally:

$$\begin{aligned} E[D_i|z_i = 0] &= 0 \times Pr(NT) + 0 \times Pr(Comp) \\ &\quad + 1 \times Pr(AT) + 1 \times Pr(Def) \end{aligned}$$

$$\begin{aligned} E[D_i|z_i = 1] &= 0 \times Pr(NT) + 1 \times Pr(Comp) \\ &\quad + 1 \times Pr(AT) + 0 \times Pr(Def) \end{aligned}$$

If there are no defiers AND the probabilities of the 4 groups are the same in the $z = 0$ and $z = 1$ populations, then

$$E[D_i|z_i = 1] - E[D_i|z_i = 0] = Pr(Comp) = \pi_1$$

Only the compliers change their smoking behavior in the $z = 1$ environment relative to the $z = 0$ environment, so the extra fraction who stop smoking is $Pr(Comp)$

Clearly the difference in birth weights between the $z = 0$ and $z = 1$ groups depends on the changes experienced by the compliers as a result of the intervention in the $z = 1$ environment. To formalize that we introduce *potential outcomes* for y :

$$Y_i(d, z) = \text{outcome for } i \text{ if } D_i = d \text{ and } z_i = z$$

Now if z has no direct effect on the outcome:

$$Y_i(0, 0) = Y_i(0, 1) = Y_{0i} \text{ the outcome if } D_i = 0$$

$$Y_i(1, 0) = Y_i(1, 1) = Y_{1i} \text{ the outcome if } D_i = 1$$

(Aside). Part of what it means to be a “good instrument” is that $-z$ only works through its effect on D . We will assume this “exclusion restriction” is true. When we write

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_i + u_i && \text{(structural model)} \\x_i &= \pi_0 + \pi_1 z_i + \eta_i && \text{(1st stage)}\end{aligned}$$

we are *building in the exclusion restriction*.

Now let's think about our 4 groups:

- never takers: $D_i = 0$ regardless of $z_i \Rightarrow Y_i = Y_{0i}$ for either z_i

- compliers: $D_i = 0$ if $z_i = 0 \Rightarrow$ observe $Y_i = Y_{0i}$ if $z_i = 0$

$D_i = 1$ if $z_i = 1 \Rightarrow$ observe $Y_i = Y_{1i}$ if $z_i = 1$

- always takers: $D_i = 1$ regardless of $z_i \Rightarrow Y_i = Y_{1i}$ for either z_i

- defiers: $D_i = 1$ if $z_i = 0 \Rightarrow$ observe $Y_i = Y_{1i}$ if $z_i = 0$

$D_i = 0$ if $z_i = 1 \Rightarrow$ observe $Y_i = Y_{0i}$ if $z_i = 1$

Now let's think about the mean outcomes we *observe* in the $z = 0$ world:

$$\begin{aligned} E[y_i|z_i = 0] &= E[Y_{i0}|NT] \times Pr(NT) \\ &\quad + E[Y_{i0}|Comp] \times Pr(Comp) \\ &\quad + E[Y_{i1}|AT] \times Pr(AT) \\ &\quad + E[Y_{i1}|Def] \times Pr(Def) \end{aligned}$$

Note that we don't have to condition the probabilities of the 4 types on the value of z because we are assuming random assignment.

Likewise in the $z = 1$ world:

$$\begin{aligned}
 E[y_i|z_i = 1] &= E[Y_{i0}|NT] \times Pr(NT) \\
 &\quad + E[Y_{i1}|Comp] \times Pr(Comp) \\
 &\quad + E[Y_{i1}|AT] \times Pr(AT) \\
 &\quad + E[Y_{i0}|Def] \times Pr(Def)
 \end{aligned}$$

So: if there are no defiers:

$$\begin{aligned}
 E[y_i|z_i = 1] - E[y_i|z_i = 0] &= E[Y_{i1}|Comp] \times Pr(Comp) \\
 &\quad - E[Y_{i0}|Comp] \times Pr(Comp) \\
 &= E[Y_{i1} - Y_{i0}|Comp] \times Pr(Comp)
 \end{aligned}$$

Thus the population regression version of the reduced form model yields

$$\delta_1 = E[y_i|z_i = 1] - E[y_i|z_i = 0] = E[Y_{i1} - Y_{i0}|Comp] \times Pr(Comp)$$

and the population regression verison of the 1st stage model yields

$$\pi_1 = E[D_i|z_i = 1] - E[D_i|z_i = 0] = Pr(Comp)$$

So

$$\frac{\delta_1}{\pi_1} = E[Y_{i1} - Y_{i0}|Comp]$$

Under our assumptions, the ratio of the reduced form effect δ_1 to the 1st stage effect π_1 is $E[Y_{i1} - Y_{i0}|Comp]$.

Let's review. We have a model of the form:

$$\begin{aligned}y_i &= \beta_0 + \beta_1 D_i + u_i && \text{(structural model)} \\D_i &= \pi_0 + \pi_1 z_i + \eta_i && \text{(1st stage)} \\\Rightarrow y_i &= \beta_0 + \beta_1 \pi_0 + \beta_1 \pi_1 z_i + u_i + \beta_1 \eta_i \\&= \delta_0 + \delta_1 z_i + \nu_i && \text{(reduced form)}\end{aligned}$$

In this case, we know

$$\hat{\beta}_1^{iv} = \frac{\bar{y}^1 - \bar{y}^0}{\bar{D}^1 - \bar{D}^0}$$

We are trying to understand what the IV model actually tells us!

2 potential outcomes for the first stage: (D_{0i}, D_{1i}) . In general there are 4 “types” of people:

- $D_{0i} = D_{1i} = 0$ – “never takers”
- $D_{0i} = 0, D_{1i} = 1$ – “compliers”
- $D_{0i} = 1, D_{1i} = 1$ – “always takers”
- $D_{0i} = 1, D_{1i} = 0$ – “defiers”

In theory, 4 potential outcomes $Y(d, z)$ for the outcome: $Y(0, 0) \dots Y(1, 1)$
BUT we assume z does not affect the outcome conditional on D , so we only have 2: (Y_{0i}, Y_{1i}) . (the “exclusion” restriction).

We also assume the joint distribution of $(Y_{0i}, Y_{1i}, D_{0i}, D_{1i})$ is the same whether $z_i = 0$ or $z_i = 1$ – random instrument

With that setup, all we do is take expectations of D and y conditional on $z = 0$ and $z = 1$.

$$\begin{aligned} E[D_i|z_i = 0] &= 0 \times Pr(NT) + 0 \times Pr(Comp) \\ &\quad + 1 \times Pr(AT) + 1 \times Pr(Def) \end{aligned}$$

$$\begin{aligned} E[D_i|z_i = 1] &= 0 \times Pr(NT) + 1 \times Pr(Comp) \\ &\quad + 1 \times Pr(AT) + 0 \times Pr(Def) \end{aligned}$$

$$\Rightarrow E[D_i|z_i = 1] - E[D_i|z_i = 0] = Pr(Comp)$$

(assuming no defiers)

And:

$$\begin{aligned} E[y_i|z_i = 0] &= E[Y_{i0}|NT] \times Pr(NT) \\ &\quad + E[Y_{i0}|Comp] \times Pr(Comp) \\ &\quad + E[Y_{i1}|AT] \times Pr(AT) \\ &\quad + E[Y_{i1}|Def] \times Pr(Def) \end{aligned}$$

$$\begin{aligned} E[y_i|z_i = 1] &= E[Y_{i0}|NT] \times Pr(NT) \\ &\quad + E[Y_{i1}|Comp] \times Pr(Comp) \\ &\quad + E[Y_{i1}|AT] \times Pr(AT) \\ &\quad + E[Y_{i0}|Def] \times Pr(Def) \end{aligned}$$

$$\Rightarrow E[y_i|z_i = 1] - E[y_i|z_i = 0] = (E[Y_{i1}|Comp] - E[Y_{i0}|Comp]) \times Pr(Comp)$$

So with (i) exclusion (ii) as good as random z (iii) no defiers, we have

$$\pi_1 = E[D_i|z_i = 1] - E[D_i|z_i = 0] = Pr(Comp) \quad (1)$$

$$\delta_1 = E[D_i|z_i = 1] - E[D_i|z_i = 0] = E[Y_{i1} - Y_{i0}|Comp] \times Pr(Comp) \quad (2)$$

And so the “population” IV is

$$\beta_1^{IV} = \frac{\delta_1}{\pi_1} = E[Y_{i1} - Y_{i0}|Comp]$$

The IV estimate provides an estimate of the “gain” in y that compliers obtain as a result of being in the $z = 1$ setting and getting $D_i = 1$ instead of $D_i = 0$.

Interpretation:

$$y_i = \beta_0 + \beta_1 D_i + u_i \quad (\text{structural model})$$

$$D_i = \pi_0 + \pi_1 z_i + \eta_i \quad (\text{1st stage})$$

When you look at the structural model, you are tempted to think that it means that everyone with $D = 1$ experiences a gain of β_1 in the outcome. But:

$$y_i = Y_{i0} + (Y_{i1} - Y_{i0})D_i$$

so we should think of (Y_{i0}, Y_{i1}) as random variables, with

$$E[Y_{i0}] = \beta_0, \quad E[Y_{i1}] = \beta_0 + \beta_1,$$

$$E[Y_{i1} - Y_{i0}] = \beta_1$$

But among various subgroups, $E[Y_{i1} - Y_{i0}]$ can be different. And β_1^{IV} identifies the gain for the compliers!

Lecture 16: More on IV and LATE

$$y_i = \beta_0 + \beta_1 D_i + u_i \quad (\text{structural model})$$

$$D_i = \pi_0 + \pi_1 z_i + \eta_i \quad (\text{1st stage}) \quad z_i = 0/1$$

$$\Rightarrow y_i = (\beta_0 + \beta_1 \pi_0) + \beta_1 \pi_1 z_i + (u_i + \beta_1 \eta_i)$$

$$\widehat{\beta}_1^{iv} = \widehat{\delta}_1 / \widehat{\pi}_1 = \frac{\bar{y}^1 - \bar{y}^0}{\bar{D}^1 - \bar{D}^0} \rightarrow E[Y_{1i} - Y_{0i} | \text{complier}]$$

Example: experiment to study the effect of smoking by mothers on birthweight.

Can't command mothers-to-be to smoke or not. So: conduct an "encouragement design" experiment among pregnant women who smoke and want to quit.

z_i = indicator if given "treatment" (counselling, etc)

D_i = dummy for stop smoking (1=stop, 0=keep smoking)

y_i = weight of newly born infant

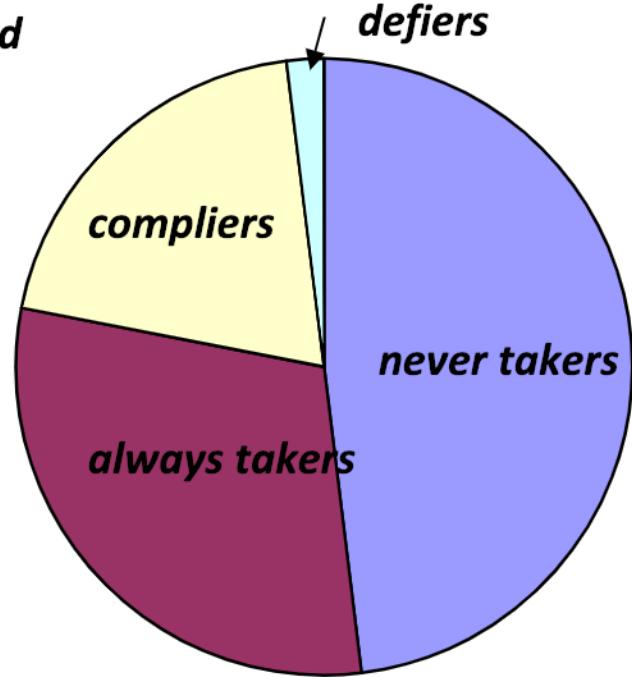
2 potential outcomes for the first stage: (D_{0i}, D_{1i}) . In general there are 4 “types” of people:

- $D_{0i} = D_{1i} = 0$ – “never takers”
- $D_{0i} = 0, D_{1i} = 1$ – “compliers”
- $D_{0i} = 1, D_{1i} = 1$ – “always takers”
- $D_{0i} = 1, D_{1i} = 0$ – “defiers”

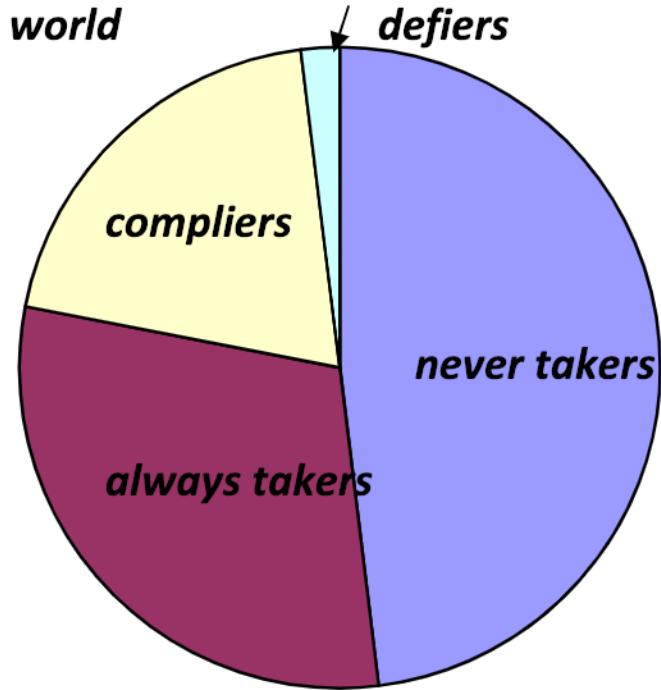
In general, 4 potential outcomes for i : $\{Y_{dzi}\}$, $d = 0, 1$; $z = 0, 1$. BUT we assume z does not affect the outcome conditional on D , so we only have 2: (Y_{0i}, Y_{1i}) . (*Exclusion restriction*).

We also assume the joint distribution of $(Y_{0i}, Y_{1i}, D_{0i}, D_{1i})$ is the same whether $z_i = 0$ or $z_i = 1$ – “as good as random” *instrument*.

$z=0$ world



$z=1$ world



Four types: never takers; compliers; always takers; defiers

If z is "as good as random":

1. $\Pr(\text{type}=j | z) = \Pr(\text{type}=j)$
2. $E[Y_{1i} | \text{type}, z] = E[Y_{1i} | \text{type}]$; $E[Y_{0i} | \text{type}, z] = E[Y_{0i} | \text{type}]$

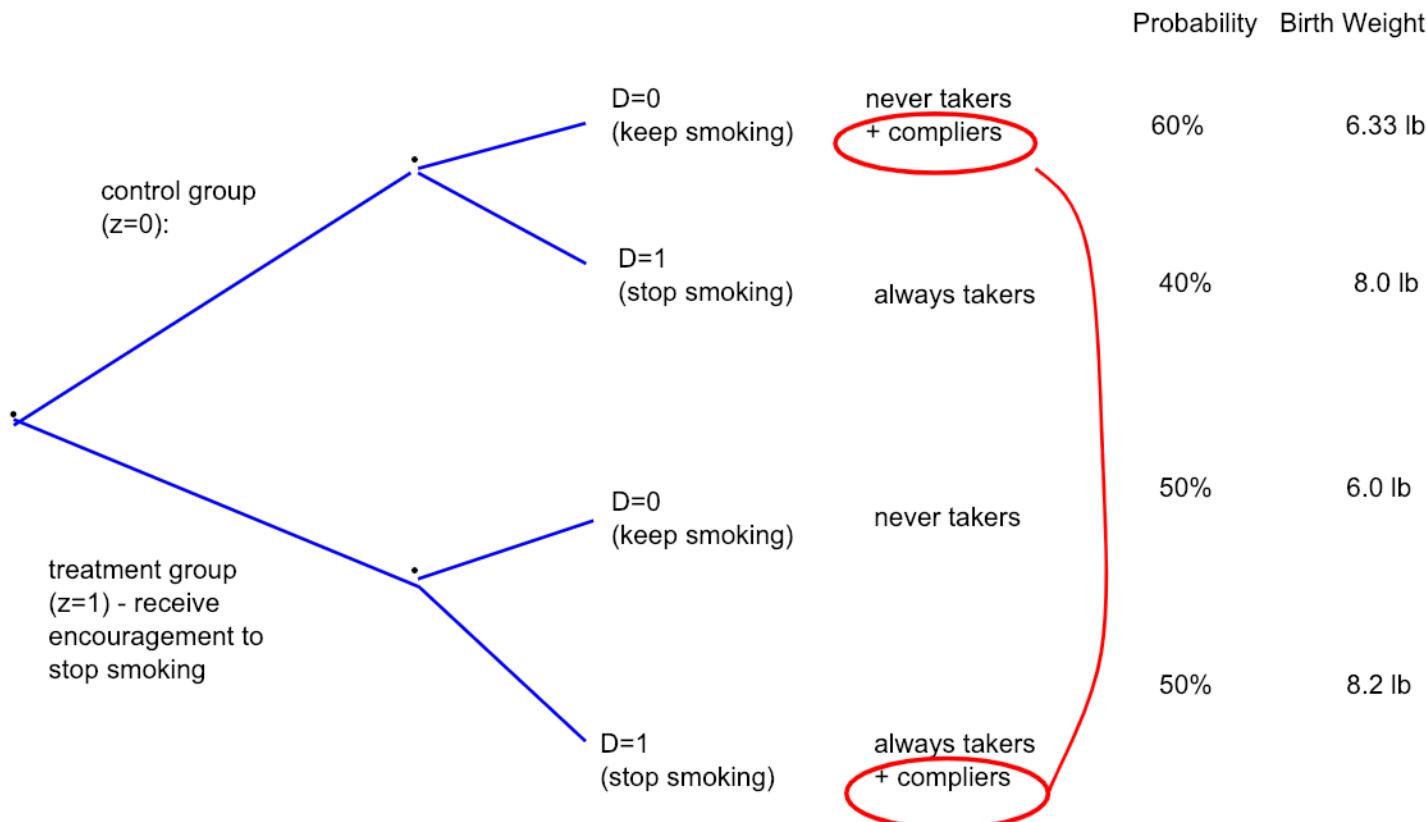
With that setup, all we do is take expectations of D and y conditional on $z = 0$ and $z = 1$.

$$\begin{aligned} E[D_i|z_i = 0] &= 0 \times Pr(\text{nevertaker}) + 0 \times Pr(\text{complier}) \\ &\quad + 1 \times Pr(\text{alwaystaker}) + 1 \times Pr(\text{defier}) \end{aligned}$$

$$\begin{aligned} E[D_i|z_i = 1] &= 0 \times Pr(\text{nevertaker}) + 1 \times Pr(\text{complier}) \\ &\quad + 1 \times Pr(\text{alwaystaker}) + 0 \times Pr(\text{defier}) \end{aligned}$$

$$\Rightarrow E[D_i|z_i = 1] - E[D_i|z_i = 0] = Pr(\text{complier})$$

If there are no defiers...



$$E[D | z=1] - E[D | z=0] = 0.1 \rightarrow \Pr(\text{Complier})=.1$$

$$E[Y | z=1] = 7.1 \text{ lb}; \quad E[D | z=0] = 7.0 \text{ lb}$$

$$\beta_{IV} = (7.1 - 7.0) / (0.50 - 0.40) = 1.0$$

$$\begin{aligned}
E[y_i|z_i = 0] &= E[Y_{i0}|nevertaker] \times Pr(nevertaker) \\
&\quad + E[Y_{i0}|complier] \times Pr(complier) \\
&\quad + E[Y_{i1}|alwaystaker] \times Pr(alwaystaker) \\
E[y_i|z_i = 1] &= E[Y_{i0}|nevertaker] \times Pr(nevertaker) \\
&\quad + E[Y_{i1}|complier] \times Pr(complier) \\
&\quad + E[Y_{i1}|alwaystaker] \times Pr(alwaystaker)
\end{aligned}$$

$$\Rightarrow E[y_i|z_i = 1] - E[y_i|z_i = 0] = E[Y_{i1} - Y_{i0}|complier] \times Pr(complier)$$

So with (i) exclusion (ii) as good as random z (iii) no defiers, we have

$$\pi_1 = E[D_i|z_i = 1] - E[D_i|z_i = 0] = Pr(\text{complier})$$

$$\delta_1 = E[D_i|z_i = 1] - E[D_i|z_i = 0] = E[Y_{i1} - Y_{i0}|\text{complier}] \times Pr(\text{complier})$$

So the “population IV” is

$$\beta_1^{IV} = \frac{\delta_1}{\pi_1} = E[Y_{i1} - Y_{i0}|\text{complier}]$$

β_1^{IV} identifies the gain for the compliers, which may be different from $E[Y_{i1} - Y_{i0}]$

Example:

Self Sufficiency Project (SSP) experiment, conducted in Canada,
early 1990s

- earnings subsidy for people who enter FT work
- subsidy = $0.5 \times (\text{T-monthly earnings})$ $\text{T}=\$2500/\text{mo}$ or $\$3083/\text{mo}$

ONLY for “FT” (30+ hrs/week) at min. wage or more

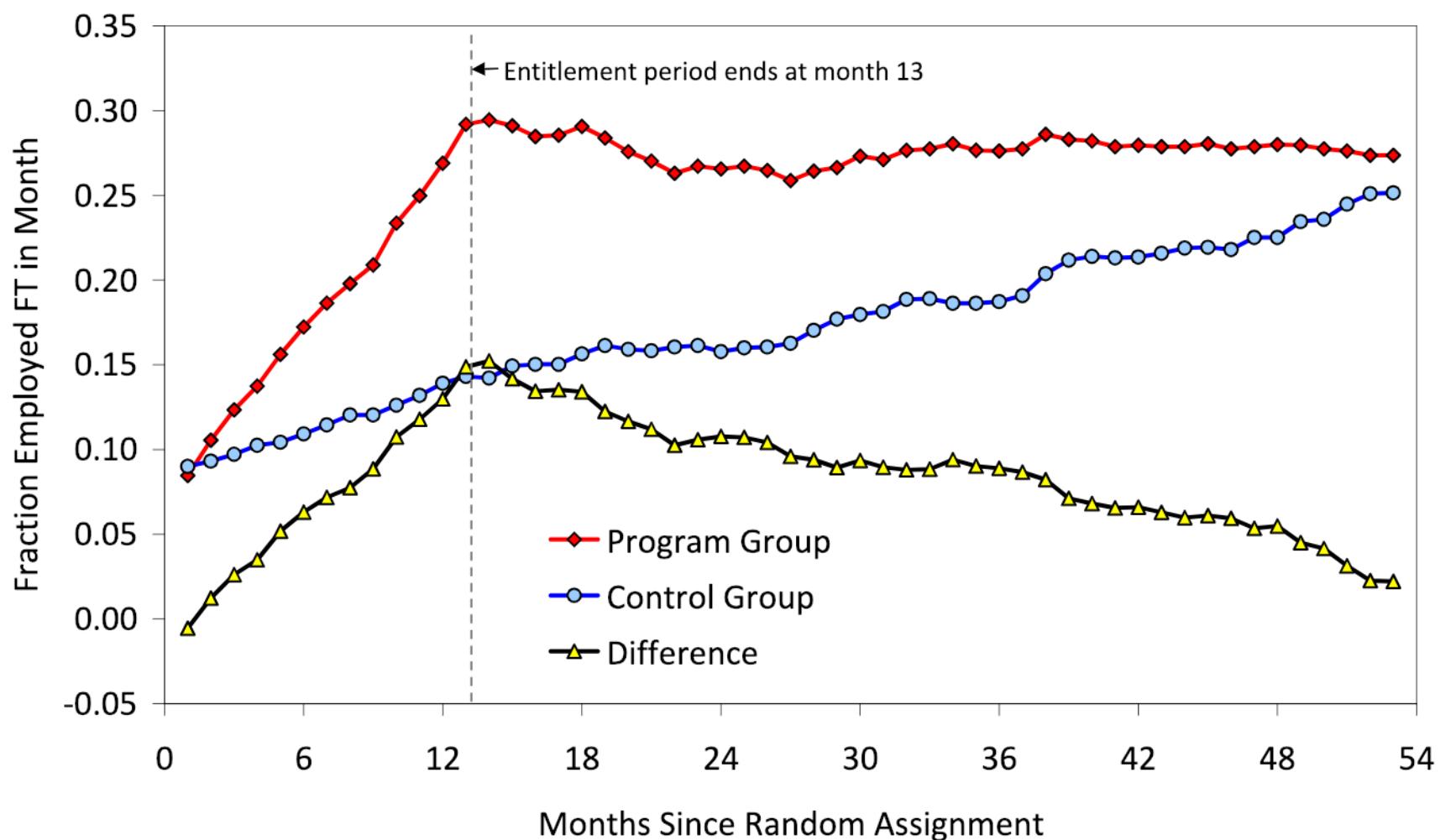
- subsidy lasts for 3 years
- have to start receiving subsidy within 12 mo. of RA

- SSP subsidy = strong incentive to enter FT work and
- BUT: can long-term welfare recipients find *and* maintain FT job?
- and: will they be better off after 3 years of subsidy?

Experiment:

- welfare recipients with kids (95% females), 1+ years on welfare
- randomly allocated to treatment group ($z = 1$) or control ($z = 0$)
- series of surveys measure employment etc.

Monthly FT Empl. Rates of SSP Control and Treatment Groups



Structural model for IA at month 15:

$$IA15_i = \beta_0 + \beta_1 FT15_i + u_i$$

First stage model for FT empl at month 15:

$$FT15_i = \pi_0 + \pi_1 TREAT_i + \eta_i$$

Reduced form model for IA at month 15:

$$IA15_i = \delta_0 + \delta_1 TREAT_i + \nu_i$$

Let's think about SSP:

Is “as good as random” satisfied? YES - really random!

Is exclusion satisfied? *SSP subsidy* \Rightarrow *FT work* \Rightarrow *leave IA*

seems plausible?

Is “no defiers” satisfied? Likely since SSP would never cause you to stop working FT (unless you are really weird).

Summary of SSP Outcomes at Month 15

	Employed FT (1)	On Welfare (IA) (2)
Control Group (n=2718)	0.1472	0.8098
Treatment Group (n=2762)	0.2848	0.6687
Difference: T-C (std error)	0.1376 (0.011)	-0.1411 (0.012)
IV estimate:	-1.027 (0.082)	

We know that there is only partial compliance with the subsidy incentive.

$$Pr(\text{nevertakers, mo.15}) = 1 - 0.2846 = 0.7154$$

$$Pr(\text{always takers, mo.15}) = 0.1471$$

$$Pr(\text{compliers, mo.15}) = 0.1375$$

Note that these fractions are changing as the experiment goes on....

Q: who are the compliers? Are they “like” other people? Are they more or less advantaged?

Let's consider x_i = dummy for HS education

mean for overall control group = 0.446

mean for controls not FT at mo 15 = 0.4146 (compliers+never takers)

mean for controls working FT at mo 15 = 0.6150 (always takers)

mean for overall treat group = 0.457

mean for treats not FT at mo 15 = 0.4140 (never takers)

mean for treats working FT at mo 15 = 0.5763 (compliers+always takers)

So: mean(AT)>mean(NT) and compliers in between?

Let's formalize this. z =treatment status, D =FT at mo. 15

$$E[x_i|z_i = 0, D_i = 1] = E[x_i|AT] = 0.6150 \quad (\text{FT controls})$$

$$E[x_i|z_i = 1, D_i = 1] = E[x_i|AT \text{ or } C] = 0.5763 \quad (\text{FT treatment group})$$

BUT:

$$E[x_i|AT \text{ or } C] = \frac{E[x_i|AT] \times Pr(AT) + E[x_i|C] \times Pr(C)}{Pr(AT \text{ or } C)}$$

So:

$$E[x_i|C] = \frac{E[x_i|AT \text{ or } C] \times Pr(AT \text{ or } C) - E[x_i|AT] \times Pr(AT)}{Pr(C)}$$

$$E[x_i|C] = \frac{E[x_i|AT \text{ or } C] \times Pr(AT \text{ or } C) - E[x_i|AT] \times Pr(AT)}{Pr(C)}$$

$$Pr(AT) = 0.1471$$

$$Pr(AT \text{ or } C) = 0.2841$$

$$Pr(C) = 0.137$$

$$E[x_i|AT] = 0.6150$$

$$E[x_i|AT \text{ or } C] = 0.5763$$

$$\text{So } E[x_i|C] = (0.5763 \times 0.2841 - 0.6150 \times 0.1471) / 0.137 = 0.535$$

$$E[x_i|C] = \frac{E[x_i|AT \text{ or } C] \times Pr(AT \text{ or } C) - E[x_i|AT] \times Pr(AT)}{Pr(C)}$$

Consider the “goofy” 2sls model:

$$\begin{aligned} x_i D_i &= a + b D_i + e_i \quad (\text{structural model}) \\ D_i &= \pi_0 + \pi_1 z_i + \eta_i \quad (\text{1st stage}) \end{aligned}$$

We know that the population IV estimate for b will be:

$$b^{iv} = \frac{E[x_i D_i | z_i = 1] - E[x_i D_i | z_i = 0]}{E[D_i | z_i = 1] - E[D_i | z_i = 0]}$$

$$b^{iv} = \frac{E[x_i D_i | z_i = 1] - E[x_i D_i | z_i = 0]}{E[D_i | z_i = 1] - E[D_i | z_i = 0]}$$

$$1. E[x_i D_i | z_i = 1] = E[x_i | D_i = 1, z_i = 1] \times Pr(D_i = 1 | z_i = 1)$$

$$= E[x_i | AT \text{ or } C] \times Pr(AT \text{ or } C)$$

$$2. E[x_i D_i | z_i = 0] = E[x_i | D_i = 1, z_i = 0] \times Pr(D_i = 1 | z_i = 0)$$

$$= E[x_i | AT] \times Pr(AT)$$

$$3. E[D_i | z_i = 1] - E[D_i | z_i = 0]$$

So b^{iv} estimates the mean of x for the compliers!

Summary: for the simplest possible model with as good as random z :

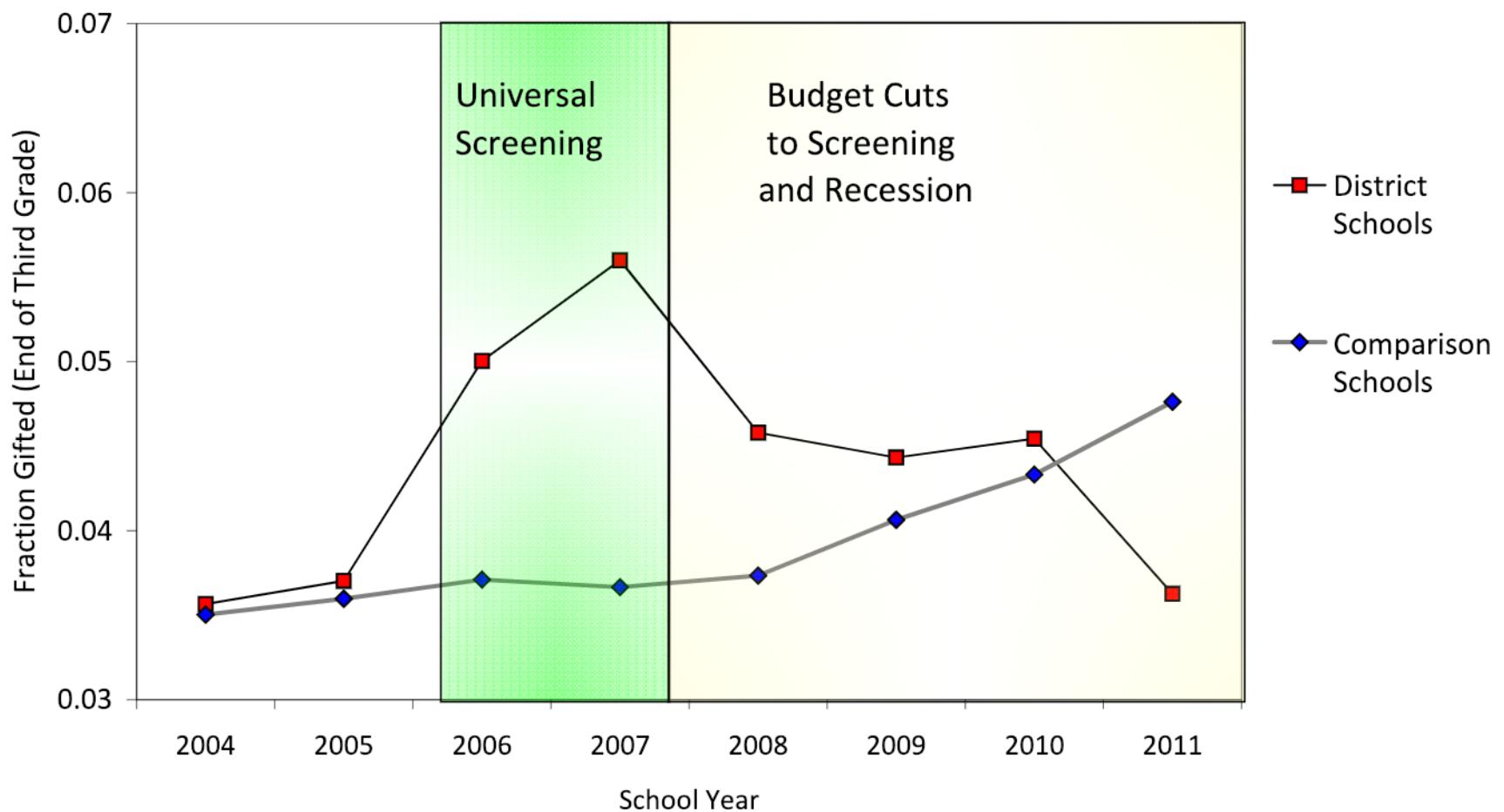
$$y_i = \beta_0 + \beta_1 D_i + u_i \quad (\text{structural model})$$

$$D_i = \pi_0 + \pi_1 z_i + \eta_i \quad (\text{1st stage}) \quad z_i = 0/1$$

$$\hat{\beta}_1^{iv} = \hat{\delta}_1 / \hat{\pi}_1 = \frac{\bar{y}^1 - \bar{y}^0}{\bar{D}^1 - \bar{D}^0} \rightarrow E[Y_{1i} - Y_{0i} | \text{complier}]$$

and we can estimate the characteristics of the compliers.

Figure 1: Trends in Fraction Gifted by End of Third Grade
District Schools versus Matched Comparison Schools



Note: school level data weighted by mean school enrollment in all years. Comparison schools weighted by propensity score \times enrollment. Screening program was eliminated in 2010, affecting students in 2011.

Characteristics of Gifted Students Before and After Screening Program

	Gifted in 2004-2005 (Always Takers)	Gifted in 2006-2007 (Compliers + AT's)	Compliers (Newly Identified Gifted)
Female	0.45 (0.01)	0.47 (0.01)	0.56 (0.12)
White	0.61 (0.01)	0.43 (0.01)	0.08 (0.14)
Black	0.12 (0.01)	0.17 (0.01)	0.23 (0.10)
Hispanic	0.16 (0.01)	0.27 (0.01)	0.46 (0.11)
FRL	0.20 (0.01)	0.35 (0.01)	0.67 (0.13)
School FRL Rate	0.28 (0.01)	0.34 (0.01)	0.47 (0.08)
School Minority Share	0.45 (0.01)	0.54 (0.01)	0.70 (0.08)
Math+Reading (3rd Grade)	1.39 (0.02)	1.22 (0.02)	0.97 (0.16)

Lecture 17: Regression Discontinuity (RD) Methods

- discontinuous assignment models
- Sharp RD
- fuzzy RD

Model based analysis methods:

Interested in “causal effect” of x on y , holding other things constant

Gold standard: randomized experiment

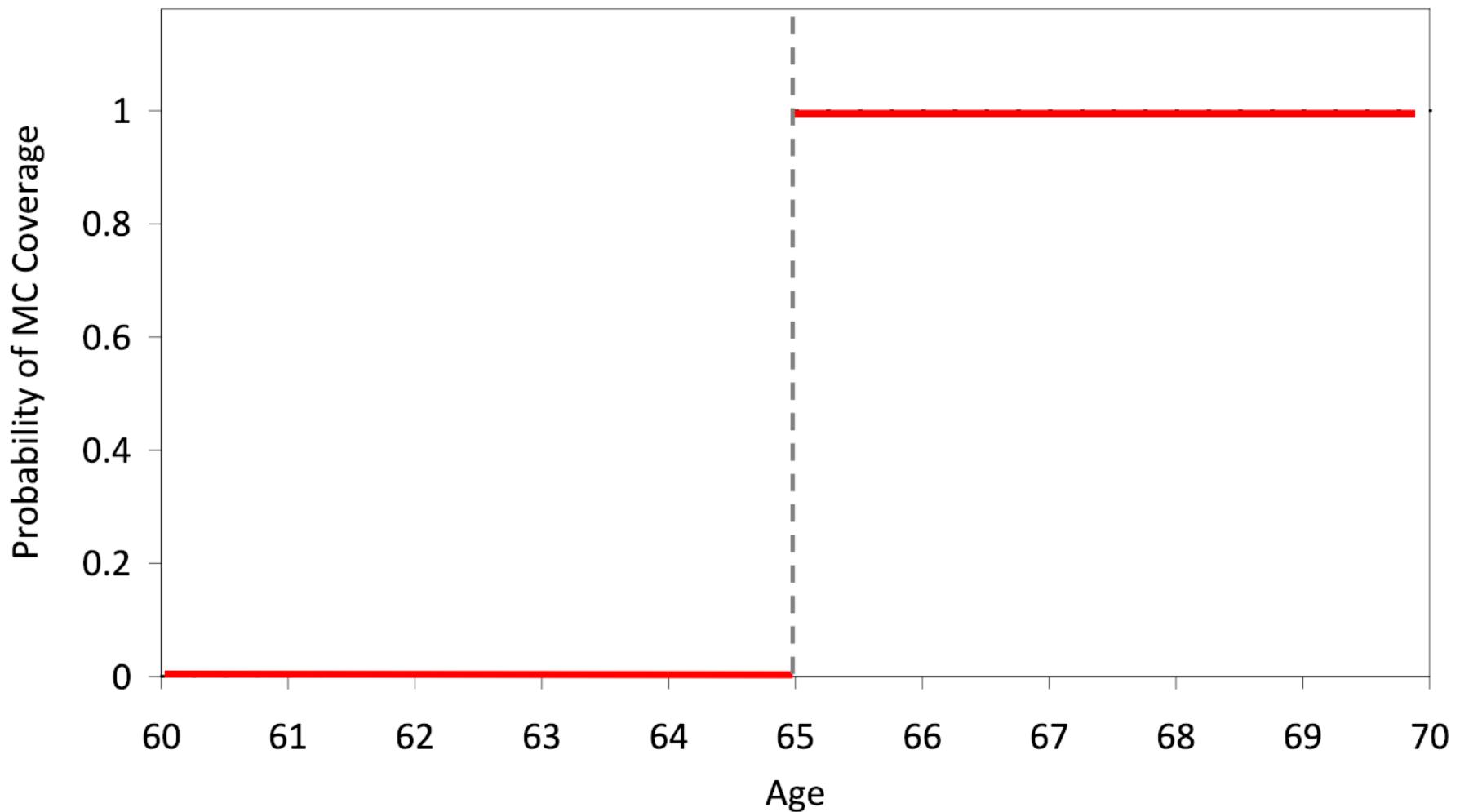
Non-experimental methods:

- 1) OLS and reweighting counterfactuals
- 2) Diff in diff
- 3) IV
- 4) RD

RD arises when assignment is based on a discontinuous function of some “random” or “as good as random” *running variable*:

- 1) election outcomes ($rv=$ vote share)
- 2) admission or financial aid based on test scores ($rv=$ score)
- 3) rules based on job tenure, duration of marriage etc ($rv=$ duration)
- 4) rules based on age ($rv=$ exact day of birth) e.g. Medicare

Medicare Coverage with Age



x_i =running variable

D_i =assignment status: $D_i = 1$ if “treated” otherwise 0

Sharp RD: $D_i = 0$ if $x_i < c$; $D_i = 1$ if $x_i \geq c$ (set $c = 0$)

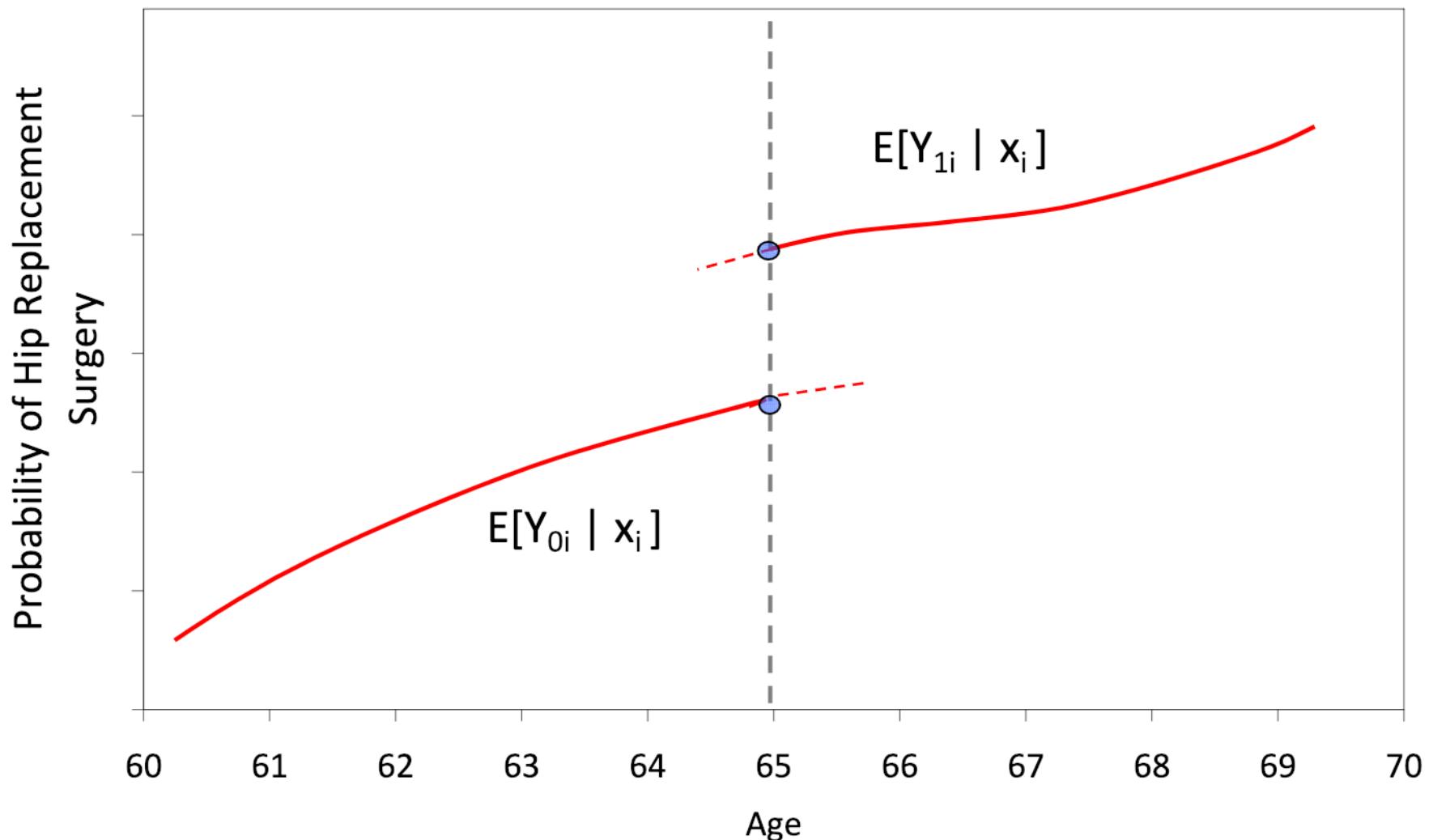
y_i =outcome; potential outcomes Y_{0i} if $D_i = 0$; Y_{1i} if $D_i = 1$

Key assumption:

$E[Y_{0i}|x_i]$ is continuous function of x at $x = 0$ (the cutoff)

$E[Y_{1i}|x_i]$ is continuous function of x at $x = 0$

Probability of Hip Replacement Surgery by Age



Define:

$$\beta_1 = E[Y_{1i}|x_i = 0] - E[Y_{0i}|x_i = 0]$$

This is the treatment effect on agents with $x = 0$.

We estimate $E[Y_{0i}|x_i = 0]$ using data for $x < 0$

and we estimate $E[Y_{1i}|x_i = 0]$ using data for $x \geq 0$

Then we form $\hat{\beta}_1$.

Benchmark case:

$$E[Y_{0i}|x] = \alpha_0 + \alpha_1 x_i \quad \text{and} \quad E[Y_{1i}|x] = E[Y_{0i}|x] + \beta_1.$$

i.e.: *linear* conditional expectations

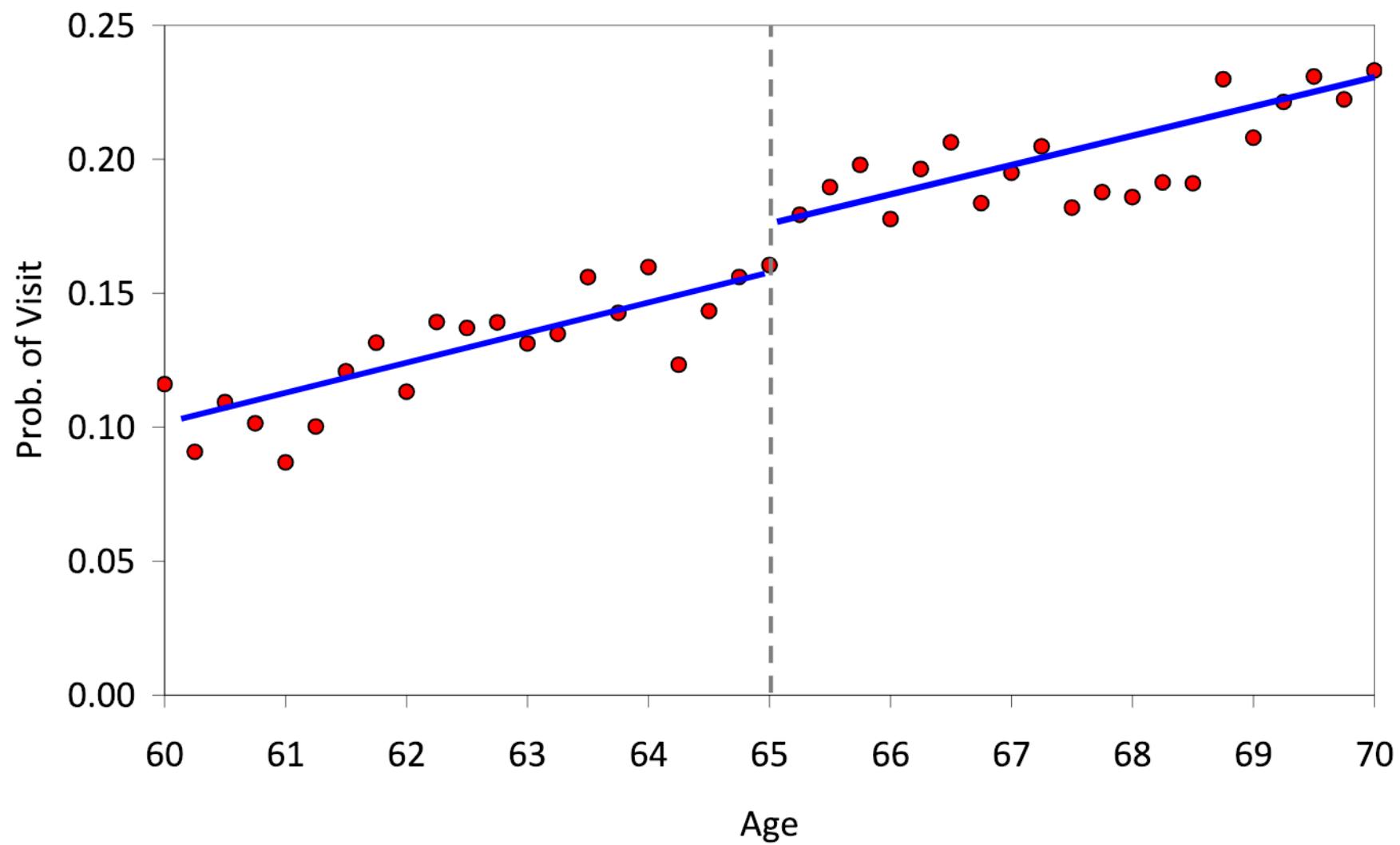
and constant treatment effect: $E[Y_{1i} - Y_{0i}] = \beta_1$.

In this case we can estimate a very simple model:

$$y_i = \alpha_0 + \alpha_1 x_i + \beta_1 D_i + u_i$$

and $E[u_i|x_i, D_i] = E[u_i|x_i] = 0$ (because the regression model is the correct model of conditional expectations).

Probability of Hospital Visit by Age



What does it mean to assume

$E[Y_{0i}|x_i]$ and $E[Y_{1i}|x_i]$ are continuous functions of x at $x = 0$?

Consider samples of $x_i \in (-\theta, 0)$ and associated (Y_{0i}, Y_{1i})

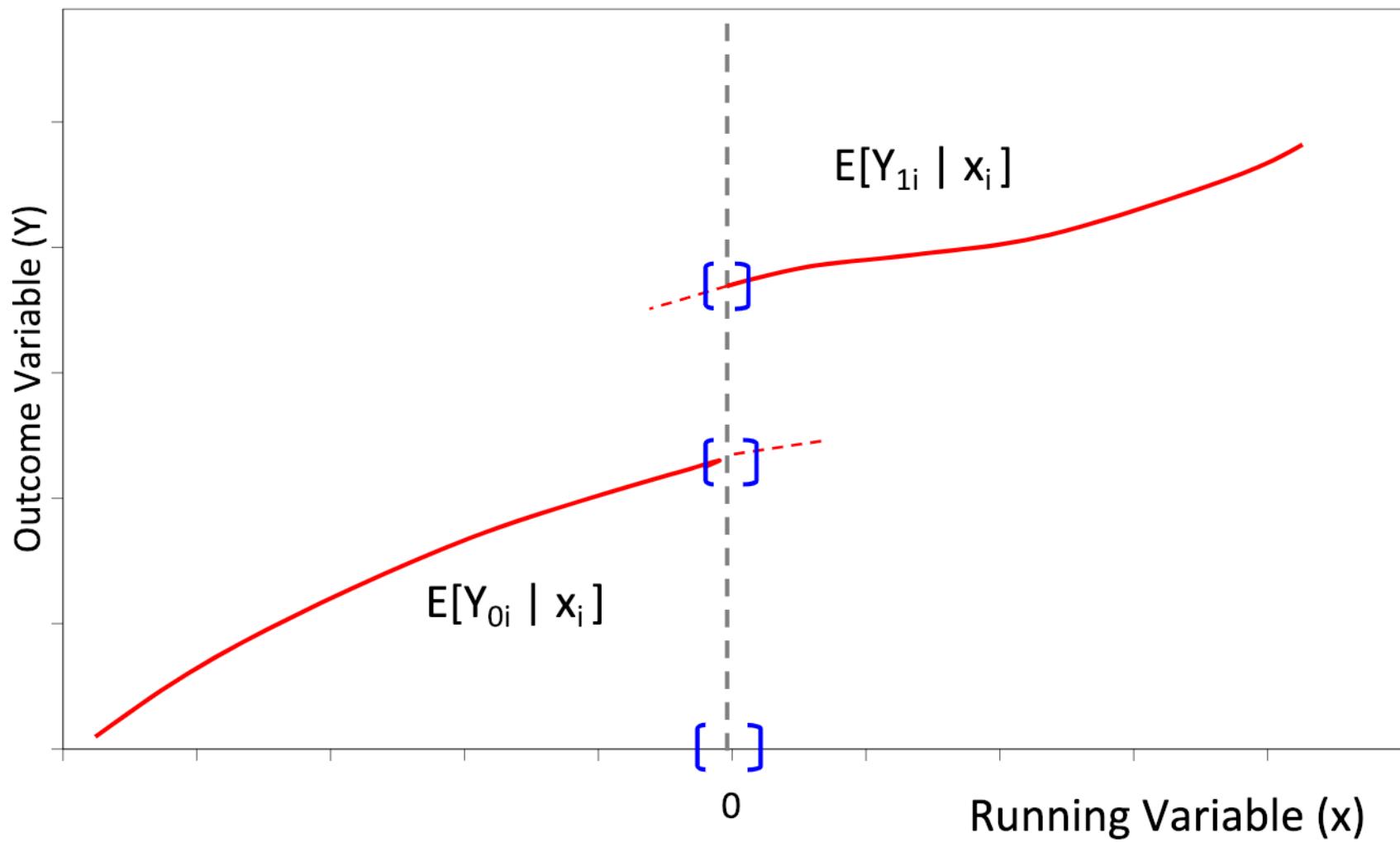
vs. samples of $x_i \in [0, \theta)$ and associated (Y_{0i}, Y_{1i})

⇒ distributions of Y_{0i} from 2 samples are centered at ~same mean

and distributions of Y_{1i} from 2 samples are centered at ~same mean

i.e., as good as random w.r.t means of the pot. outcomes.

Continuity Assumptions of RD Model



A sufficient condition for the required continuity is that around $x = 0$, individual assignment to the plus or minus side is as good as random. This condition has a testable implication:

- if w_i is some predetermined characteristic (determined prior to assignment)

then $E[w_i|x_i]$ will be continuous at $x = 0$.

- we also expect the relative number of observations to trend smoothly at $x = 0$

How could the RD assumptions break down?

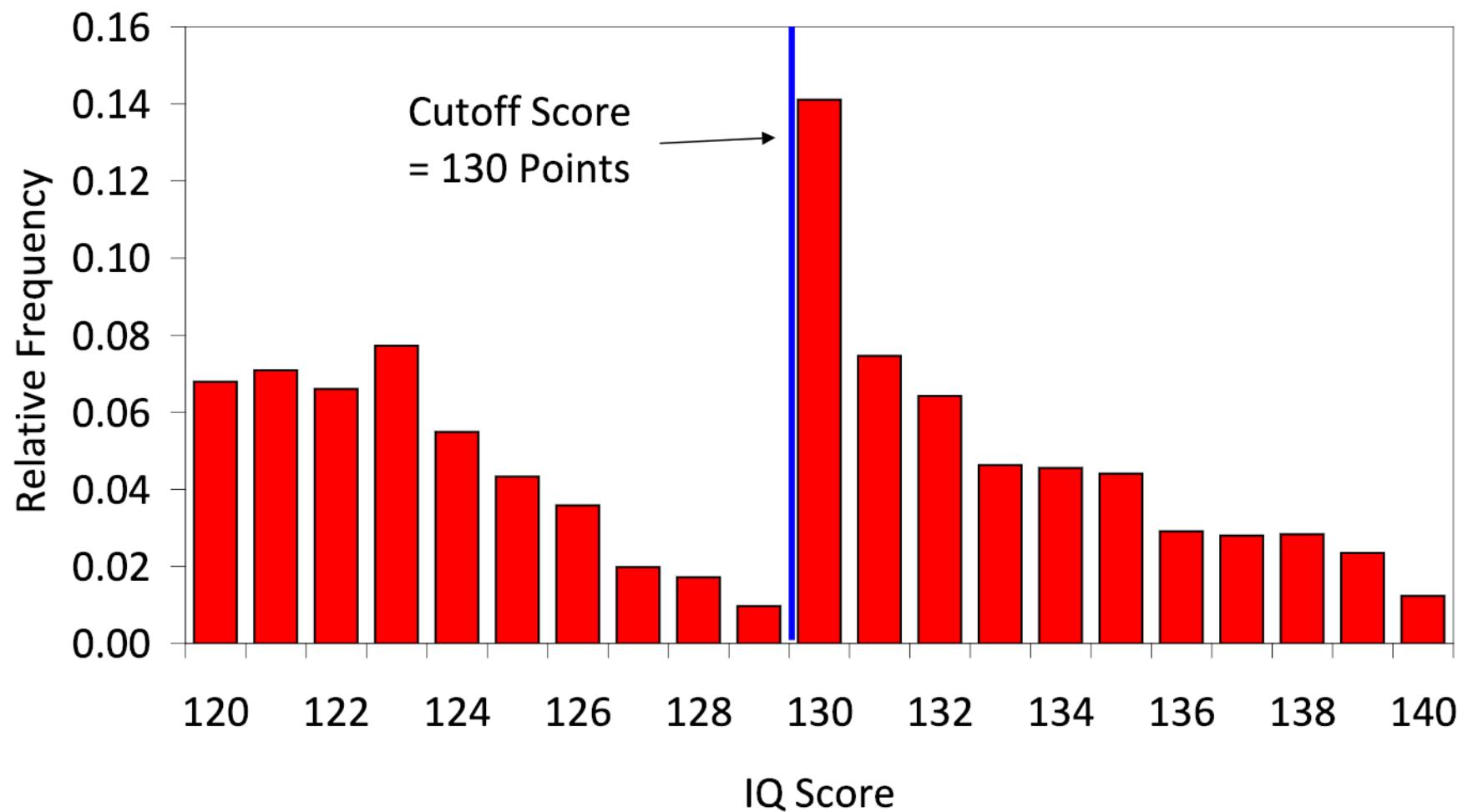
Think of the case where $D_i = 1$ is desirable (you win a scholarship if your test score is above a cutoff). If people can bribe the test-taker – or take the test again – we would expect to see:

- (a) differences in characteristics of people on either side of the cutoff
- (b) a “spike” of people just to the right

Intuitively, if assignment to the plus or minus side is as good as random, we can think of the people on either side as treatments and controls in a randomized experiment!

Researchers therefore look for evidence of “manipulation” of the running variable.

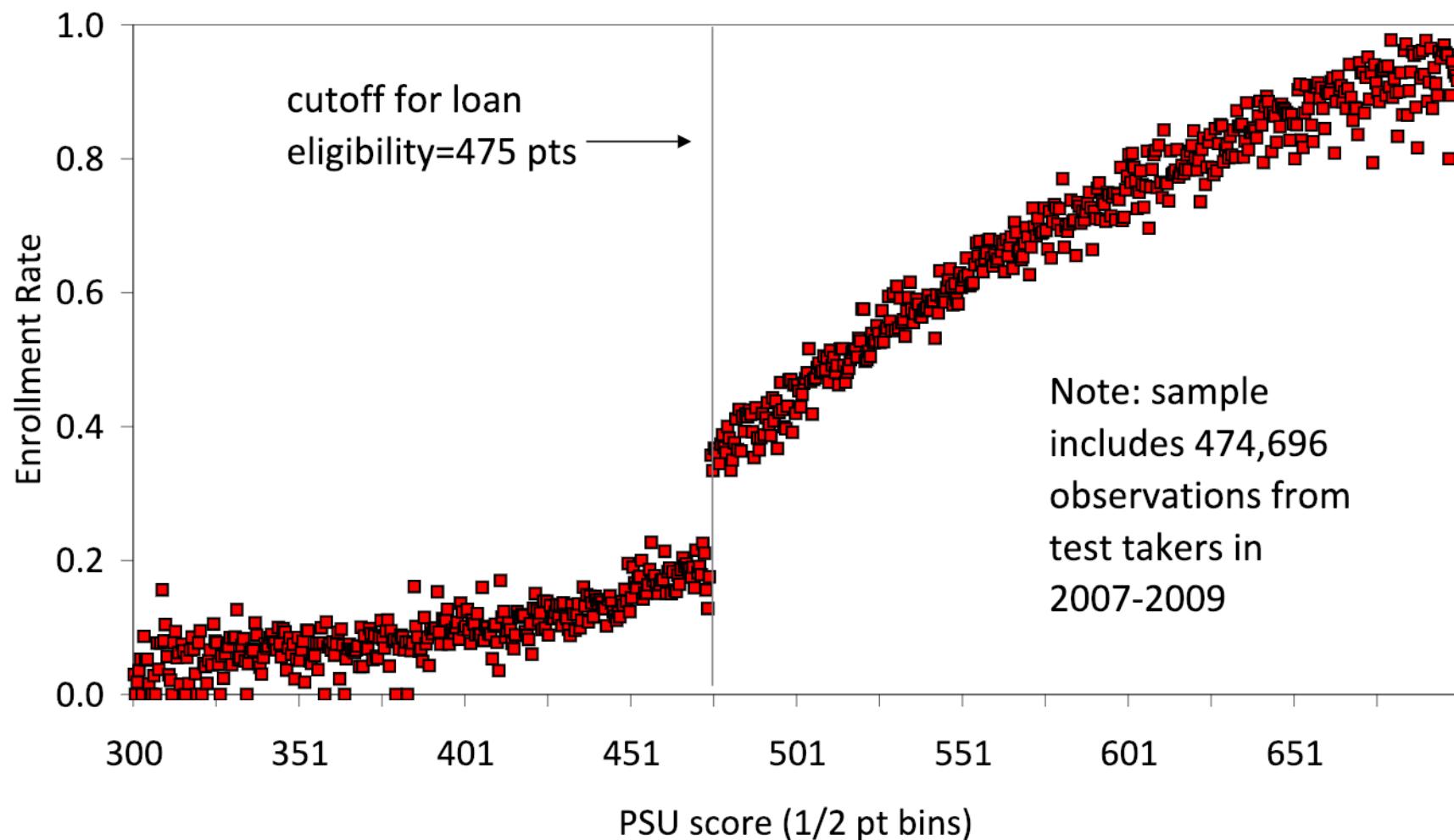
Distribution of IQ Scores - Children Tested for Gifted Status
(Excluding Free/Reduced Price Lunch and ELL Participants)



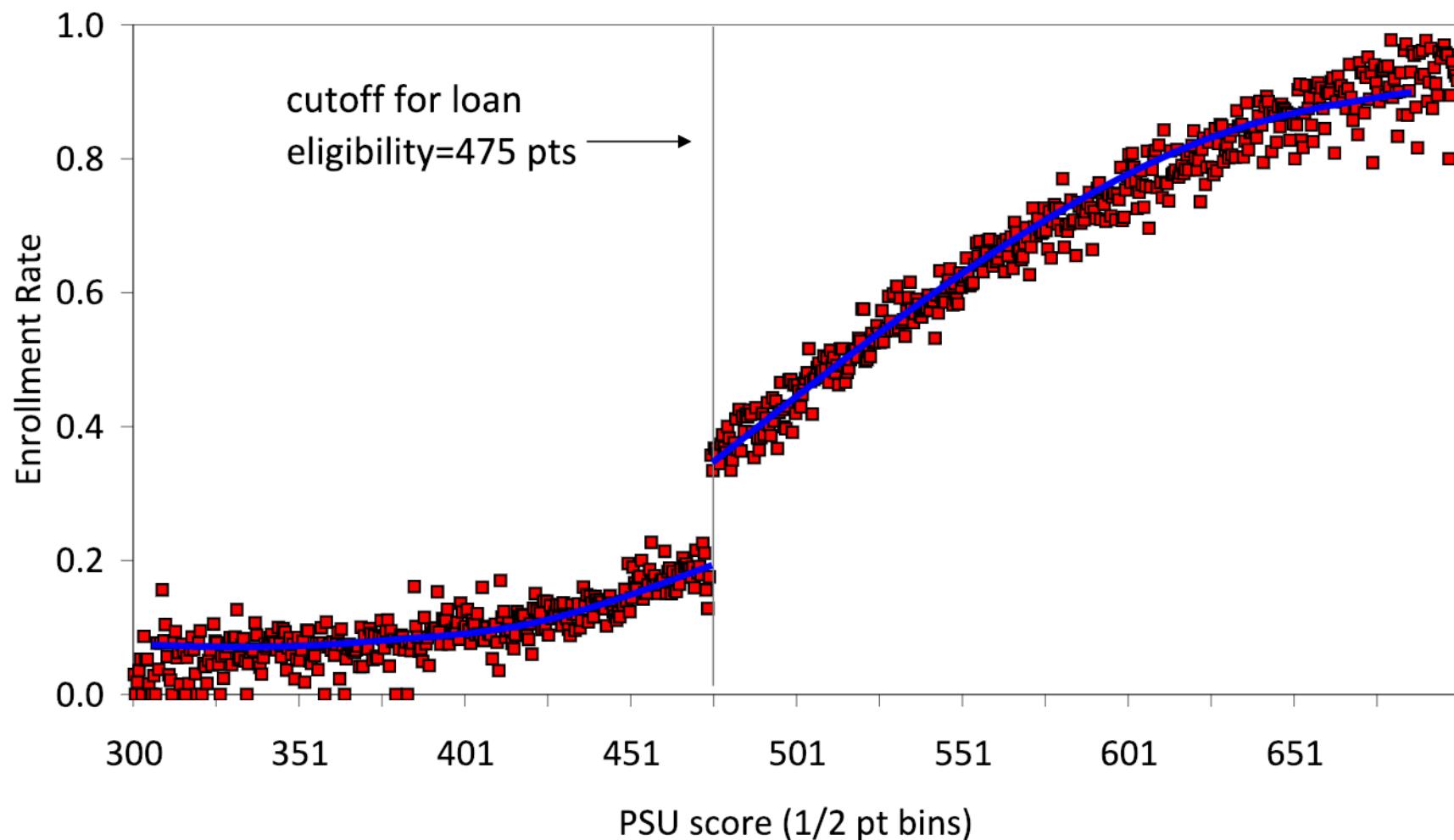
Example: “Credit Access and College Enrollment”, Alex Solis

- in Chile, students take “PSU” test (96% of HS grads take the test)
- colleges and universities use PSU + GPA to rank students and admit
- students with family taxable income < 75th percentile can use a loan program if they score at least 475 points (combined math and language) on PSU
- does access to loans matter? RD time!
- sample =475,000 students who wrote the test 2007-2009. Administrative data followup to measure enrollment in first year of college/university.

College Enrollment of Chilean Students and PSU Test Scores



College Enrollment of Chilean Students and PSU Test Scores



Fuzzy RD

In many cases, we don't have a deterministic assignment rule. Instead the probability of assignment to "treatment" depends on the running variable x and jumps discontinuously at $x = 0$. This leads to:

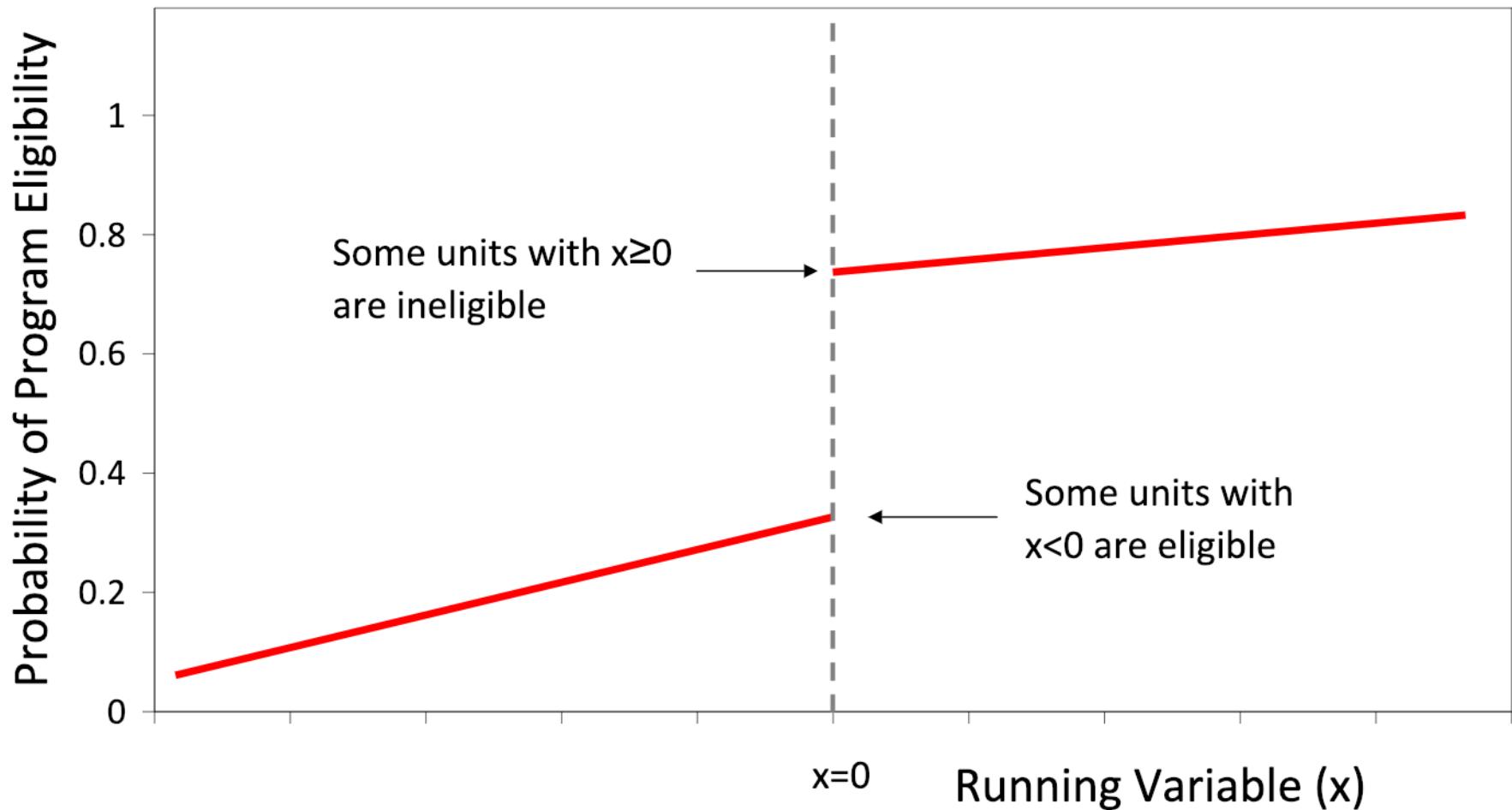
Fuzzy RD: $\Pr(D_i = 1|x_i) = g_0(x_i)$ for $x_i < 0$

and $\Pr(D_i = 1|x_i) = g_1(x_i)$ for $x_i \geq 0$ with $g_1(0) \neq g_0(0)$

Define $z_i = 1[x_i \geq 0]$. Then

$$\Pr(D_i = 1|x_i) = g_0(x_i) + (g_1(x_i) - g_0(x_i)) z_i$$

Fuzzy RD Based on Discontinuity in Program Eligibility



Key assumptions for fuzzy RD:

$E[Y_{0i}|x_i]$ and $E[Y_{1i}|x_i]$ are continuous functions of x at $x = 0$

$g_1(0) - g_0(0) \neq 0$, g_1 continuous from above, g_0 continuous from below

Define:

$$\begin{aligned}\delta_1 &= \lim_{x \rightarrow 0^+} E[y_i|x_i] - \lim_{x \rightarrow 0^-} E[y_i|x_i] \\ \pi_1 &= \lim_{x \rightarrow 0^+} E[D_i|x_i] - \lim_{x \rightarrow 0^-} E[D_i|x_i]\end{aligned}$$

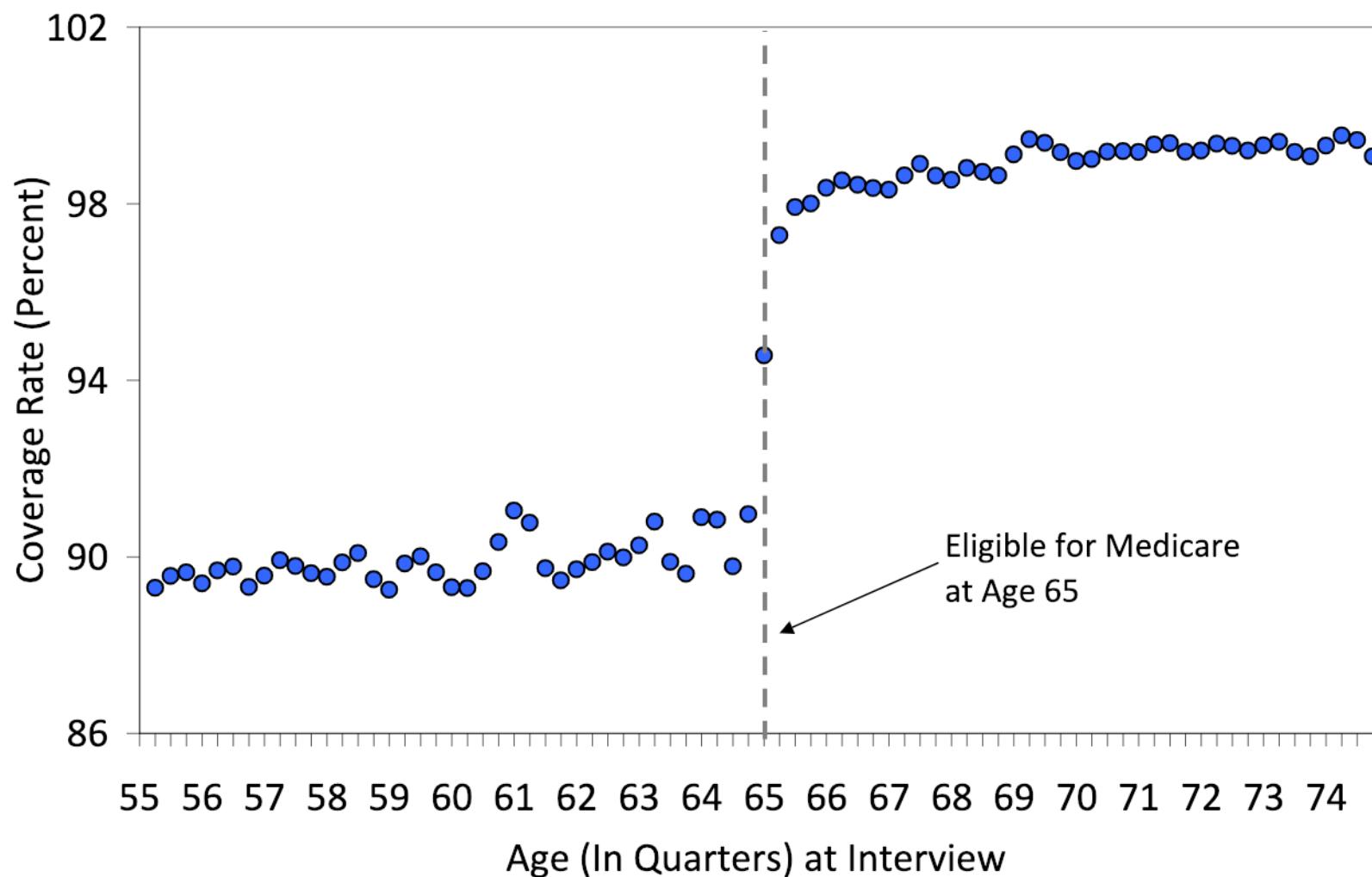
Then we estimate δ_1 and π_1 and form the ratio δ_1/π_1 . Note that for a sharp RD, $\pi_1 = 1$. Otherwise, $\pi_1 = g_1(0) - g_0(0)$

Example: effect of health insurance on use of health care services.

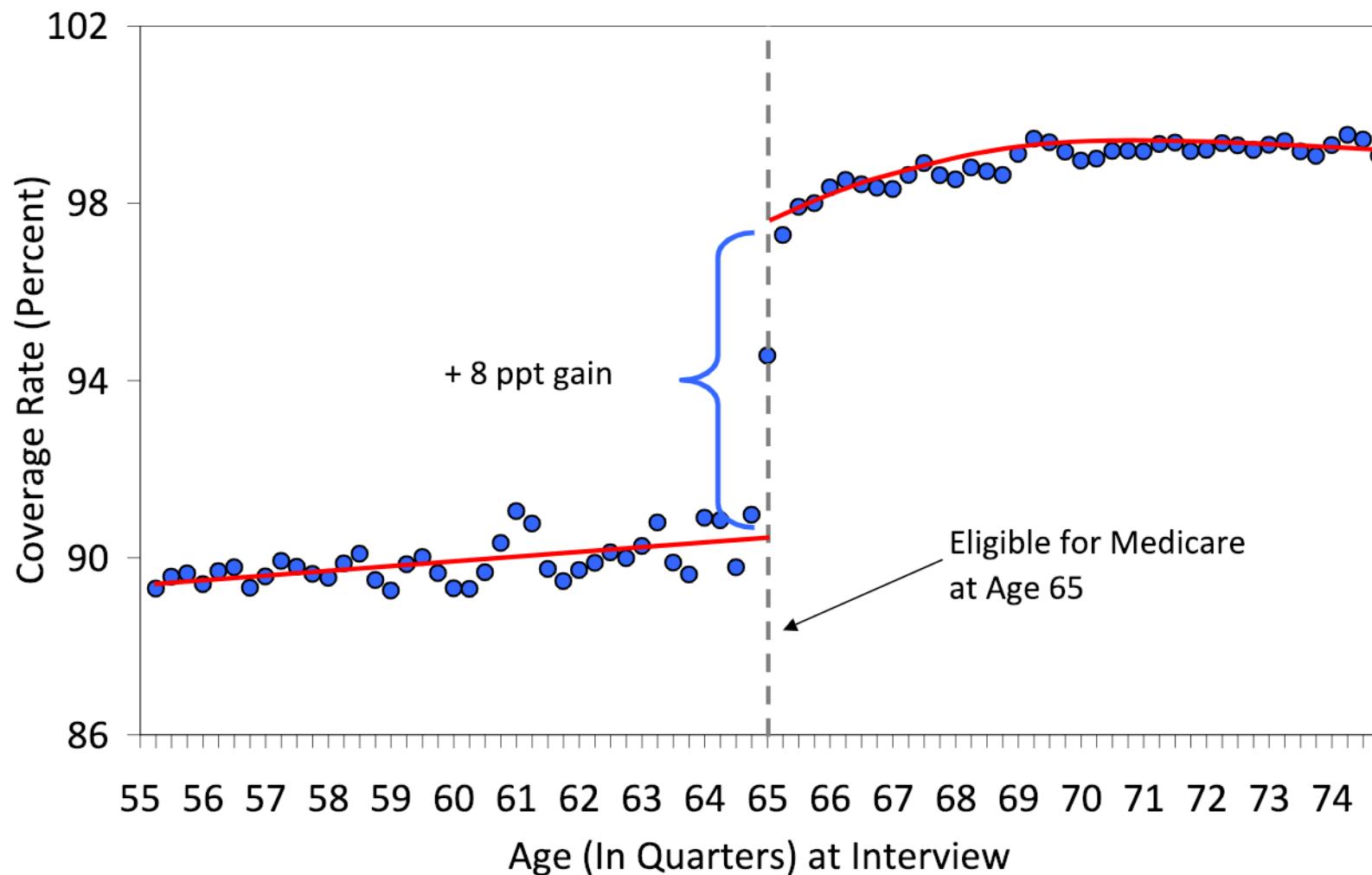
- most people become eligible for Medicare at age 65
- at that point, many people who previously lacked insurance get coverage
- Card, Dobkin Maestas study: what happens?

This is a fuzzy design because some people get on Medicare pre-65 (DI program) and some people over 65 are ineligible (or delay entering). Also, effect on insurance status is smaller than rise in Medicare enrollment.

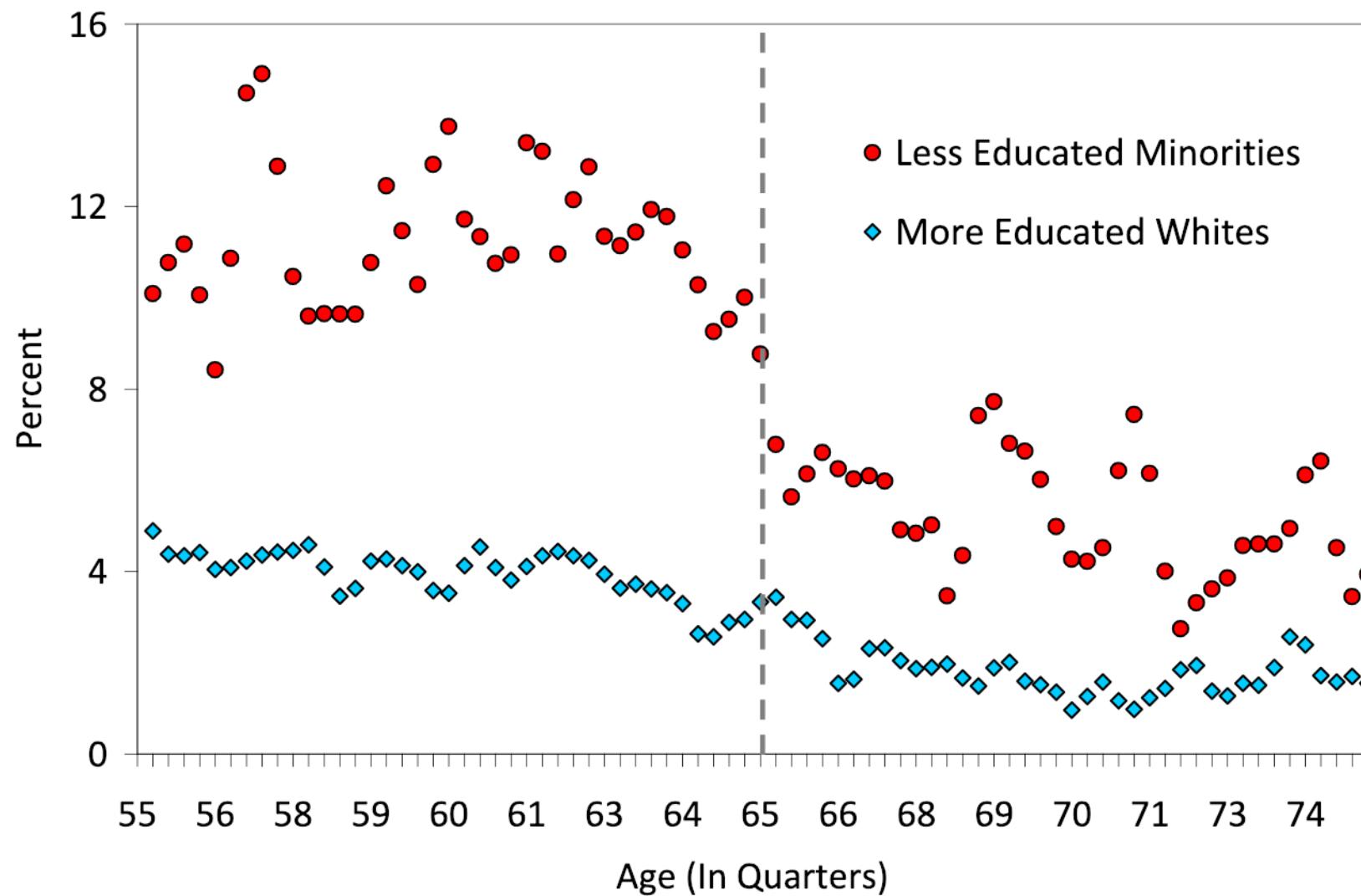
Figure 1: Health Insurance Coverage Rates by Age, HIS



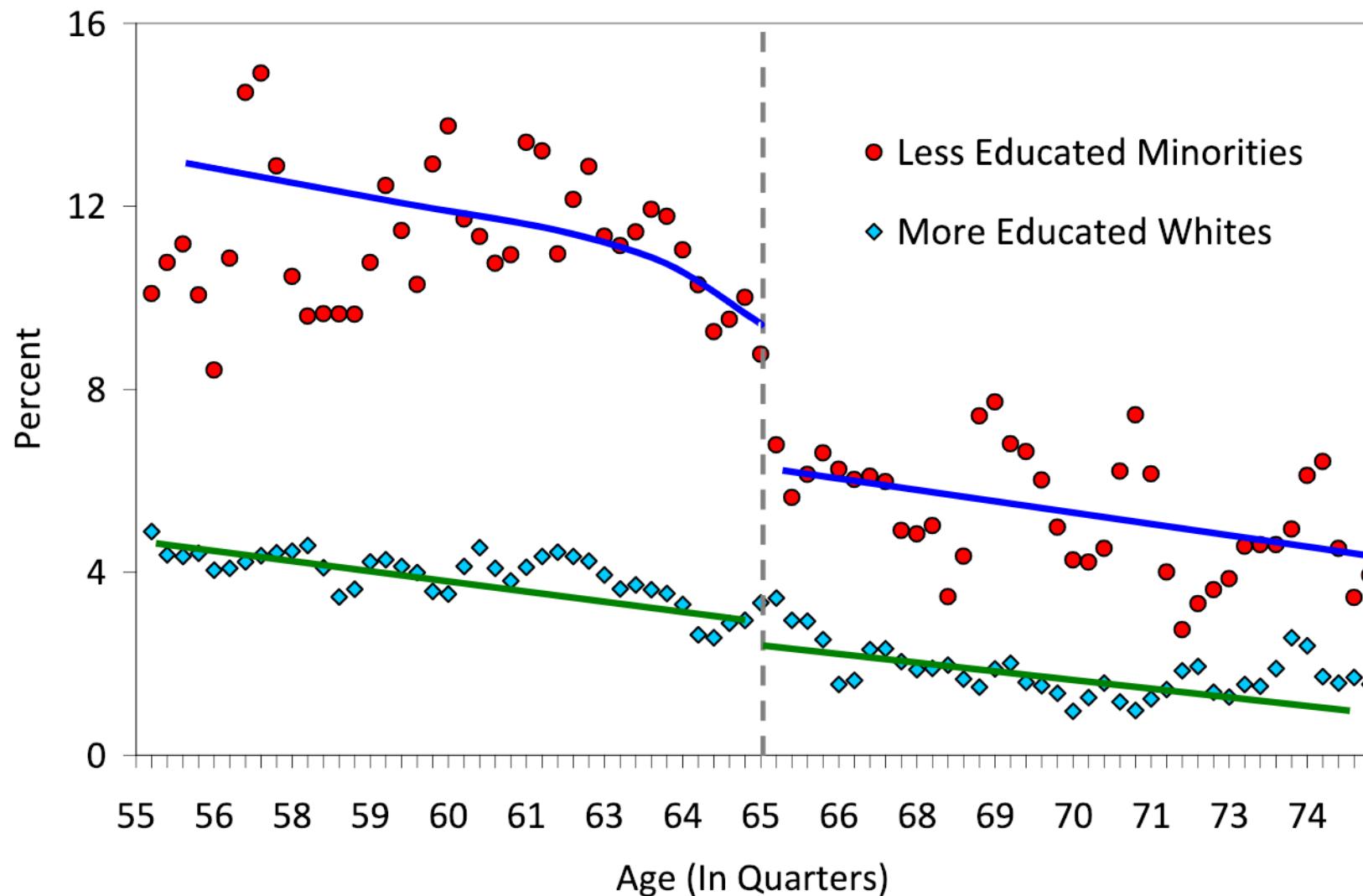
Health Insurance Coverage Rates by Age, HIS



Fraction Who Did Not Get Care for Cost Reasons by Age and Group



Fraction Who Did Not Get Care for Cost Reasons by Age and Group



Lecture 18: Regression Discontinuity continued

- Sharp RD (“full compliance”)
- fuzzy RD (“imperfect compliance”)
- implementation details

x_i = running variable (vote share, age, test score) *re-centered*

D_i = assignment status

Sharp RD: $D_i = 0 \text{ if } x_i < 0; \quad D_i = 1 \text{ if } x_i \geq 0$

potential outcomes: we see $y_i = Y_{0i}$ if $D_i = 0$; $y_i = Y_{1i}$ if $D_i = 1$

Key assumptions:

$E[Y_{0i}|x_i]$ is continuous function of x at $x = 0$

$E[Y_{1i}|x_i]$ is continuous function of x at $x = 0$

To left of cutoff we see $E[Y_{0i}|x_i]$

To right of cutoff we see $E[Y_{1i}|x_i]$

local treatment effect: $\beta_1 = E[Y_{1i}|x_i = 0] - E[Y_{0i}|x_i = 0]$

We estimate $E[Y_{0i}|x_i = 0]$ using data for $x < 0$

And we estimate $E[Y_{0i}|x_i = 0]$ using data for $x > 0$

Then we form $\hat{\beta}_1$ using the estimates at the “boundary” $x = 0$

Estimation: “Global” quadratic (old-fashioned)

Assume: $E[Y_{0i}|x] = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2$ (only see this for $x < 0$)

and $E[Y_{1i}|x] = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2$ (only see this for $x \geq 0$)

$$\beta_1 = \gamma_0 - \alpha_0 = E[Y_{1i}|x_i = 0] - E[Y_{0i}|x_i = 0]$$

Use “fully interacted” regression

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \beta_1 D_i + \alpha'_1 D_i x_i + \alpha'_2 D_i x_i^2 + u_i$$

where $\alpha'_1 = \gamma_1 - \alpha_1$, $\alpha'_2 = \gamma_2 - \alpha_2$. Fit to wider range of the data

Local linear (“frontier”):

$$E[Y_{0i}|x] = \alpha_0 + \alpha_1 x_i \text{ (only see this for } x < 0)$$

$$E[Y_{1i}|x] = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 \text{ (only see this for } x \geq 0)$$

$$\beta_1 = \gamma_0 - \alpha_0 = E[Y_{1i}|x_i = 0] - E[Y_{0i}|x_i = 0]$$

Implement this by fitting an “interacted” regression

$$y_i = \alpha_0 + \alpha_1 x_i + \beta_1 D_i + \alpha'_1 D_i x_i + u_i$$

where $\alpha'_1 = \gamma_1 - \alpha_1$.

Fit the model for $-h \leq x \leq h$ and examine $\hat{\beta}_1$. How to choose L ? Many recent papers.

Example: CDM (QJE, 2009) “Does Medicare Save Lives”

- look at mortality of patients admitted to hospital with “non-deferrable” conditions
- identify conditions as diagnoses with same admission rate week-end/weekday
- then look at mortality changes at age 65

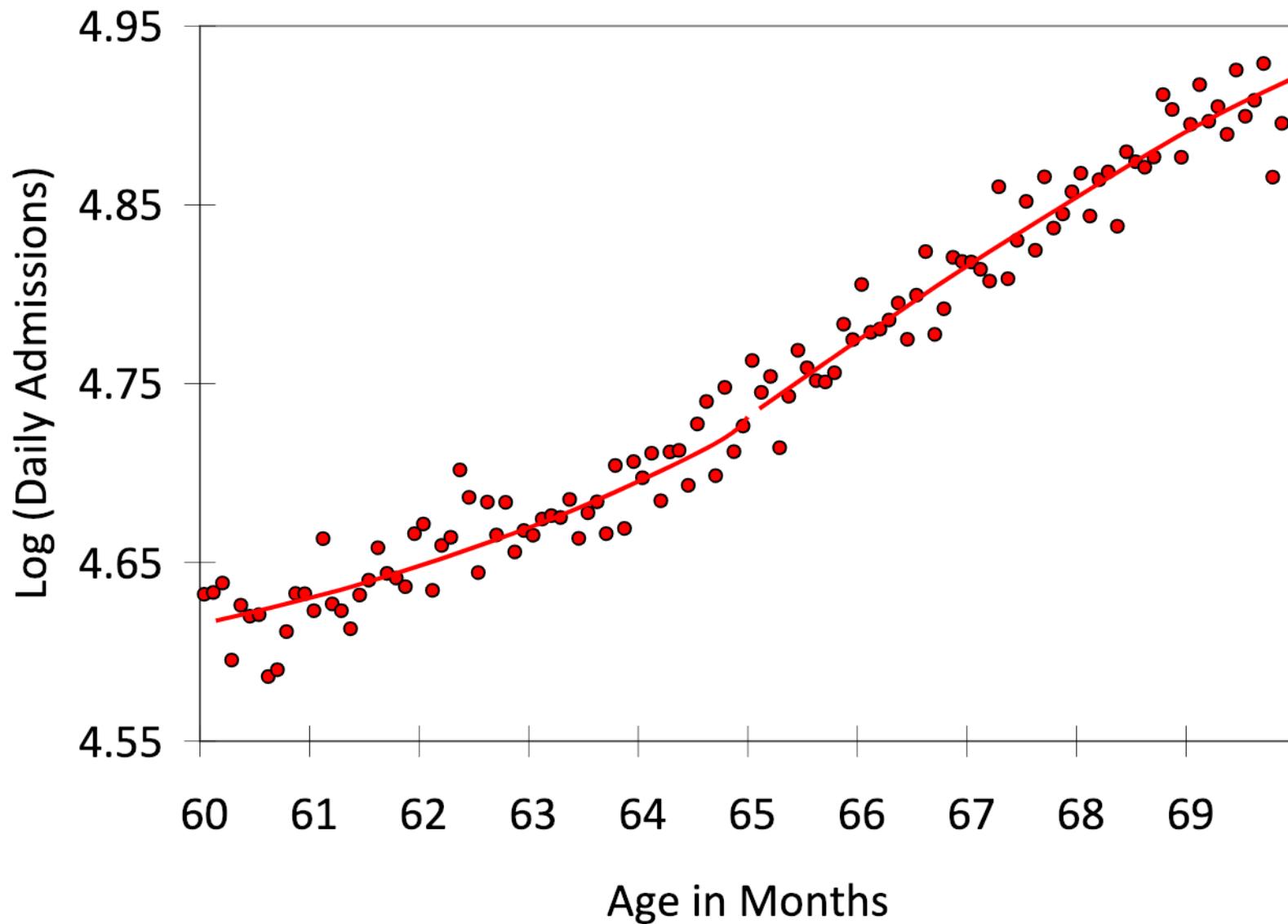
In Lecture 17 we discussed the idea that $E[Y_{0i}|x_i]$ and $E[Y_{1i}|x_i]$ will be continuous functions of x at $x = 0$ if there is an assignment process determining $x_i \geq 0$ or $x_i < 0$ that is “as good as random” for people close to $x = 0$. This condition has 2 implications:

1. the relative number of observations should trend smoothly at $x = 0$
2. if w_i is some predetermined characteristic (determined prior to assignment)

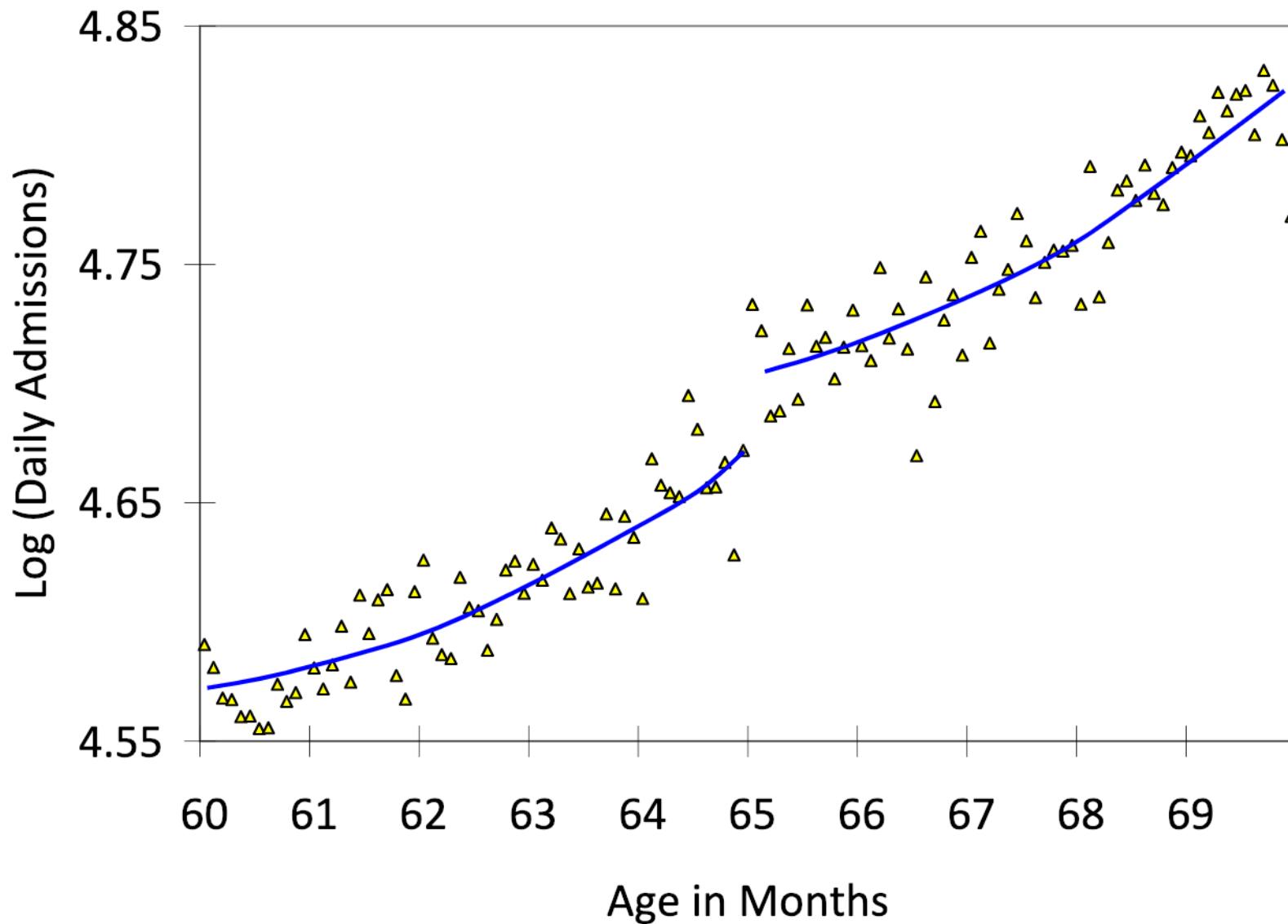
then $E[w_i|x_i]$ will be continuous at $x = 0$.

Let's look at numbers of admissions and “co-morbidity” scores

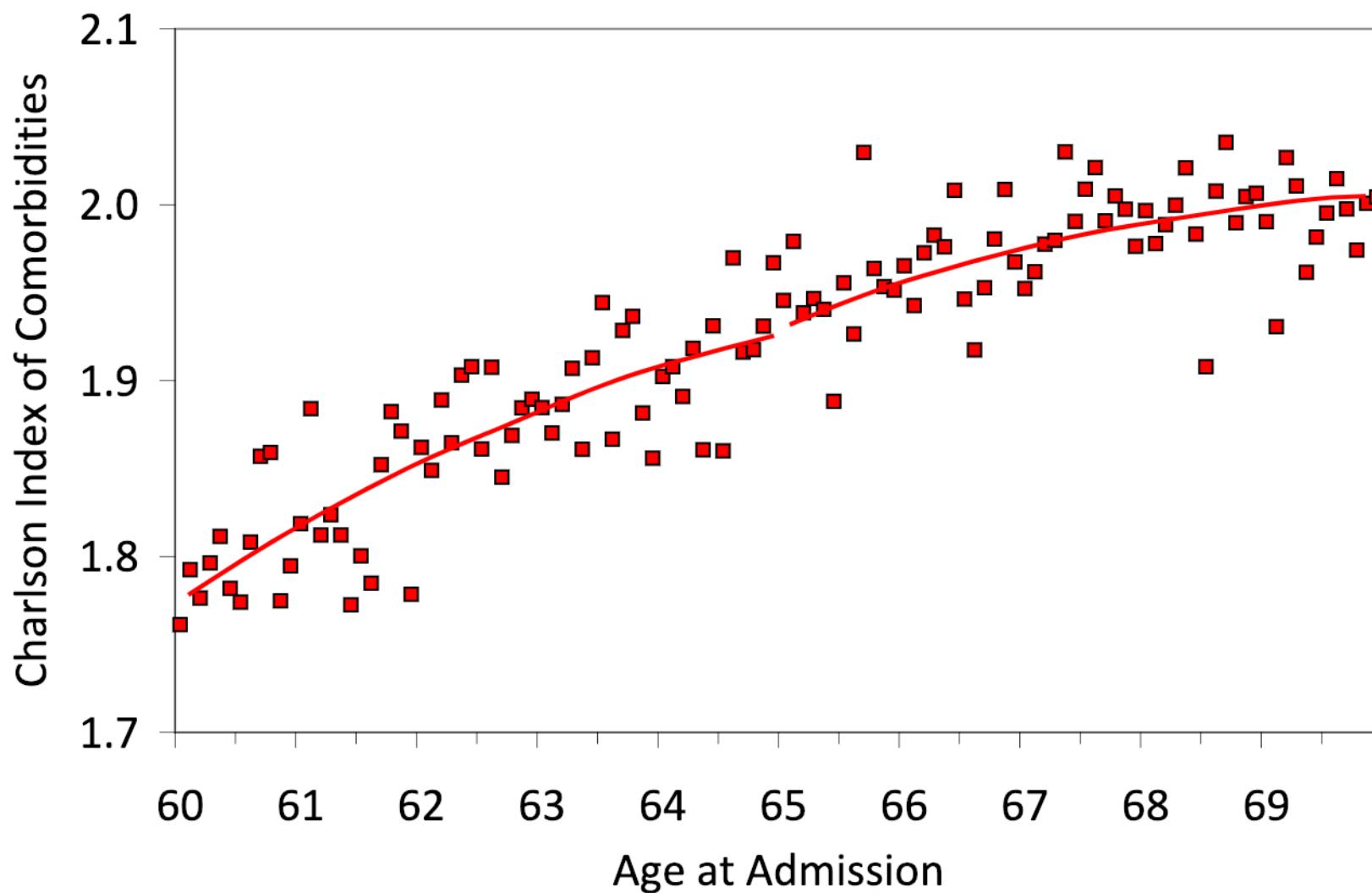
Daily Admissions through the ED: Non-Deferrable Conditions



Daily Admissions through the ED: Deferrable Conditions



Age Profile of Charlson Comorbidity Scores

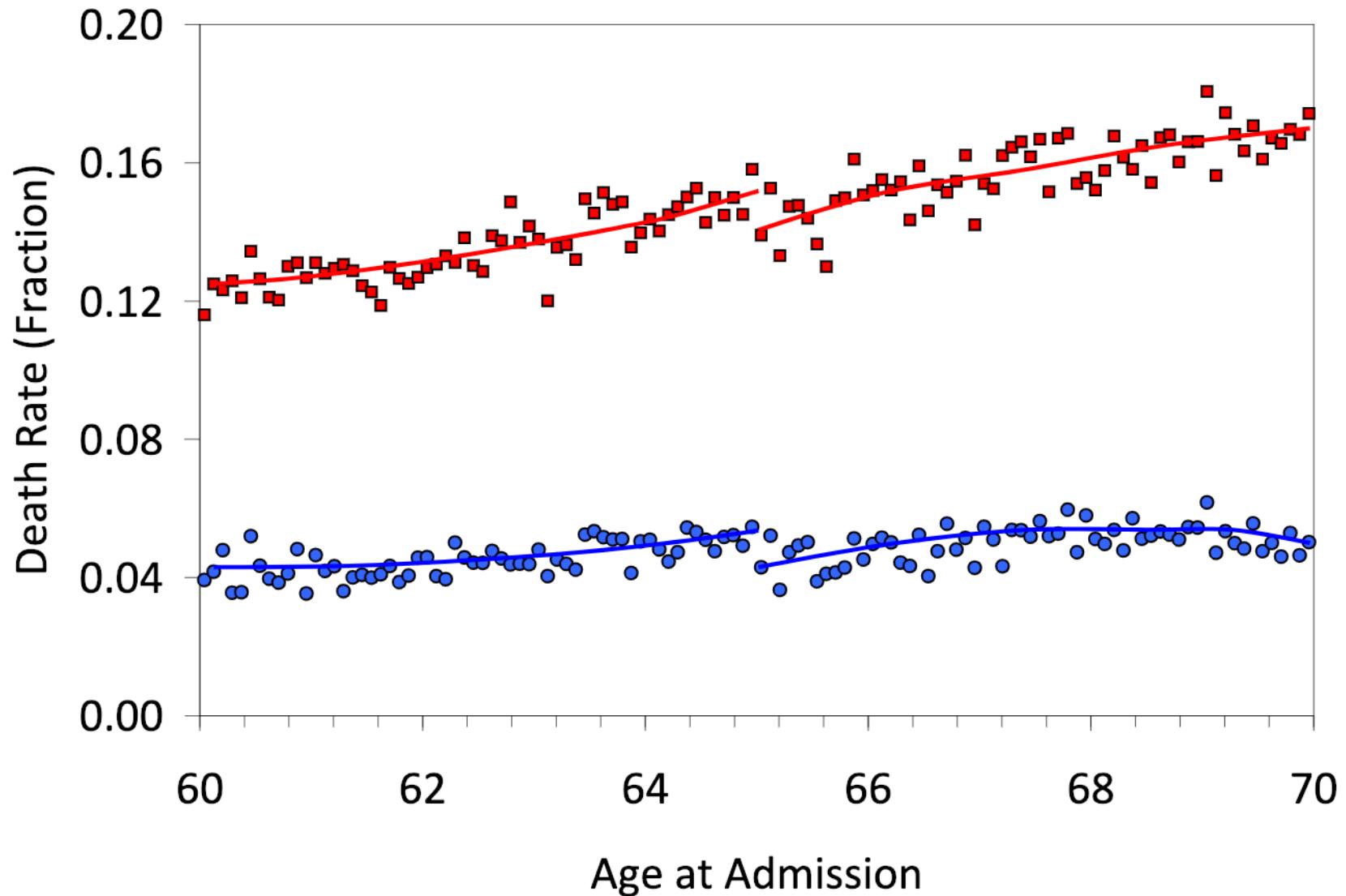


What about mortality?

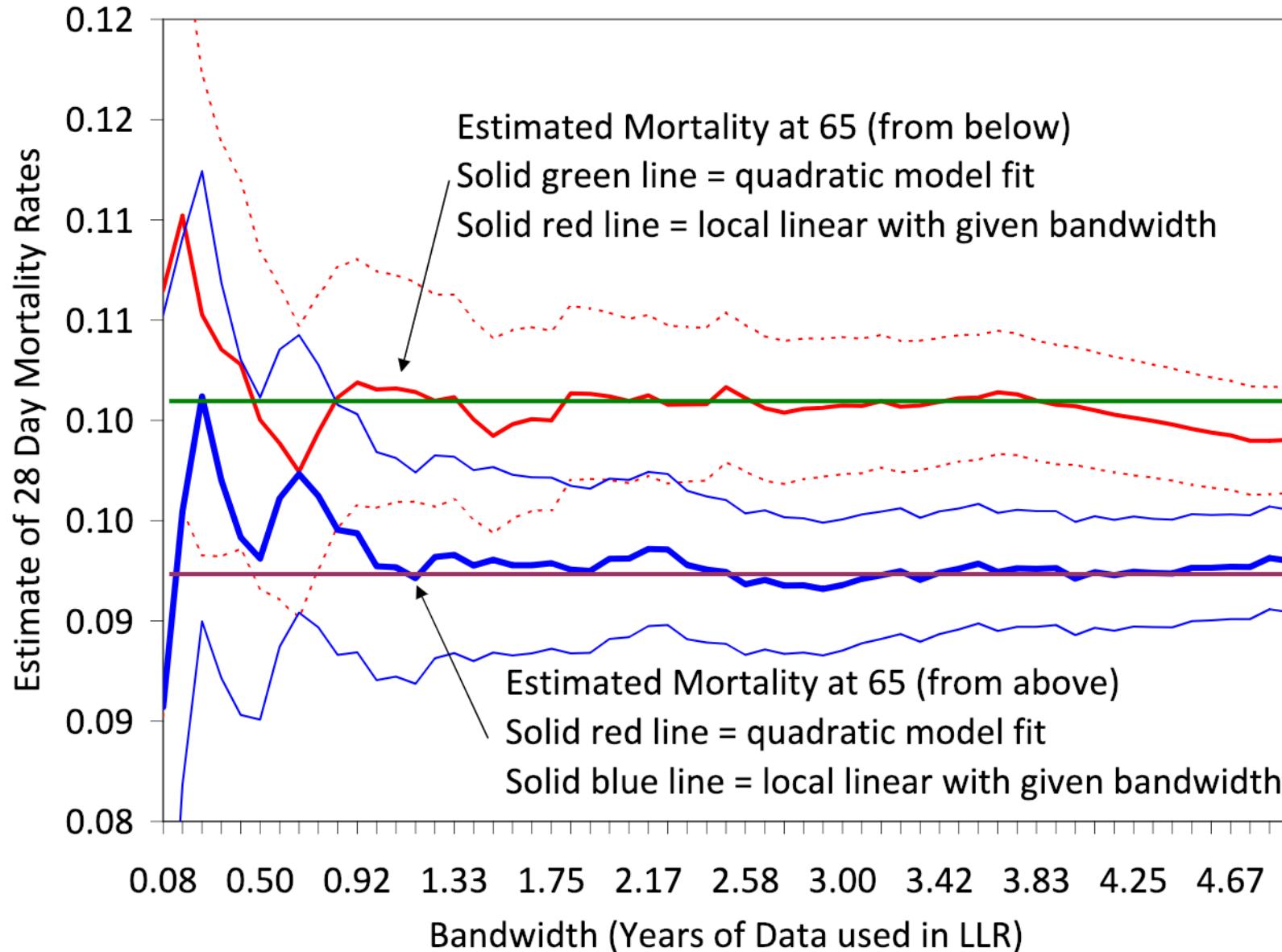
-global quadratic

-local linear

Patient Mortality Rates, 7 and 90 Days



Effect of Bandwidth Choices on 28 Day Mortality



Fuzzy RD: $\Pr(D_i = 1|x_i) = g_0(x_i)$ for $x_i < 0$

and $\Pr(D_i = 1|x_i) = g_1(x_i)$ for $x_i \geq 0$

Note that $1[x_i \geq 0]$ and D_i are now distinct

Key assumptions for fuzzy RD:

$E[Y_{0i}|x_i]$ and $E[Y_{1i}|x_i]$ both continuous at $x = 0$;

$g_1(x)$ continuous from right at $x = 0$

$g_0(x)$ continuous from left at $x = 0$

$g_1(x) \neq g_0(x)$.

Define:

$$\begin{aligned}\delta_1 &= \lim_{x \rightarrow 0^+} E[y_i|x_i] - \lim_{x \rightarrow 0^-} E[y_i|x_i] \\ \pi_1 &= \lim_{x \rightarrow 0^+} E[D_i|x_i] - \lim_{x \rightarrow 0^-} E[D_i|x_i]\end{aligned}$$

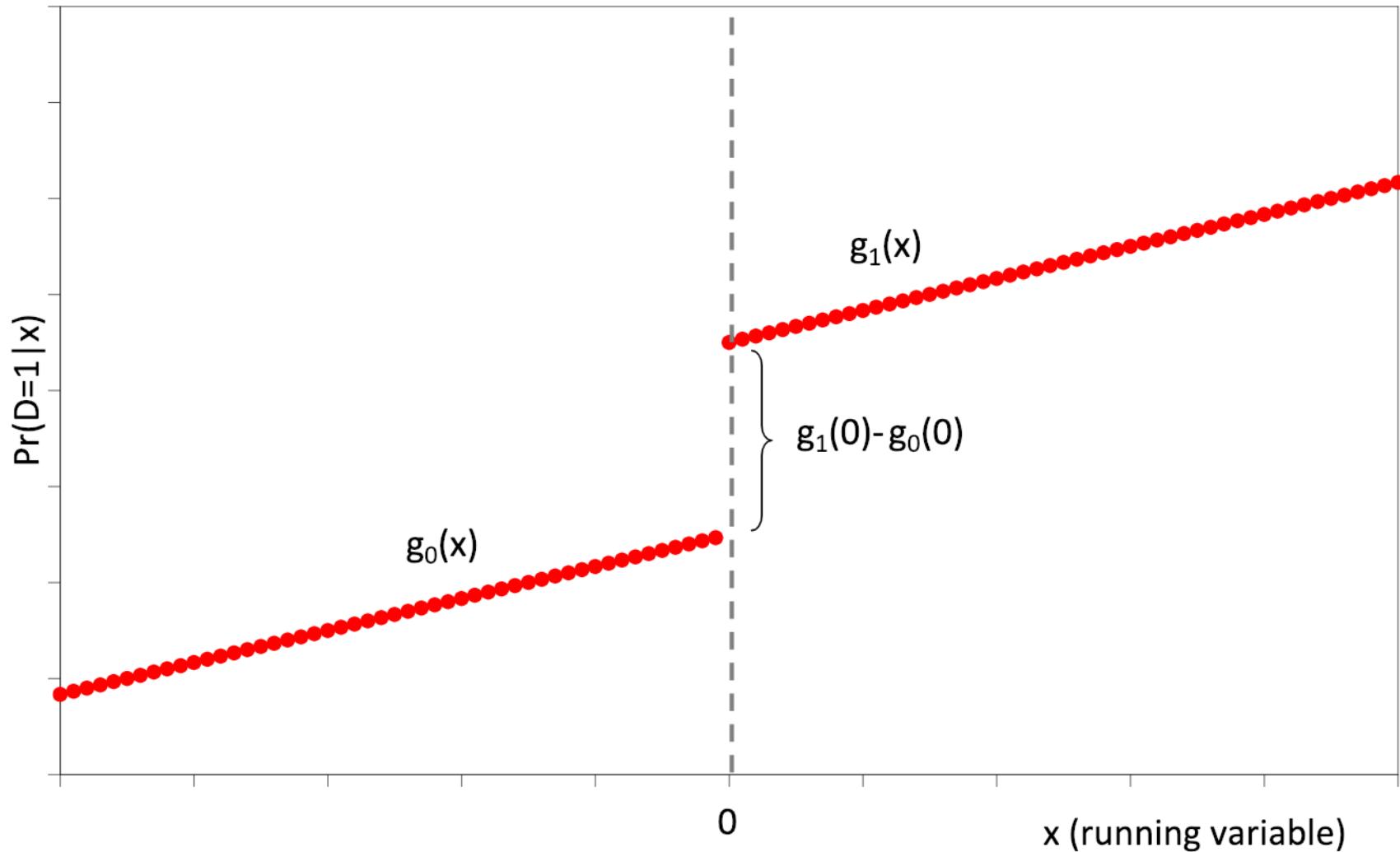
Then we estimate δ_1 and π_1 and form the ratio δ_1/π_1 .

The fuzzy RD estimator is the ratio of the jump in $E[y_i|x_i]$ at $x_i = 0$ to the corresponding jump in $Pr(D_i = 1|x_i)$.

Why does this work?

$$\begin{aligned}\pi_1 &= \lim_{x \rightarrow 0^+} E[D_i|x_i] - \lim_{x \rightarrow 0^-} E[D_i|x_i] \\ &= \lim_{x \rightarrow 0^+} Pr(D_i = 1|x_i) - \lim_{x \rightarrow 0^-} Pr(D_i = 1|x_i) \\ &= g_1(0) - g_0(0)\end{aligned}$$

First Stage Model



$$\delta_1 = \lim_{x \rightarrow 0^+} E[y_i|x_i] - \lim_{x \rightarrow 0^-} E[y_i|x_i]$$

Now: $E[y_i|x_i] = E[y_i|D_i = 0, x_i] \times Pr(D_i = 0|x_i)$

$$+ E[y_i|D_i = 1, x_i] \times Pr(D_i = 1|x_i)$$

For $x_i < 0$: $E[y_i|x_i] = E[Y_{0i}|x_i](1 - g_0(x_i)) + E[Y_{1i}|x_i]g_0(x_i)$

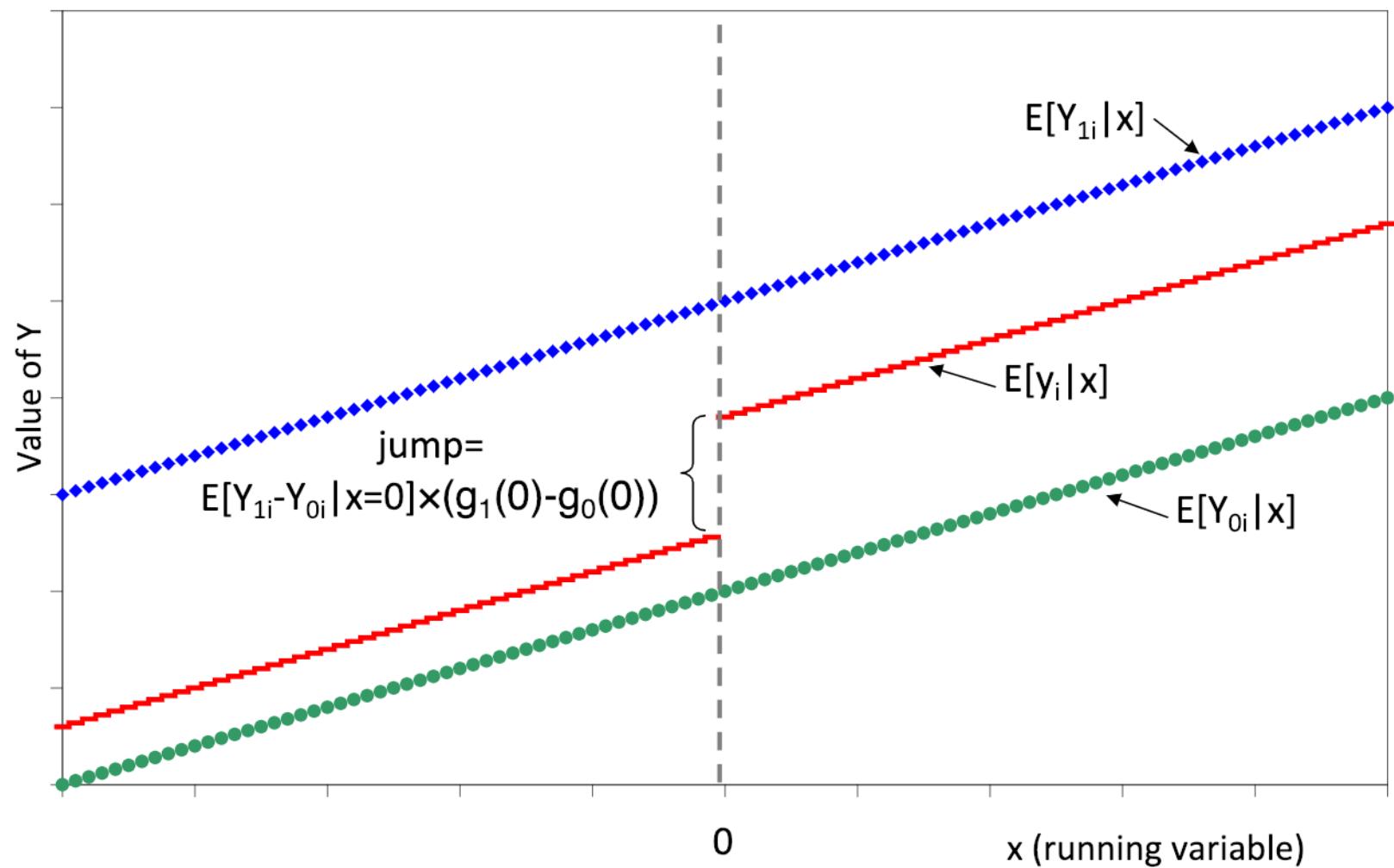
$$\Rightarrow \lim_{x \rightarrow 0^-} E[y_i|x_i] = E[Y_{0i}|x_i = 0](1 - g_0(0)) + E[Y_{1i}|x_i = 0]g_0(0)$$

For $x_i \geq 0$: $E[y_i|x_i] = E[Y_{0i}|x_i](1 - g_1(x_i)) + E[Y_{1i}|x_i]g_1(x_i)$

$$\Rightarrow \lim_{x \rightarrow 0^+} E[y_i|x_i] = E[Y_{0i}|x_i = 0](1 - g_1(0)) + E[Y_{1i}|x_i = 0]g_1(0)$$

$$\Rightarrow \delta_1 = (g_1(0) - g_0(0)) \times (E[Y_{1i}|x_i = 0] - E[Y_{0i}|x_i = 0])$$

Reduced Form Model



So we've shown

$$\begin{aligned}\pi_1 &= g_1(0) - g_0(0) \\ \delta_1 &= (g_1(0) - g_0(0)) \times (E[Y_{1i}|x_i = 0] - E[Y_{0i}|x_i = 0]) \\ \Rightarrow \delta_1/\pi_1 &= E[Y_{1i}|x_i = 0] - E[Y_{0i}|x_i = 0] \\ &= E[Y_{1i} - Y_{0i}|x_i = 0]\end{aligned}$$

Notice that β_1 is a “local” effect. If there are no defiers, it is the change for the “local compliers” who switch from $D = 0$ to $D = 1$ as x goes through 0. Their proportion in the overall population is $g_1(0) - g_0(0)$.

The ratio form of the fuzzy IV estimator suggests an IV approach. Define

$$z_i = 1[x_i \geq 0]$$

We are going to use z as an instrument for D . But unlike the “experimental” case, we have to include controls for the running variable.

Model for outcome - parametric quadratic (simplified):

$$E[Y_{0i}|x] = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2$$

$$E[Y_{1i}|x] = E[Y_{0i}|x] + \beta_1 ; \beta_1 = \text{tr eff.}$$

$y_i = Y_{0i} + D_i(Y_{1i} - Y_{0i})$. So our structural model is:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \beta_1 D_i + u_i$$

First stage model – parametric quadratic (simplified):

$$\begin{aligned} E[D_i|x_i] &= g_0(x_i) = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 \quad (x_i < 0) \\ &= g_1(x_i) = g_0(x_i) + \pi_1 \quad (x_i \geq 0) \end{aligned}$$

So first stage model is:

$$D_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \pi_1 z_i + \eta_i$$

where $\pi_1 = g_1(0) - g_0(0)$

$$\begin{aligned}
 y_i &= \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \beta_1 D_i + u_i \\
 D_i &= \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \pi_1 z_i + \eta_i \\
 \Rightarrow y_i &= \psi_0 + \psi_1 x_i + \psi_2 x_i^2 + \delta_1 z_i + e_i
 \end{aligned}$$

In the reduced form model, $\delta_1 = \beta_1 \pi_1$. So we can estimate the first stage model and the reduced form and take the ratio of the coefficients on z_i . Or we can do a 2SLS model where we instrument D_i with z_i . Notice that the structural model, the first stage and the reduced form all have a quadratic in the running variable included.

Example: Card-Giuliano study of gifted education program.

-separate classroom in 4th grade, 5th grade

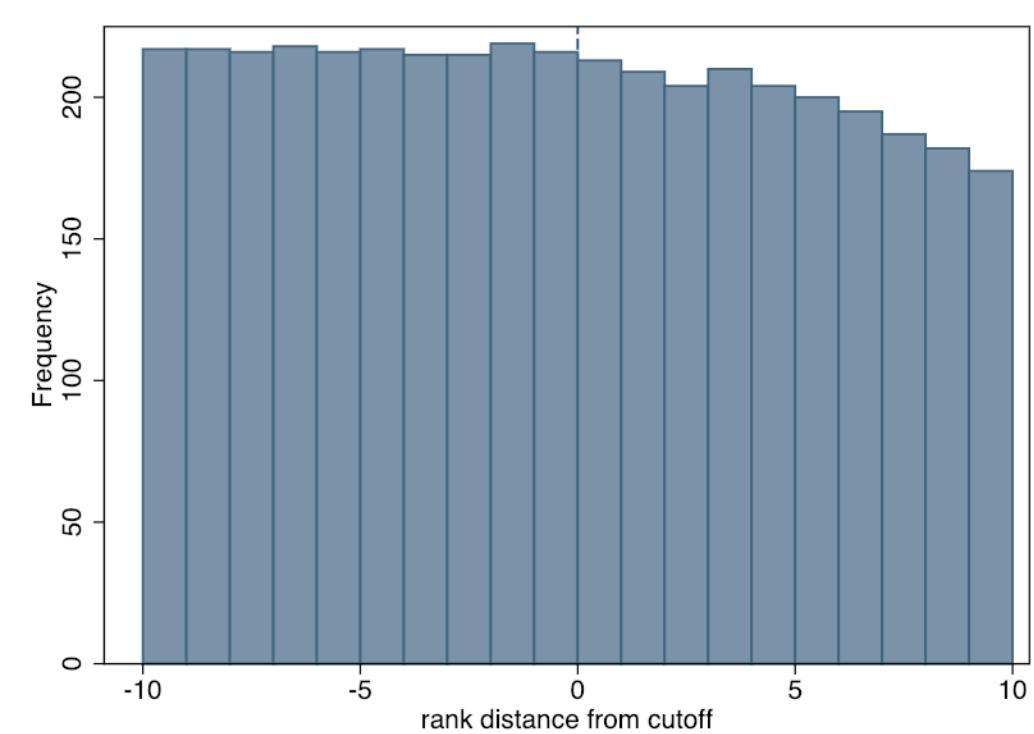
-3 groups in the class (21-22 seats):

non-free lunch kids with $\text{IQ} \geq 130$ = “Plan A”

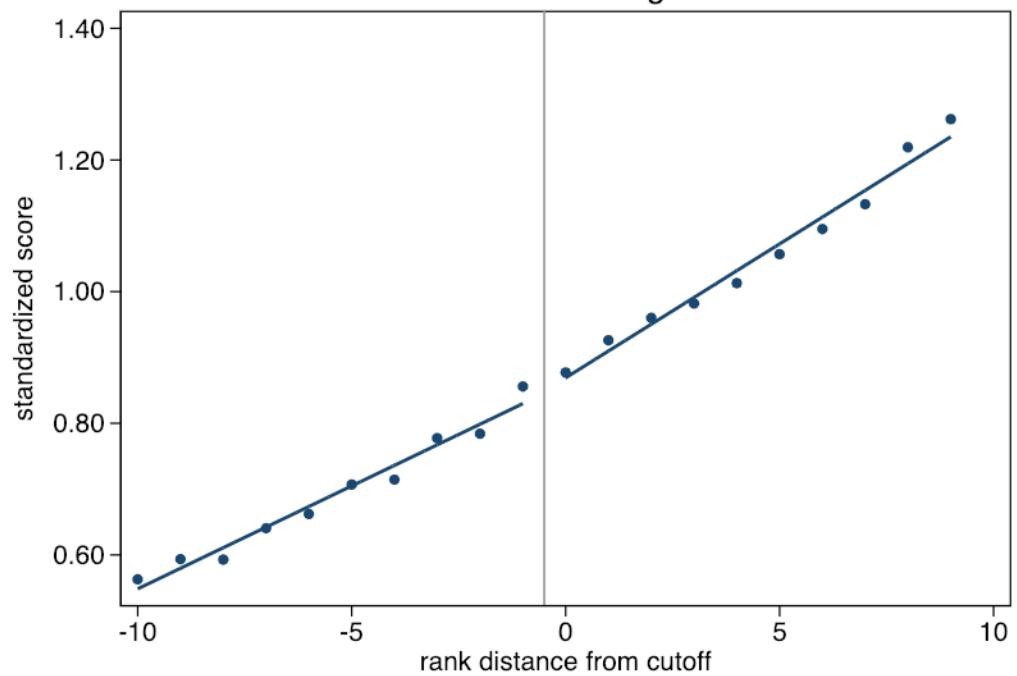
free lunch kids with $\text{IQ} \geq 116$ = “Plan B”

all other seats → top scorers on last year’s tests (high achievers)

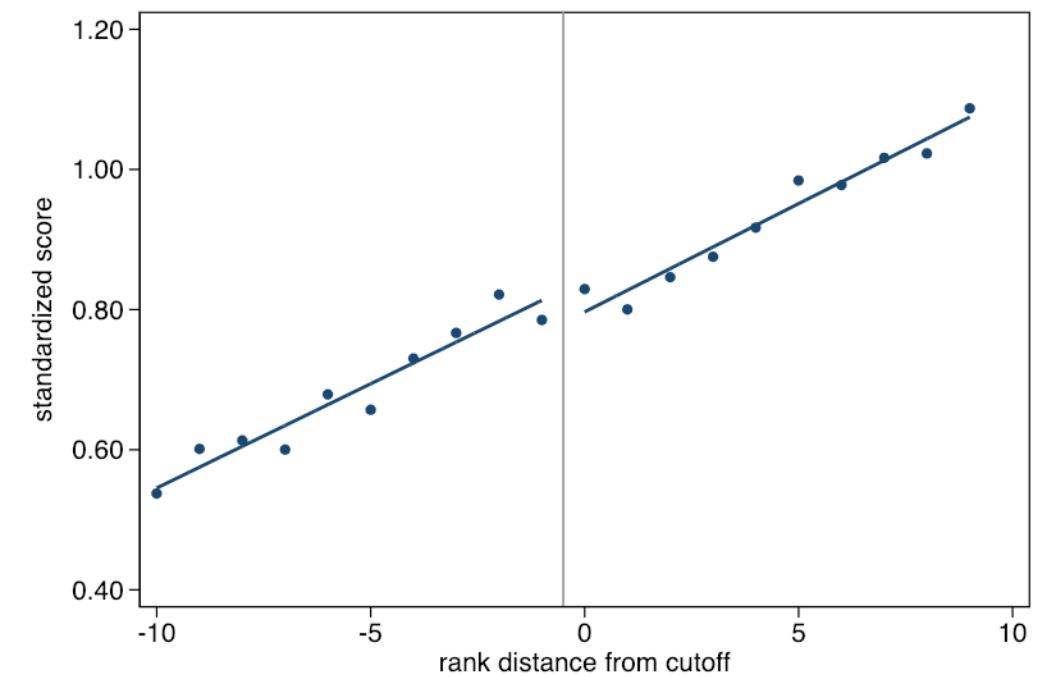
-RD analysis: use test score ranks and # open seats for HA (e.g., if there are 12 open seats then students with ranks 1-12 get in, students with ranks 13-24 do not



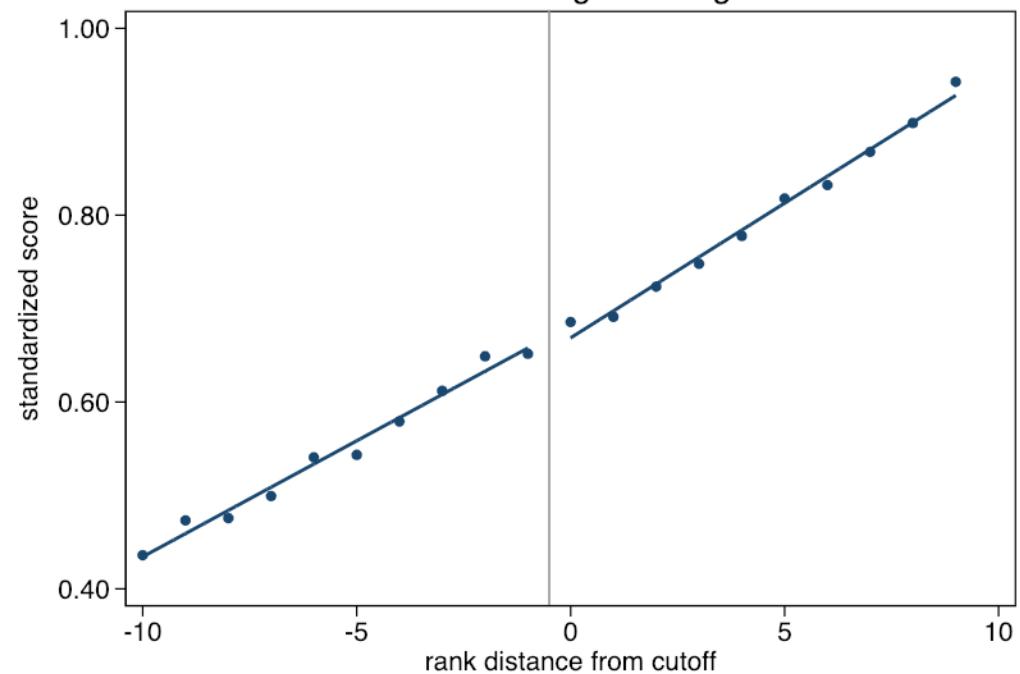
A. Baseline reading score



B. Baseline math score



D. Predicted average reading & math



A. Fraction placed
in GHA classroom

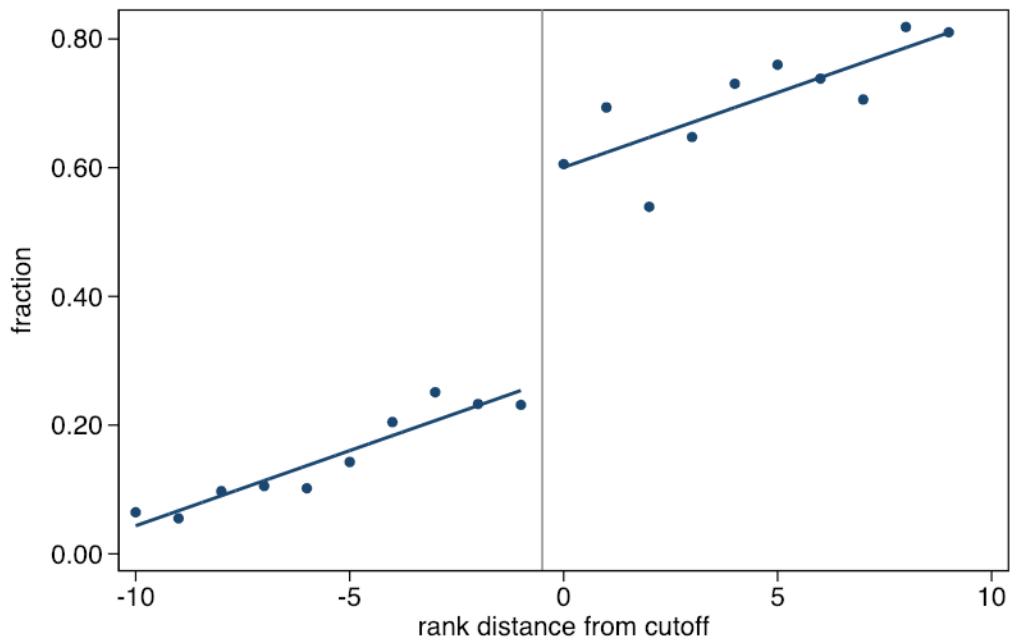
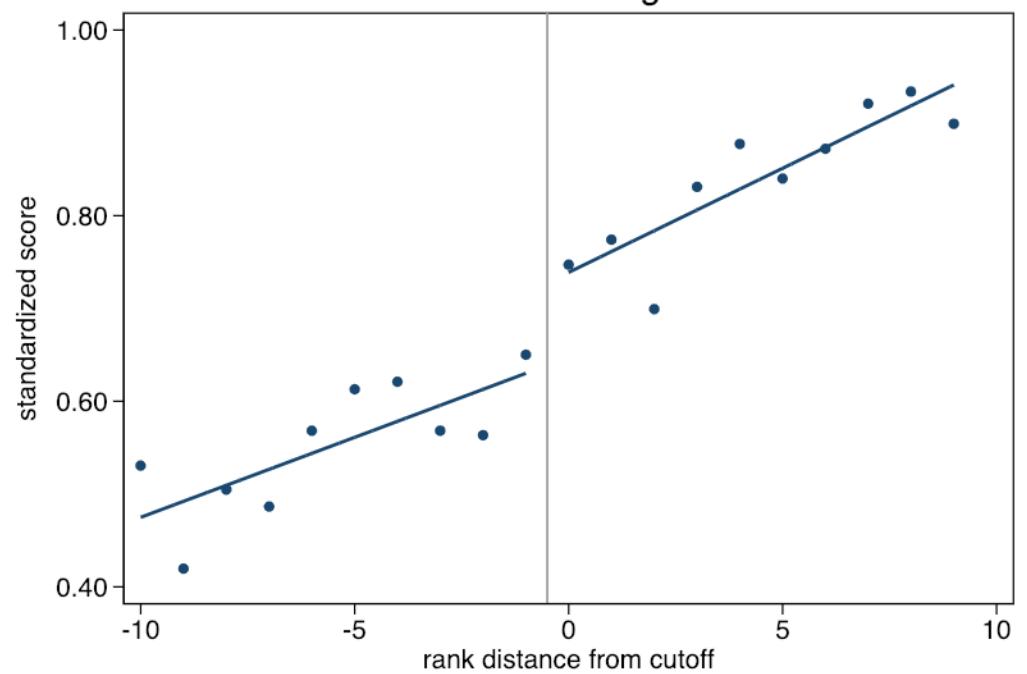


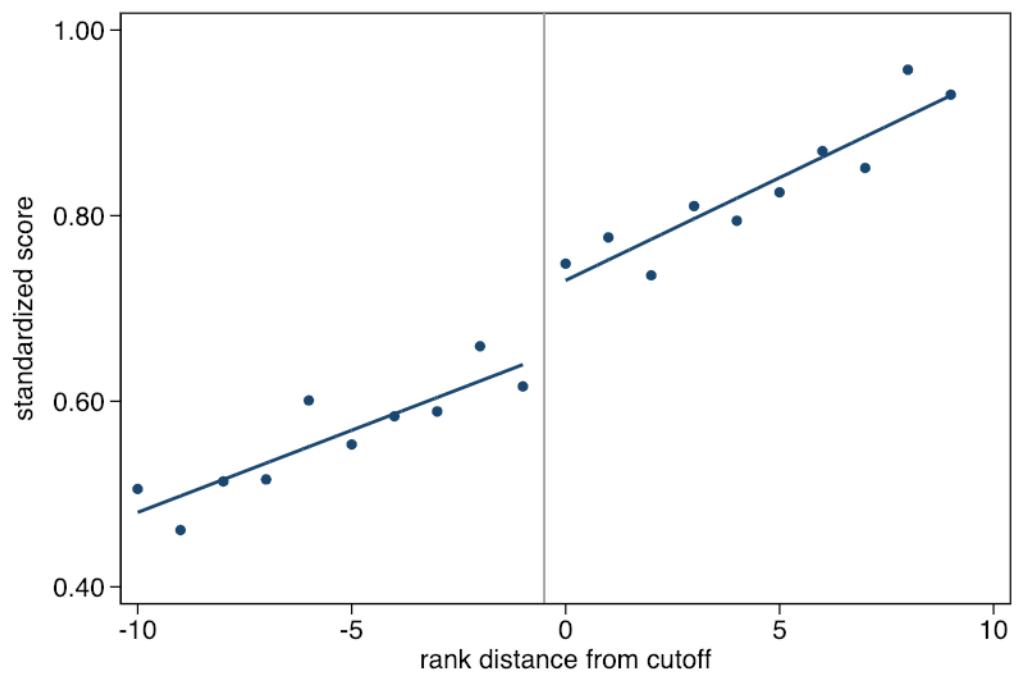
Table 2. Regression Discontinuity Estimates: Baseline Scores and 1st Stage

	Baseline Scores		First-Stage Models		
	Reading	Math	In GHA Class	Pred. Peer Scores	Frac. Peers Suspended
	(1)	(2)	(3)	(4)	(5)
1. No controls	0.008 (0.029)	-0.046 (0.043)	0.323** (0.025)	0.262** (0.027)	-0.009** (0.003)
2. School+year effects, lagged scores (col 3-8)	0.015 (0.027)	-0.044 (0.040)	0.319** (0.026)	0.265** (0.028)	-0.009** (0.003)

A. Reading



B. Math



C. Writing

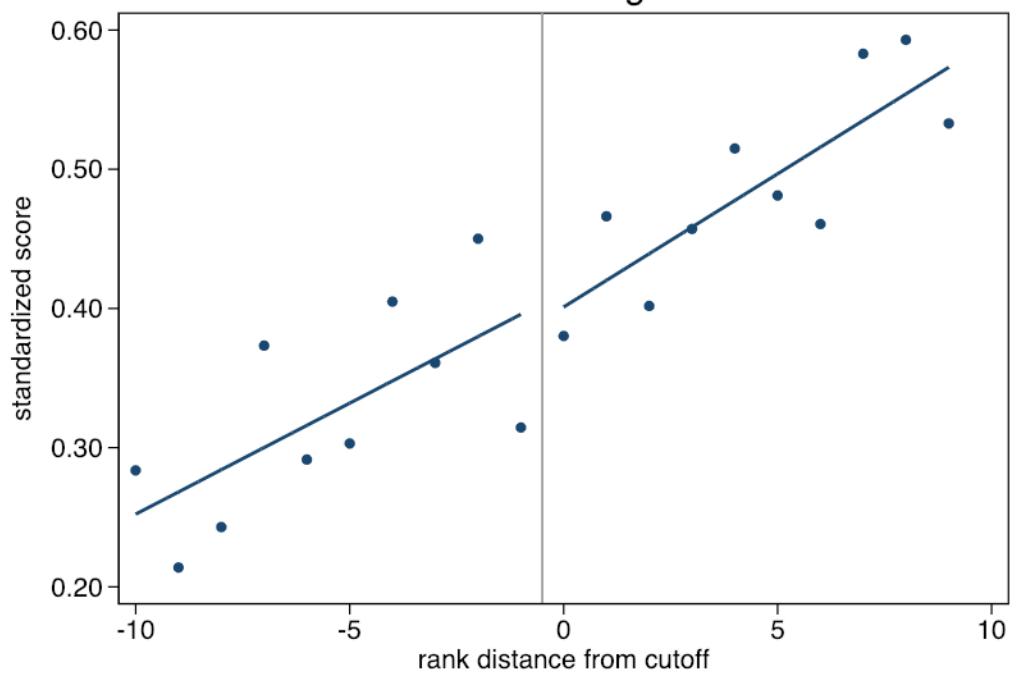
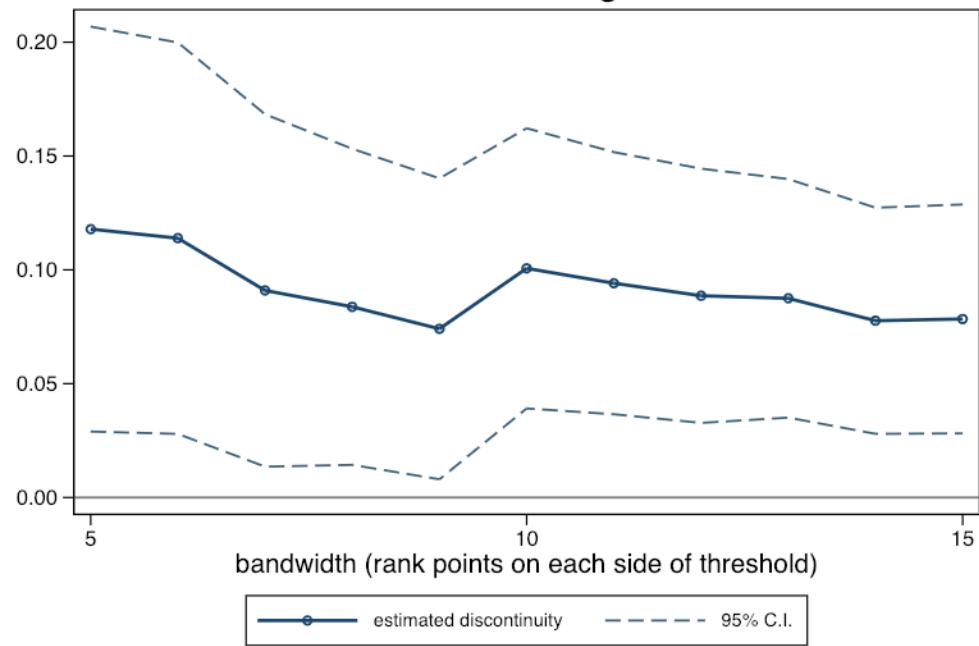


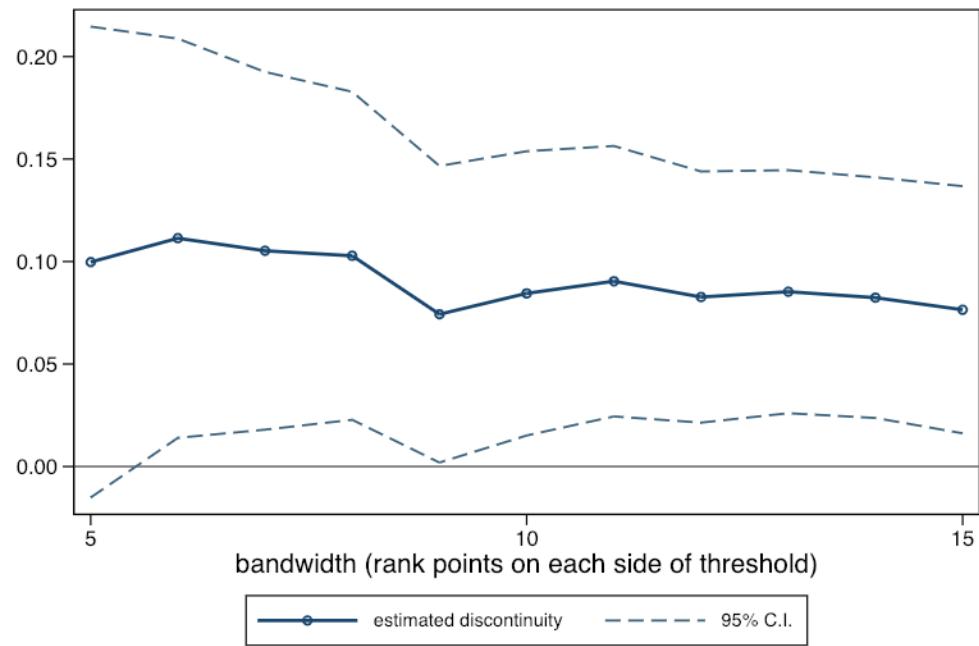
Table 2. Regression Discontinuity Estimates: Outcomes

	Reduced-Form Outcomes (4th Grade Scores)		
	Reading (6)	Math (7)	Writing (8)
1. No controls	0.092** (0.034)	0.073† (0.039)	-0.011 (0.054)
2. School+year effects, lagged scores (col 3-8)	0.093** (0.031)	0.087* (0.035)	-0.012 (0.051)
3. Differenced (change in test scores)	0.092** (0.033)	0.105* (0.041)	--
<i>Sample size</i>	4,144	4,144	4,144

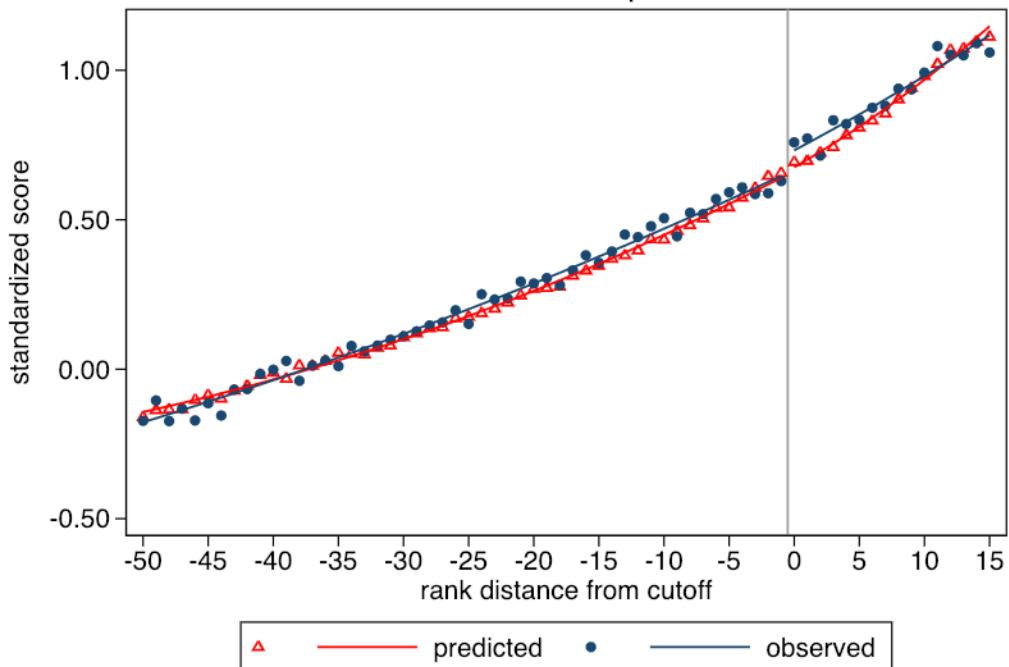
A. Reading



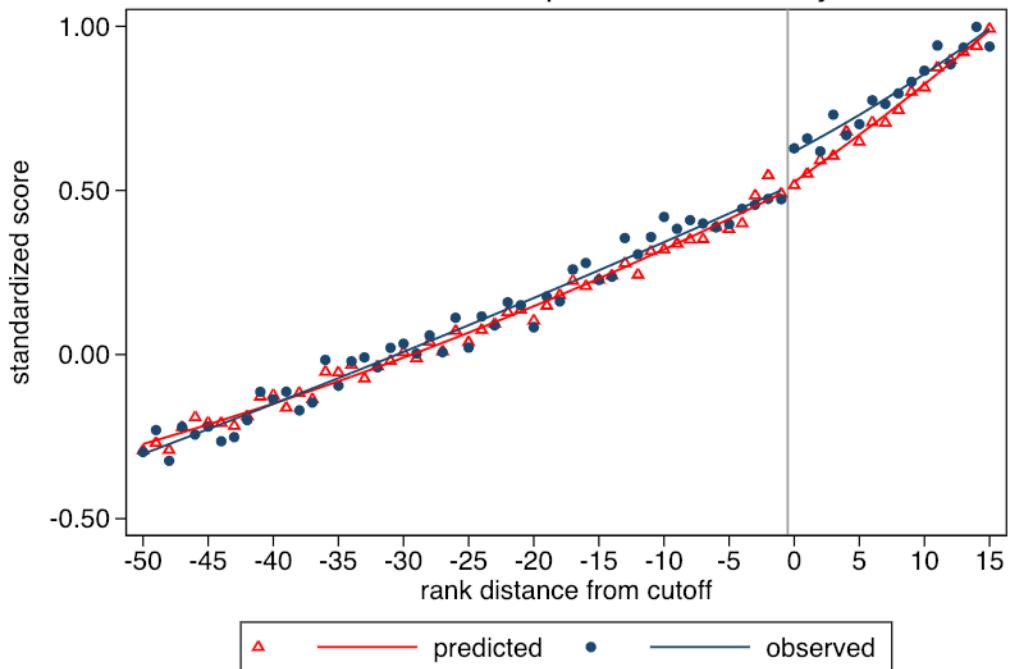
B. Math



A. Full sample



B. Black and Hispanic students only



Introduction to “Statistical Learning”

1. prediction problem: 1-dim. response / classifications
2. model selection, overfitting, regularization
3. training and test samples
4. resampling methods - k-fold

background: ISLR – chapter 2

- *Statistical Learning* is a set of techniques for estimating relationships between *inputs* (or predictors) and *outputs* (or responses)
- the focus is on forming *good predictors* of the output, rather than on evaluating a particular causal model
- two main cases:
 - 1-dimensional responses (income, health)
 - categorical responses or “classification” (product/location choice)

1-dimensional responses

- y = response; x =predictors (implicit “ i ” subscript)
- $f(x) = E[y|x]$ is the benchmark
- can write $y = f(x) + \epsilon$ where $E[\epsilon|x] = 0$ and $E[\epsilon h(x)] = 0$ for all $h(\cdot)$
- $f(x) = E[y|x]$ minimizes $E[(y - g(x))^2]$
- in practice we have to work with predictor $\hat{f}(x)$
- \hat{f} incorporates 2 things: choice of function; estimation error

- e.g. \hat{f} = polynomial of order p :

$$y = b_0 + b_1x + b_2x^2 + \dots + b_px^p + v$$

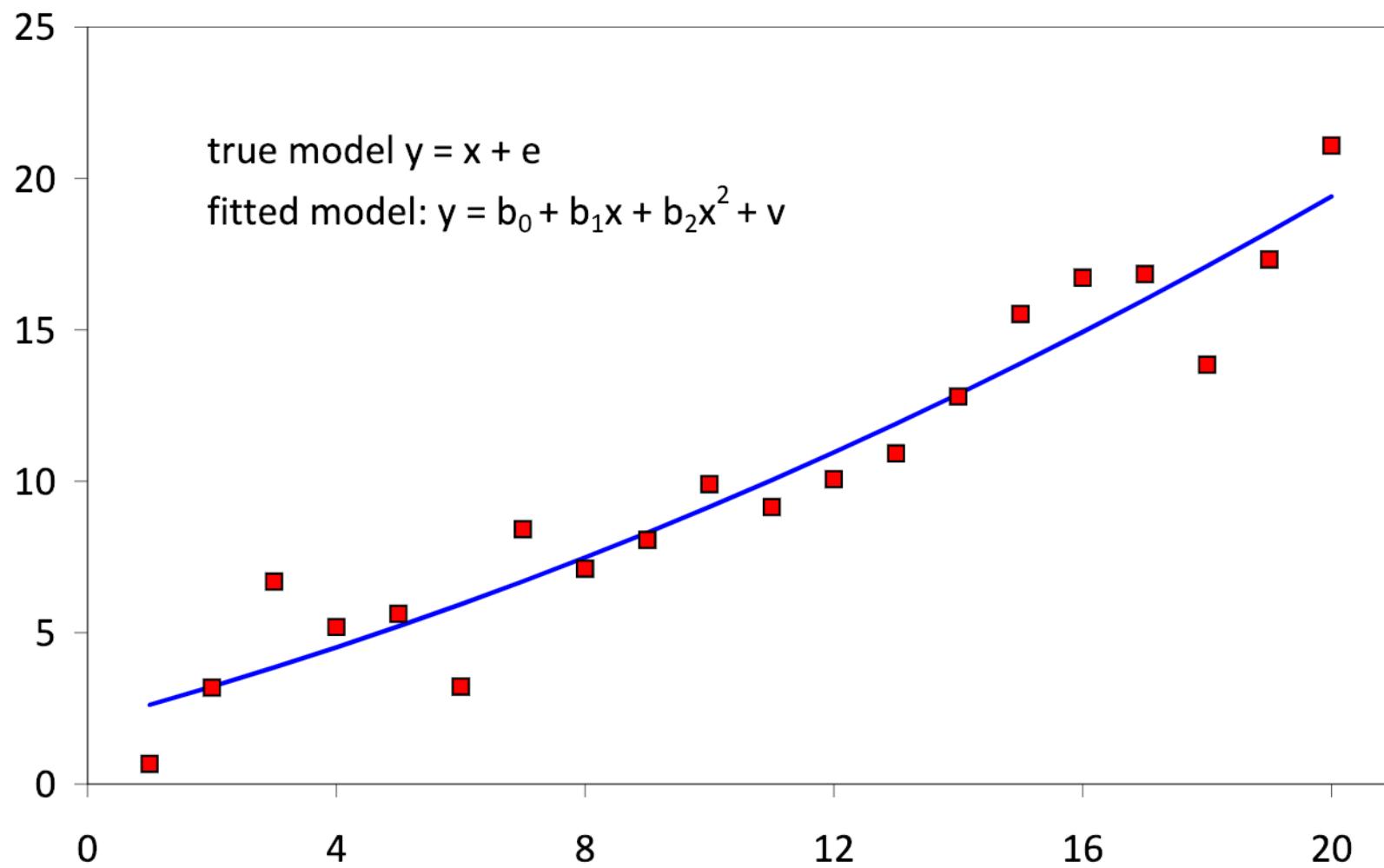
$$\hat{f}^p(x) = \hat{b}_0 + \hat{b}_1x + \hat{b}_2x^2 + \dots + \hat{b}_px^p$$

2 sources of error:

- 1) deviation of $\hat{f}^p(x)$ from $f(x)$ with ∞ sample: bigger p is better
- 2) deviation of $\hat{f}^p(x)$ from $E[\hat{f}^p(x)]$: smaller p is better

How do we trade off?

Actual and Fitted Regression



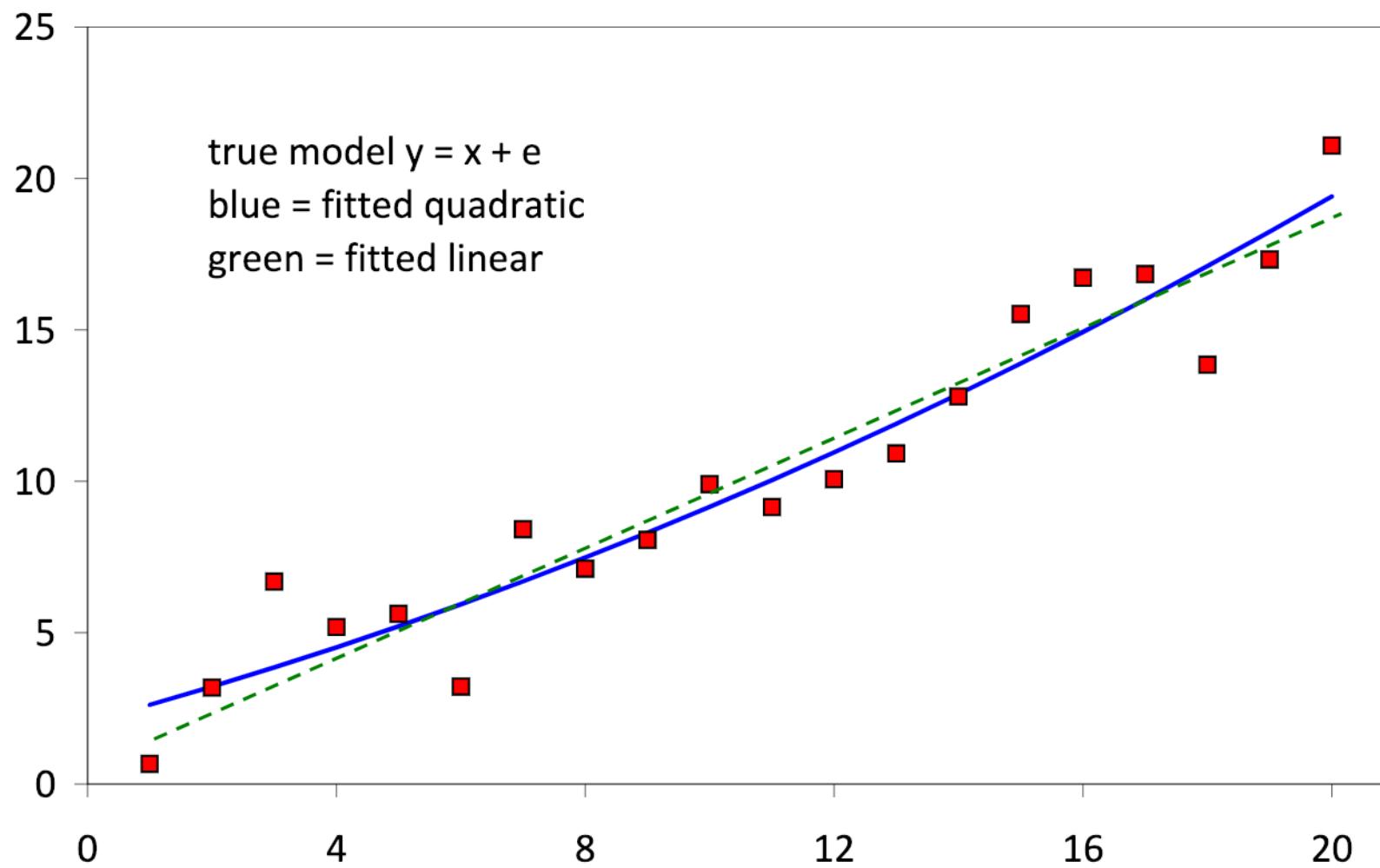
Variance vs bias

- given a model \hat{f} we can forecast y_0 associated with x_0
 - $y_0 - \hat{f}(x_0) = y_0 - f(x_0) + f(x_0) - E[\hat{f}(x_0)] + E[\hat{f}(x_0)] - \hat{f}(x_0)$
- $$\Rightarrow E[(y_0 - \hat{f}(x_0))^2] = E[\epsilon^2] \quad \text{irreducible error}$$
- $$+ E[(f(x_0) - E[\hat{f}(x_0)])^2] \quad \text{squared bias}$$
- $$+ E[(\hat{f}(x_0) - E[\hat{f}(x_0)])^2] \quad \text{variance of } \hat{f}$$

$$E[(y_0 - \hat{f}(x_0))^2] = E[\epsilon^2] + E[(f(x_0) - E[\hat{f}(x_0)])^2] + var[\hat{f}(x_0)]$$

- a simpler model may have more bias but less variance
- need to evaluate the tradeoff
- eg. polynomial $\hat{f}(x) = \hat{b}_0 + \hat{b}_1 x + \hat{b}_2 x^2 + \dots + \hat{b}_p x^p$
- larger p will have lower bias, but may be “noisier”

Actual and Fitted Regression



- standard econometric approach: emphasis on parameters
 - a) fit the model
 - b) evaluate model fit using R^2 , MSE
 - c) look at goodness of fit for certain configurations of x 's
 - d) consider augmenting/simplifying model

- statistical learning approach: emphasis on prediction
 - a) fit alt. models (different orders of p) on “training sample”
 - b) evaluate models on “test sample”
 - c) select on *out of sample* performance (test MSE)

Model selection is data driven to achieve best predictions

generalized approach: “regularization”

- very large set of potential x' s
- what if we try to minimize: $SSR + \lambda \sum_j \hat{\beta}_j^2$ (penalty function)
- for given λ : find $\hat{\beta}(\lambda)$
- now evaluate out-of-sample performance for each λ
- find best λ

Classification problems

$y \in \{y^1, y^2, \dots, y^G\}$ e.g. choice of car models

model designed to assign some value \hat{y} given predictors x

e.g.: have to “target” an advertising campaign,

or assign an occupation code to a person

For each obs i we can check if $y_i = \hat{y}_i$ (correct classification)

Error rate = $\frac{1}{N} \sum_i \mathbf{1}(y_i \neq \hat{y}_i)$: analogue to MSE

Benchmark classification:

$$\hat{y}_i = y^g \text{ if } P(y_i = y^g | x_i) = \max_h P(y_i = y^h | x_i)$$

the “Bayes classifier”. If we knew $P(y_i = y^g | x_i)$ for each outcome g we would assign the highest probability outcome. This is the analogue of $E[y|x]$ for the 1-dimensional outcome case.

Again: we don’t know $P(y_i = y^g | x_i)$ and have to estimate it given the data!

The statistical learning approach is to choose a classifier that has lowest error rate *in the test sample (or averaged test sample)*.

Flexible modeling - KNN

- how could we use an existing data set (training data) to flexibly estimate $f(x)$?
- example: KNN – “k nearest neighbors”
- suppose x is one-dimensional
- for given k and some x define $N_k(x) = \{x^1, x^2, \dots x^k\}$ that are “closest” to x
- then $\hat{f}^k(x) = \frac{1}{K} \sum_{i \in N_k(x)} y_i$
- now we can compare performance of alternative choices of k on the test sample

Re-sampling methods (ISLR, Chapter 5)

- in most cases we don't have a pre-identified test data set
- 3 basic approaches

HOLDOUT SAMPLE

Designate a subsample of the data as the “training” data set, and the remainder as the test data set. The test data set is called a holdout (or holdback) sample, since you are NOT using it to estimate the model, only to evaluate alternative models.

Two problems:

- a) different holdout samples may yield different answers
- b) the estimated model may be imprecise – using only 1/2 sample, we expect the standard errors to be $\sim \sqrt{2}$ times bigger than with the whole sample

LOOCV

Leave-one-out cross validation. For a sample of size N

- hold out i^{th} observation, estimate the model, predict \hat{y}_i , form $(y_i - \hat{y}_i)^2$
- repeat for $i = 1\dots N$ and form $\frac{1}{N} \sum_i (y_i - \hat{y}_i)^2$

Advantages:

- a) average across all the LOO sample results \Rightarrow only 1 “answer”
- b) we use $N - 1$ observations each time

BUT - there is a subtle disadvantage:

- the LOO samples are very highly correlated with each other – get nearly the same answer as the within-sample MSE
- LOOCV can result in over-fitting

this leads to the 3rd “intermediate” choice

k-FOLD CROSS VALIDATION

- choose $k = 5$ or 10
- divide the data into k *random subsets*.
- for $j = 1 \dots k$: select the j th subset as the holdout sample, fit the model, construct MSE_j on the hold out sample.
- construct $CV_k = \frac{1}{k} \sum_j MSE_j$

Advantages of k-FOLD CV:

- 1) only fit the model k times
- 2) different random partitions will yield slightly different answers, but with $k = 5, 10$ the differences tend to be very small
- 3) subsamples are correlated BUT less so than in the LOOCV case, so less over-fitting

k-fold CV can be used in lots of settings, e.g. classification settings....

Lecture 20: Shrinkage and Ridge Regression

1. prototype prediction problem: predicting group means
2. bias and variance
3. optimal combination of grand mean and group mean
4. example: gov't workers in CPS
5. ridge regression
6. choosing the “best” ridge model by cv

- we see an outcome y_{vi} for units i in buckets $v = 1 \dots V$
- we could estimate a model like:

$$y_{vi} = \sum_v D_{vi} \beta_v + u_{vi}$$

where D_{vi} = dummy for group v . BUT: this may be “over-fitting”

- the $\hat{\beta}'_v$ s will be noisy. The shrinkage idea: use a weighted average of grand mean \bar{y} and the group mean \bar{y}_v :

$$\hat{\beta}_v^{Shrink} = \theta \bar{y} + (1 - \theta) \bar{y}_v$$

OLS sets $\theta = 0$ (no weight): can we do better?

background: ISLR – Chapter 6.2 and “Lab 2” 6.6.1

Setup: outcome y , DGP has a “group” structure:

$$y_{vi} = \alpha + \beta_v + \epsilon_{vi}$$

y_{iv} = observed outcome for observation i in group $v = 1, 2, \dots, V$

Sample has N elements, with N_v observations for group v

Balanced sample: N_v same for each group; Unbalanced: N_v varies

$$y_{vi} = \alpha + \beta_v + \epsilon_{vi}$$

Assumptions:

1. $E[\beta_v] = \sum_v s_v \beta_v = 0$, $E[\beta_v^2] = \sum_v s_v \beta_v^2 = \sigma_v^2$
2. $N_v/N = s_v$ are fixed
3. ϵ_{vi} are i.i.d. with $E[\epsilon_{vi}] = 0$ and $var[\epsilon_{vi}] = \sigma_\epsilon^2$

What is our best estimate for the mean of a new observation from group v ?

Clearly, we want to get our best estimate of $\alpha + \beta_v$

$$y_{vi} = \alpha + \beta_v + \epsilon_{vi}$$

We can estimate an OLS model with dummies for each group, and form predictions. OLS prediction is just $\hat{\alpha} + \hat{\beta}_v = \bar{y}_v$. Is this the best we can do?

We could consider alternative estimators of the class:

$$\theta\bar{y} + (1 - \theta)\bar{y}_v$$

i.e., combine the *grand mean* and the *group-specific mean*. Notice that if $\theta < 1$ we are “shrinking” the prediction toward \bar{y} – so this class of models is called “shrinkage” estimators.

Why might we want to “shrink”?

- if N_v is relatively small, \bar{y}_v will be noisy
- \bar{y} is a good estimate of α
- if σ_v^2 is small there is information in \bar{y} that we can use!

Consider estimator $\theta\bar{y} + (1 - \theta)\bar{y}_v$ of a “new” obs from group v .

What is bias? $E[\bar{y}] = \alpha$; $E[\bar{y}_v] = \alpha + \beta_v$

$$\Rightarrow E[\theta\bar{y} + (1 - \theta)\bar{y}_v] = \alpha + (1 - \theta)\beta_v \Rightarrow Bias = \theta\beta_v$$

$$\Rightarrow E[Bias^2] = \theta^2 E[\beta_v^2] = \theta^2 \sigma_v^2$$

Note: unbiased if $\theta = 0 \Rightarrow$ all weight on \bar{y}_v (or if $\sigma_v^2 = 0$)

What about variance?

$$\bar{y} = \frac{1}{N} \sum_v \sum_{i \in v} y_{vi} = \frac{1}{N} \sum_v \sum_{i \in v} (\alpha + \beta_v + \varepsilon_{vi}) = \alpha + \frac{1}{N} \sum_v \sum_{i \in v} \varepsilon_{vi}$$

$$\Rightarrow \text{var}[\bar{y}] = \sigma_\epsilon^2 / N$$

$$\bar{y}_v = \frac{1}{N_v} \sum_{i \in v} y_{vi} = \alpha + \beta_v + \frac{1}{N_v} \sum_{i \in v} \varepsilon_{vi} \quad \Rightarrow \text{var}[\bar{y}_v] = \sigma_\epsilon^2 / N_v$$

Note: $\bar{y} - E[\bar{y}] = \frac{N_v}{N} \left(\frac{1}{N_v} \sum_{i \in v} \varepsilon_{vi} \right) + \frac{1}{N} \sum_{v' \neq v} \sum_{i \in v'} \varepsilon_{vi}$

and $\bar{y}_v - E[\bar{y}_v] = \frac{1}{N_v} \sum_{i \in v} \varepsilon_{vi}$

$$\Rightarrow \text{cov}[\bar{y}, \bar{y}_v] = \frac{N_v}{N} \times \sigma_\epsilon^2 / N_v = \sigma_\epsilon^2 / N = \text{var}[\bar{y}]$$

$$\Rightarrow \text{var}[\theta \bar{y} + (1 - \theta) \bar{y}_v] = \theta^2 \text{var}[\bar{y}] + (1 - \theta)^2 \text{var}[\bar{y}_v] + 2\theta(1 - \theta)\text{var}[\bar{y}]$$

Recall from Lecture 19: we want to minimize $E[bias^2] + variance$

$$variance = \theta^2 var[\bar{y}] + (1 - \theta)^2 var[\bar{y}_v] + 2\theta(1 - \theta)var[\bar{y}]$$

$$= \frac{\sigma_\epsilon^2}{N} \times \{\theta^2 + (1 - \theta)^2 \frac{N}{N_v} + 2\theta - 2\theta^2\}$$

$$E[bias^2] = \theta^2 \sigma_v^2 = \theta^2 k \frac{\sigma_\epsilon^2}{N}, \text{ where } k = \sigma_v^2 / \frac{\sigma_\epsilon^2}{N}$$

$$\begin{aligned} v + b &= \frac{\sigma_\epsilon^2}{N} \left(2\theta + \theta^2(k - 1) + (1 - \theta)^2 \frac{N}{N_v} \right) \\ \frac{\partial(v + b)}{\partial\theta} &= 2\frac{\sigma_\epsilon^2}{N} \left(1 + \theta(k - 1) - (1 - \theta)\frac{N}{N_v} \right) \\ \frac{\partial^2(v + b)}{\partial\theta^2} &= 2\frac{\sigma_\epsilon^2}{N} \left(k - 1 + \frac{N}{N_v} \right) > 0 \end{aligned}$$

FOC implies $1 + \theta(k - 1) - (1 - \theta)\frac{N}{N_v} = 0$ or:

$$\begin{aligned}\theta_v^* &= \frac{\frac{N}{N_v} - 1}{\frac{N}{N_v} - 1 + k} \\ &= \frac{N - N_v}{N - N_v + N_v k}\end{aligned}$$

Observations:

- 1) if $k = 0$ (*i.e.* $\sigma_v^2 = 0$) then $\theta_v^* = 1$ – use grand mean!
- 2) as $k \rightarrow \infty$, $\theta_v^* \rightarrow 0$
- 3) if $k = 1$ (*i.e.* $\sigma_v^2 = \frac{\sigma_\epsilon^2}{N}$) $\theta_v^* = (N - N_v)/N$, *and* $1 - \theta_v^* = N_v/N$
- 4) we can easily estimate $\frac{\sigma_\epsilon^2}{N}$ so the only unknown is σ_v^2

How can we proceed?

OLD WAY: estimate σ_v^2 by looking at OLS model:

$$y_{vi} = \alpha + D_i' \beta + \epsilon_{vi}$$

where D_i is a vector of dummies for group membership (V rows).

- instead of dropping one dummy, estimate with the restriction that $\sum_v (N_v/N) \beta_v = 0$ (restricted OLS). This gets the right constant and imposes our assumption that $E[\beta_v] = 0$.
- get $\hat{\beta}_v$ and estimate $\sigma_v^2 = \sum_v (N_v/N) \hat{\beta}_v^2$
- this is too big because $\hat{\beta}_v$ is noisy!
- but we know the sampling error $\hat{\sigma}_{\beta_v}^2$ so we subtract off $\sum_v (N_v/N) \hat{\sigma}_{\beta_v}^2$.

NEW WAY: cross validation.

- select a training sample and holdout sample.
- estimate \bar{y} and \bar{y}_v in the training sample
- choose a value of $\sigma_v \Rightarrow \theta_v^*$ Note: different θ_v^* for each group
- for observation i in the holdout sample in group $v(i)$:

prediction is: $\theta_{v(i)}^* \bar{y} + (1 - \theta_{v(i)}^*) \bar{y}_{v(i)}$

- form $MSE(\sigma_v)$

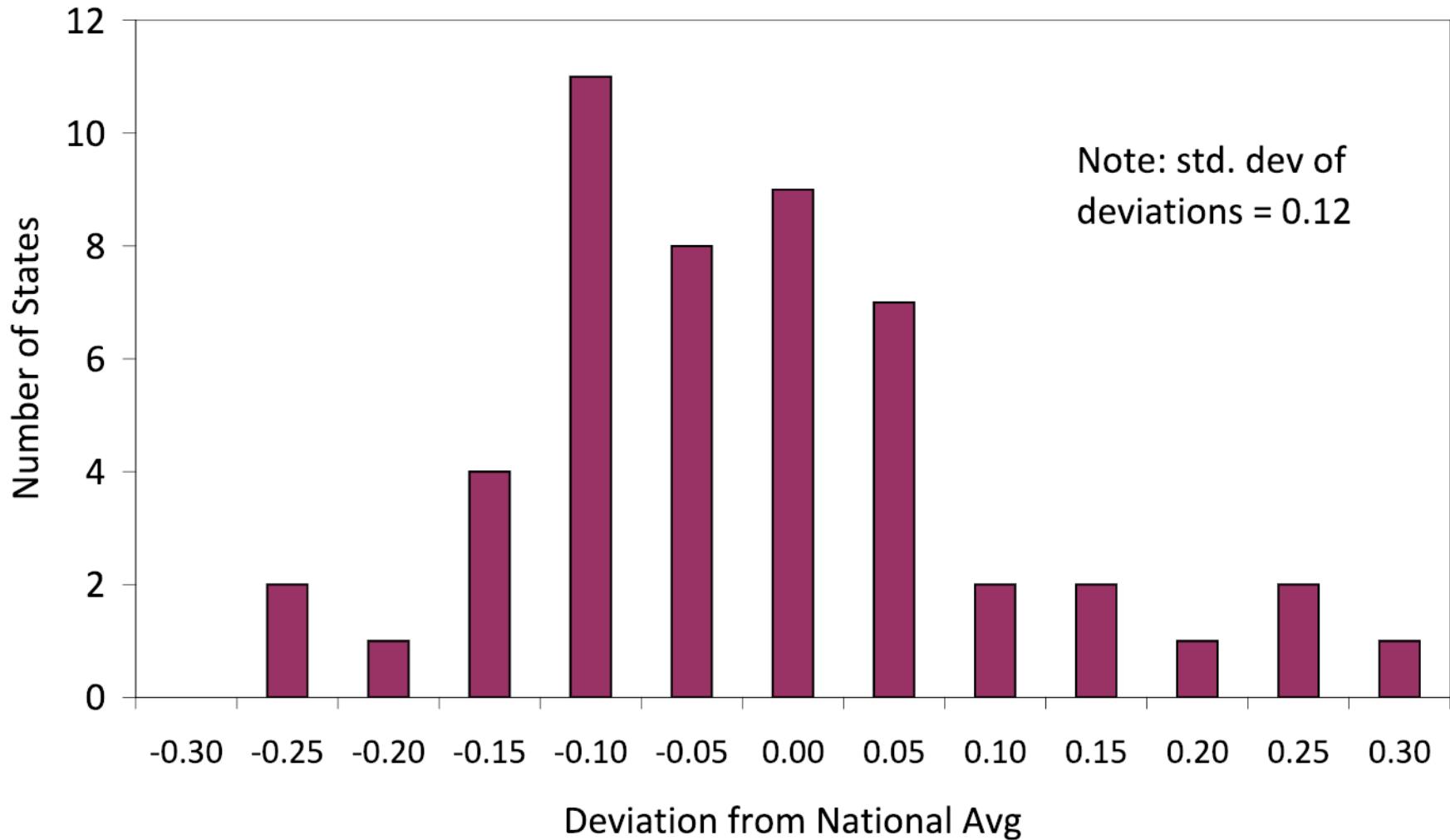
$$MSE(\sigma_v) = \frac{1}{N_H} \sum_{i \in H} (y_i - \theta_{v(i)}^* \bar{y} - (1 - \theta_{v(i)}^*) \bar{y}_{v(i)})^2$$

Example

We have CPS sample for 2012 and 2013, with data on female workers in the government sector. We are interested in forming predictions for wages for government workers in different states. The sample sizes per state range from 140 for Idaho to 1350 for California. A few relevant stats:

- N=14,666 – excluding 376 obs from DC (state=53) which is outlier
- grand mean of log hourly wage is 2.958 (\$19.26/hr).
- deviations of state from national average: std dev = 0.12 (crude estimate of σ_v^2)

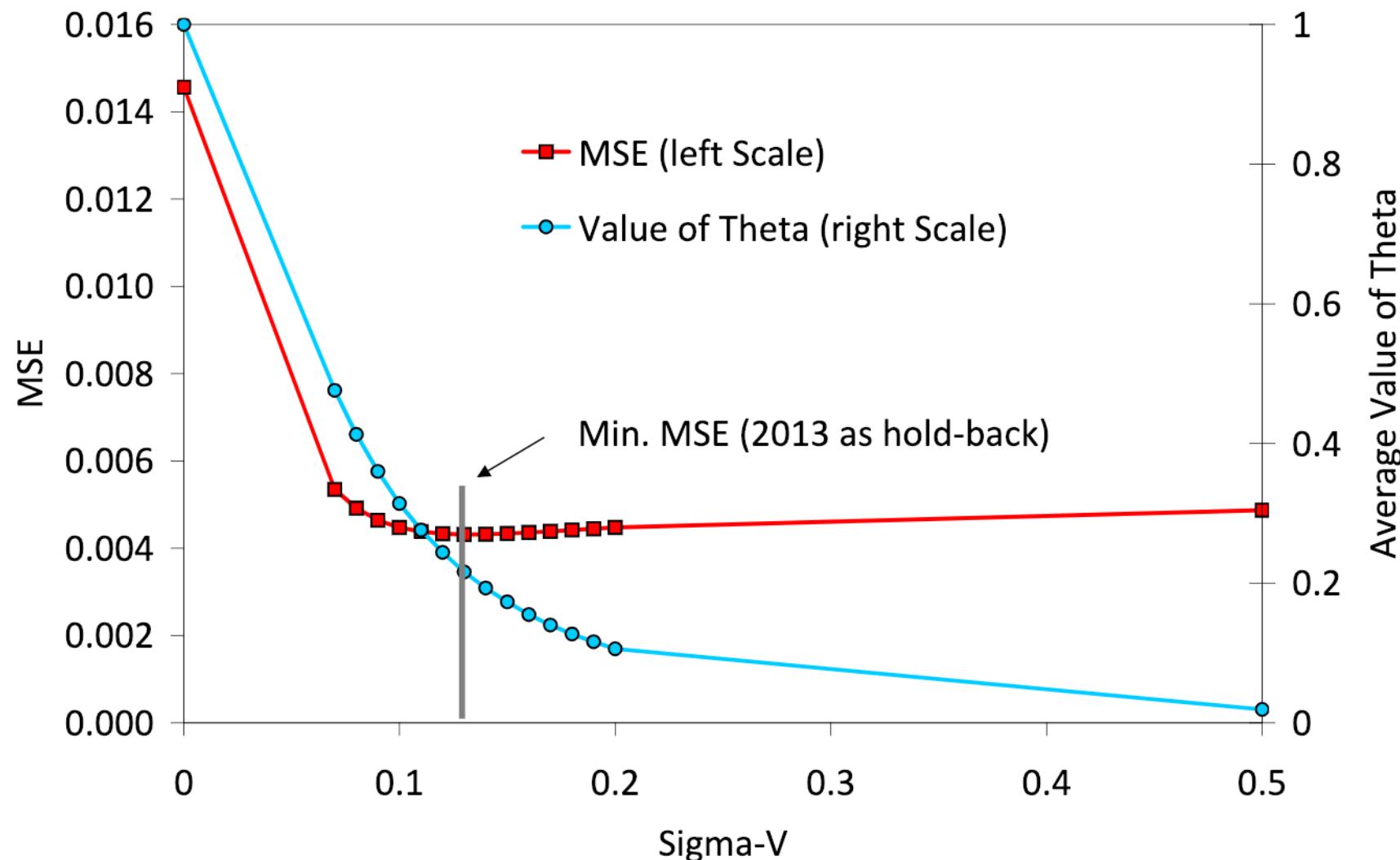
Histogram of Deviations of State Mean Wage for Government Workers from National Mean (Females only)



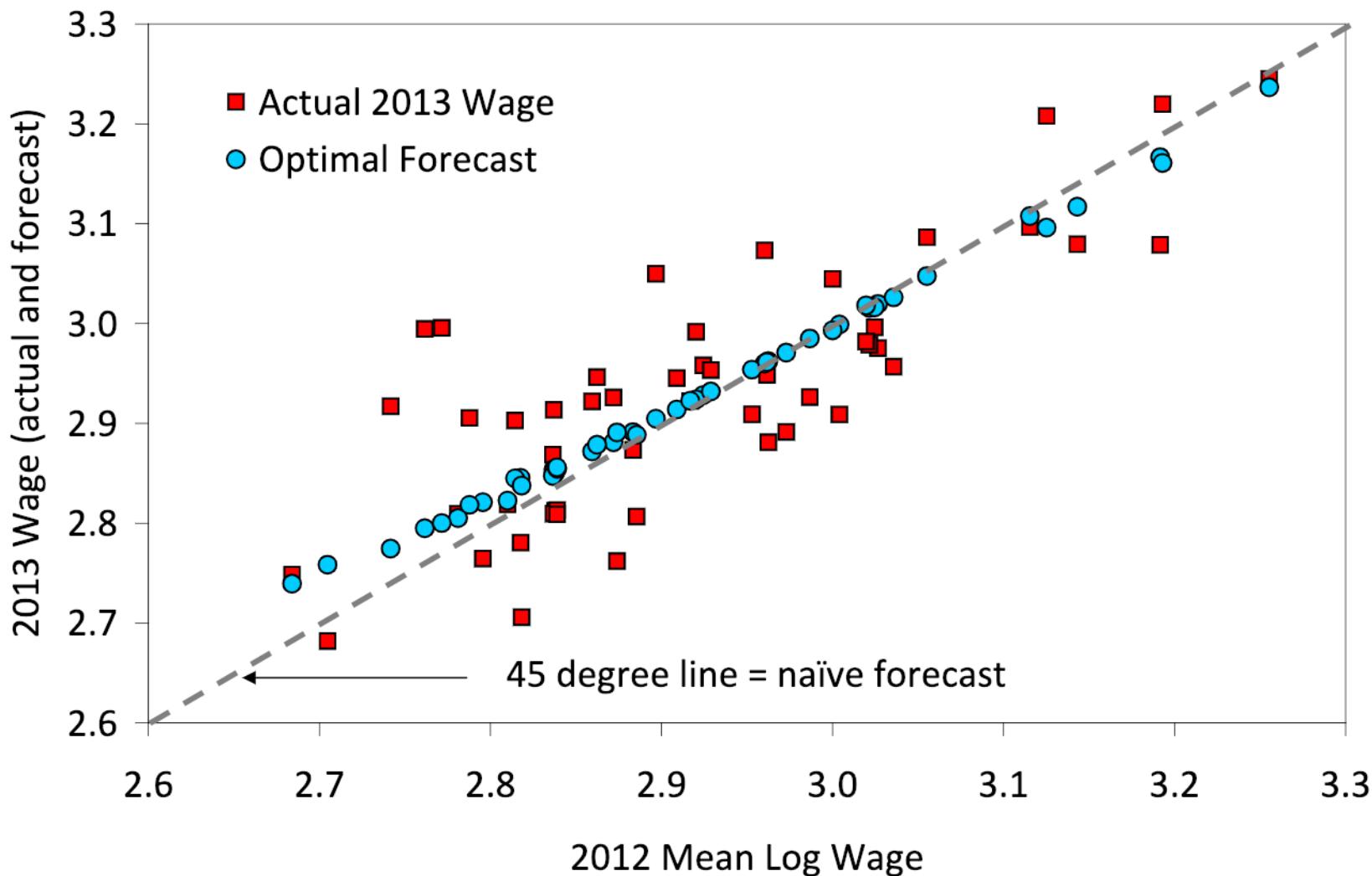
For CV: use 2012 as training sample and 2013 as test. (very slightly higher mean wages in 2013 which we'll ignore)

- compute MSE with different values for σ_v^2 – minimized for $\sigma_v^2 = 0.13$
- with this value: mean value of $\theta_{v(i)}^*$ is 0.22
- but values range from 0.339 (for Idaho) to 0.045 for California

Alternative Choices for Sigma-V



Optimal Forecast of 2013 Mean Log Wages:
Female Government Sector Workers by State



Ridge Regression

A more general version of the same class of problems
regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + \epsilon_i$$

We are worried that the estimated β 's will be “over-fit”. Idea:
penalize large $\hat{\beta}'s$

Ridge regression objective:

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

Notes:

- 1) don't penalize the constant
- 2) setting $\lambda = 0$ is the usual OLS regression model
- 3) larger values of λ penalize larger $\hat{\beta}'_j s$ more and shrink $\hat{\beta}'_j s$.
- 4) have to be careful to scale the $x'_j s$
 - in the group dummy case, all x' s are already on the same scale
 - otherwise rescale e.g.: $x_{1i} = (x_{1i} - \bar{x}_1)/\sigma(x_{1i})$ ("z-scores")

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

write in vector notation as

$$\min (y - \beta_0 - X\beta)'(y - \beta_0 - X\beta) + \lambda\beta'\beta$$

FOC for the coefficients β

$$-2X'(y - \beta_0 - X\beta) + 2\lambda\beta = 0$$

$$(X'X + \lambda I)\beta = X'(y - \beta_0)$$

$$\Rightarrow \beta = (X'X + \lambda I)^{-1} X'(y - \beta_0)$$

If $x'_{ji}s$ are rescaled to have mean 0 then $\beta_0 = \text{mean}(y_i)$

Lecture 21: Shrinkage, Ridge and Lasso

1. recap of shrinkage
2. ridge regression
3. Lasso
4. comparison of the methods

Shrinkage - recap

Outcome y has a “group” structure:

$$y_{vi} = \alpha + \beta_v + \epsilon_{vi}$$

$v = 1, 2, \dots, V$; with N_v obs. per group, $N_v/N = s_v$ fixed

1. $E[\beta_v] = \sum_v s_v \beta_v = 0$, $E[\beta_v^2] = \sum_v s_v \beta_v^2 = \sigma_v^2$

2. ϵ_{vi} are i.i.d. with $E[\epsilon_{vi}] = 0$ and $var[\epsilon_{vi}] = \sigma_\epsilon^2$

What is our best forecast of a future value from group v ?

Let our forecast be \hat{y} . Mean squared forecast error is:

$$\begin{aligned} MSFE(\hat{y}) &= E[(y_{vi} - \hat{y})^2] \\ &= E[(y_{vi} - E[y_{vi}] + E[y_{vi}] - E[\hat{y}] + E[\hat{y}] - \hat{y})^2] \\ &= E[(y_{vi} - E[y_{vi}])^2] + E[(E[y_{vi}] - E[\hat{y}])^2] \\ &\quad + E[\hat{y} - E[\hat{y}]]^2 - 2E[(E[y_{vi}] - E[\hat{y}])(\hat{y} - E[\hat{y}])] \\ &= \sigma_\epsilon^2 + E[bias(\hat{y})^2] + Var[\hat{y}] \end{aligned}$$

want to minimize expected squared bias + variance

Shrinkage estimator:

$$\hat{y}(\theta) = \theta\bar{y} + (1 - \theta)\bar{y}_v$$

What is bias? $E[\bar{y}] = \alpha$; $E[\bar{y}_v] = \alpha + \beta_v$

$$\Rightarrow E[\hat{y}(\theta)] = \alpha + (1 - \theta)\beta_v \Rightarrow Bias = \theta\beta_v$$

$$\Rightarrow E[Bias^2] = \theta^2 E[\beta_v^2] = \theta^2 \sigma_v^2$$

What is variance?

$$\begin{aligned} V[\hat{y}(\theta)] &= \theta^2 var[\bar{y}] + (1 - \theta)^2 var[\bar{y}_v] + 2\theta(1 - \theta)var[\bar{y}] \\ &= \frac{\sigma_\epsilon^2}{N} \times \left(\frac{N}{N_v} - 2\theta\left(\frac{N}{N_v} - 1\right) + \theta^2\left(\frac{N}{N_v} - 1\right) \right) \end{aligned}$$

note that $\partial V/\partial\theta = -2\left(\frac{N}{N_v} - 1\right)\frac{\sigma_\epsilon^2}{N}$ at $\theta = 0\dots$

Define

$$k = \sigma_v^2 / \frac{\sigma_\epsilon^2}{N}$$

Optimal choice of $\theta = \operatorname{argmin} E[Bias^2] + V[\hat{y}(\theta)]$

$$\begin{aligned}\theta_v^* &= \frac{\frac{N}{N_v} - 1}{\frac{N}{N_v} - 1 + k} \\ &= \frac{N - N_v}{N - N_v + N_v k}\end{aligned}$$

How do we find σ_v^2 ?

- Old way - use variation in $\hat{\beta}_v$ (adjusted for sampling variance)
- New way - cross-validation (e.g., 5-fold), searching over σ_v^2

Ridge Regression

A more general version of the same class of problems
regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_K x_{Ki} + \epsilon_i$$

We are worried that the estimated β 's will be “over-fit”. Idea:
penalize large $\hat{\beta}'s$

Ridge regression objective:

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

Notes:

- 1) don't penalize the constant
- 2) $\lambda = 0 \Rightarrow$ OLS regression model
- 3) larger λ 's penalize larger $\hat{\beta}'_j$ s more and shrink $\hat{\beta}'_j$ s
- 4) have to be careful to scale the x'_j s
 - in the group dummy case, all x' s are already on the same scale
 - otherwise rescale e.g.: $x_{1i} = (x_{1i} - \bar{x}_1)/\sigma(x_{1i})$ ("z-scores")

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

write in vector notation as

$$\min (y - \beta_0 - X\beta)'(y - \beta_0 - X\beta) + \lambda\beta'\beta$$

FOC for the coefficients β

$$-2X'(y - \hat{\beta}_0 - X\hat{\beta}) + 2\lambda\hat{\beta} = 0$$

$$(X'X + \lambda I)\hat{\beta} = X'(y - \hat{\beta}_0)$$

$$\Rightarrow \widehat{\beta}(\lambda) = (X'X + \lambda I)^{-1} X'(y - \widehat{\beta}_0)$$

Ridge regression estimate of the coefficients for the X 's :

$$\hat{\beta}(\lambda) = (X'X + \lambda I)^{-1} X'(y - \hat{\beta}_0)$$

What about $\hat{\beta}_0$?

$$\min (y - \beta_0 - X\beta)'(y - \beta_0 - X\beta) + \lambda\beta'\beta$$

So FOC is:

$$-\mathbf{1}'(y - \hat{\beta}_0 - X\hat{\beta}) - \mathbf{1}'(y - \hat{\beta}_0 - X\hat{\beta}) = 0$$

where $\mathbf{1}$ is the column vector of 1's, i.e.

$$\sum_i (y_i - \hat{\beta}_0 - x_{1i}\hat{\beta}_1 - x_{2i}\hat{\beta}_2 - \dots x_{Ji}\hat{\beta}_J) = 0$$

But if the x'_{ji} s are rescaled to have mean 0 then $\hat{\beta}_0 = \bar{y}$.

Ridge regression objective:

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

$$\Rightarrow \hat{\beta}(\lambda) = (X'X + \lambda I)^{-1} X'(y - \bar{y})$$

What if the x'_{ji} s are mutually exclusive dummies? Then the (v, v) element of $(X'X + \lambda I)$ is

$$\sum_i x_{vi}^2 + \lambda = N_v + \lambda$$

and

$$\hat{\beta}_v(\lambda) = \frac{\sum_i x_{vi}(y_i - \bar{y})}{N_v + \lambda} = \frac{N_v \bar{y}_v - N_v \bar{y}}{N_v + \lambda} = \frac{\bar{y}_v - \bar{y}}{1 + \lambda/N_v}$$

From lecture 20, $\hat{\beta}_0^{OLS} = \bar{y}$, and $\hat{\beta}_0^{OLS} + \hat{\beta}_v^{OLS} = \bar{y}_v$, so

$$\hat{\beta}_v(\lambda) = \frac{\hat{\beta}_v^{OLS}}{1 + \lambda/N_v}$$

$$\hat{\beta}_v(\lambda) = \frac{\hat{\beta}_v^{OLS}}{1 + \lambda/N_v}$$

So in the case where the x'_{ji} s are mutually exclusive dummies, ridge regression estimate “shrinks” each OLS coefficient toward 0 (by the same proportion).

Note that after fitting the ridge regression, the forecast of a new observation from group v is:

$$\bar{y} + \hat{\beta}_v(\lambda) = \bar{y} + \frac{\bar{y}_v - \bar{y}}{1 + \lambda/N_v} = \frac{\bar{y}_v + \frac{\lambda}{N_v}\bar{y}}{1 + \frac{\lambda}{N_v}}$$

which is a weighted average of \bar{y}_v and \bar{y} . So this is very similar to shrinkage.

Lasso

Lasso is a variant of ridge regression in which the penalty is based on the absolute value of the coefficients. Recall that ridge regression objective is:

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K \beta_j^2$$

This will “shrink” coefficients toward 0, but won’t eliminate them completely.

Lasso:

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K |\beta_j|$$

In general the solution to the Lasso minimization is to set some of the coefficients to 0 – thus Lasso is a “model selection” algorithm.

The reason why some coefficients are set to 0 can be seen from the nature of the Lasso objective function:

$$\begin{aligned} L(\beta) &= \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_K x_{Ki})^2 + \lambda \sum_{j=1}^K |\beta_j| \\ &= RSS(\beta) + P(\beta) \end{aligned}$$

Consider the nature of this function, focusing on variation wrt β_1 . The function $RSS(\beta)$ is quadratic in β_1 , whereas the penalty function $P(\beta)$ is a “check” function. If $\partial RSS(\beta)/\partial \beta_1$ evaluated at $\beta_1 = 0$ is not too big, it will be optimal to set $\beta_1 = 0$. By comparison, the ridge penalty function is also quadratic in β , so there is a smooth tradeoff between RSS and P .

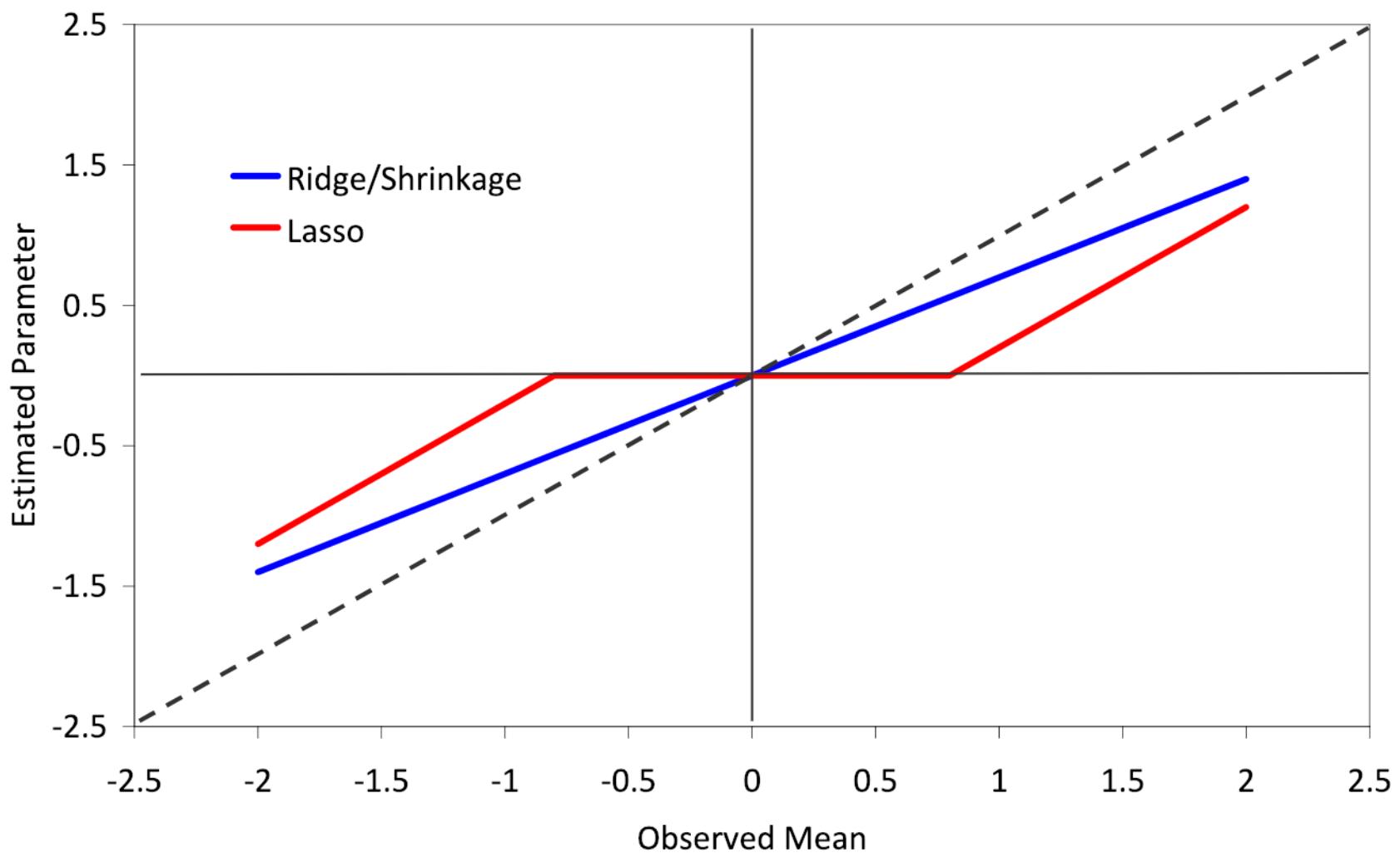
Comparing Shrinkage, Ridge and Lasso

In the special case of a group design (with equal numbers of observations per group) ridge regression is equivalent to the shrinkage estimator, and the best estimate for the coefficient for group v is:

$$\hat{\beta}_v(\lambda) = \frac{\hat{\beta}_v^{OLS}}{1 + \lambda/N_v} = \frac{\bar{y}_v - \bar{y}}{1 + \lambda/N_v}$$

Lasso performs a different kind of shrinkage: groups with \bar{y}_v close to the grand mean get $\hat{\beta}_v = 0$, and groups farther away get $\hat{\beta}_v = (\bar{y}_v - \bar{y}) - \delta$, where δ depends on the penalty λ .

Shrinkage: Ridge vs Lasso



Lecture 22: Model Selection (con't); Nonlinearity

1. more on Lasso
2. model selection (very quick overview)
3. comparison of methods – dummy variable example
4. non-linearities; splines

Lasso

$$\min_{\beta} \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_J x_{Ji})^2 + \lambda \sum_{j=1}^J |\beta_j|$$

variant of ridge regression, but $||$ penalty means that some (many) coefficients will be set to 0 – thus Lasso is a “model selection” algorithm.

Lasso objective function:

$$\begin{aligned} L(\beta) &= \sum_i (y_i - \beta_0 - \beta_1 x_{1i} - \dots - \beta_J x_{Ji})^2 + \lambda \sum_{j=1}^J |\beta_j| \\ &= RSS(\beta) + P(\beta) \\ &= RSS(\beta^{OLS}) + [RSS(\beta) - RSS(\beta^{OLS})] + P(\beta) \end{aligned}$$

What is $RSS(\beta) - RSS(\beta^{OLS})$? For any β :

$$\begin{aligned} RSS(\beta) &= (y - X\beta)'(y - X\beta) \\ &= y'y + \beta X'X\beta - 2y'X\beta \end{aligned}$$

$$\begin{aligned} RSS(\beta) - RSS(\beta^{OLS}) &= \beta X'X\beta - (\beta^{OLS})'X'X\beta^{OLS} \\ &\quad + 2y'X\beta^{OLS} - 2y'X\beta \end{aligned}$$

But $X'X\beta^{OLS} = X'y \Rightarrow 2y'X\beta^{OLS} = 2(\beta^{OLS})'X'X\beta^{OLS}$ so:

$$\begin{aligned} RSS(\beta) - RSS(\beta^{OLS}) &= \beta X'X\beta + (\beta^{OLS})'X'X\beta^{OLS} - 2(\beta^{OLS})'X'X\beta \\ &= (\beta - \beta^{OLS})'X'X(\beta - \beta^{OLS}) \end{aligned}$$

Intuition: Taylor expansion:

$$\begin{aligned} RSS(\beta) - RSS(\beta^{OLS}) &= (\beta - \beta^{OLS})' \frac{\partial RSS(\beta^{OLS})}{\partial \beta} \\ &\quad + \frac{1}{2} (\beta - \beta^{OLS})' \frac{\partial^2 RSS(\beta^{OLS})}{\partial \beta \partial \beta'} (\beta - \beta^{OLS}) \end{aligned}$$

BUT: $\frac{\partial RSS(\beta^{OLS})}{\partial \beta} = 0$ and $\frac{\partial^2 RSS(\beta^{OLS})}{\partial \beta \partial \beta'} = 2X'X$ so:

$$\begin{aligned} RSS(\beta) - RSS(\beta^{OLS}) &= (\beta - \beta^{OLS})' X' X (\beta - \beta^{OLS}) \\ &= N \cdot (\beta - \beta^{OLS}) V^x (\beta - \beta^{OLS}) \\ V^x &= \frac{1}{N} X' X = \frac{1}{N} \sum_i x_i x_i' \end{aligned}$$

So we have:

$$L(\beta) = RSS(\beta^{OLS}) + N \cdot (\beta - \beta^{OLS})' V^x (\beta - \beta^{OLS}) + \lambda \sum_{j=1}^J |\beta_j|$$

Lasso “trades off” the quadratic deviation $\beta - \beta^{OLS}$ vs. the loss function.

Notice that:

$$\begin{aligned} (\beta - \beta^{OLS})' V_{xx} (\beta - \beta^{OLS}) &= \sum_{j=1}^J V_{jj}^x (\beta_j - \beta_j^{ols})^2 + \\ &\quad 2 \sum_{j=1}^J \sum_{k \neq j}^K V_{jk}^x (\beta_j - \beta_j^{ols})(\beta_k - \beta_k^{ols}) \end{aligned}$$

So we are trying to minimize:

$$L^* = N \sum_{j=1}^J V_{jj}^x (\beta_j - \beta_j^{ols})^2 + 2N \sum_{j=1}^J \sum_{k \neq j}^K V_{jk}^x (\beta_j - \beta_j^{ols})(\beta_k - \beta_k^{ols}) + \lambda \sum_{j=1}^J |\beta_j|$$

Suppose $V_{jk}^x = 0$ (x' s uncorrelated). Then the part of L^* that varies with β_j is:

$$L_j^* = N \cdot v_j (\beta_j - \beta_j^{ols})^2 + \lambda |\beta_j|$$

where $v_j = V_{jj}^x$. Suppose that $\beta_j^{ols} > 0$. Differentiating from the right:

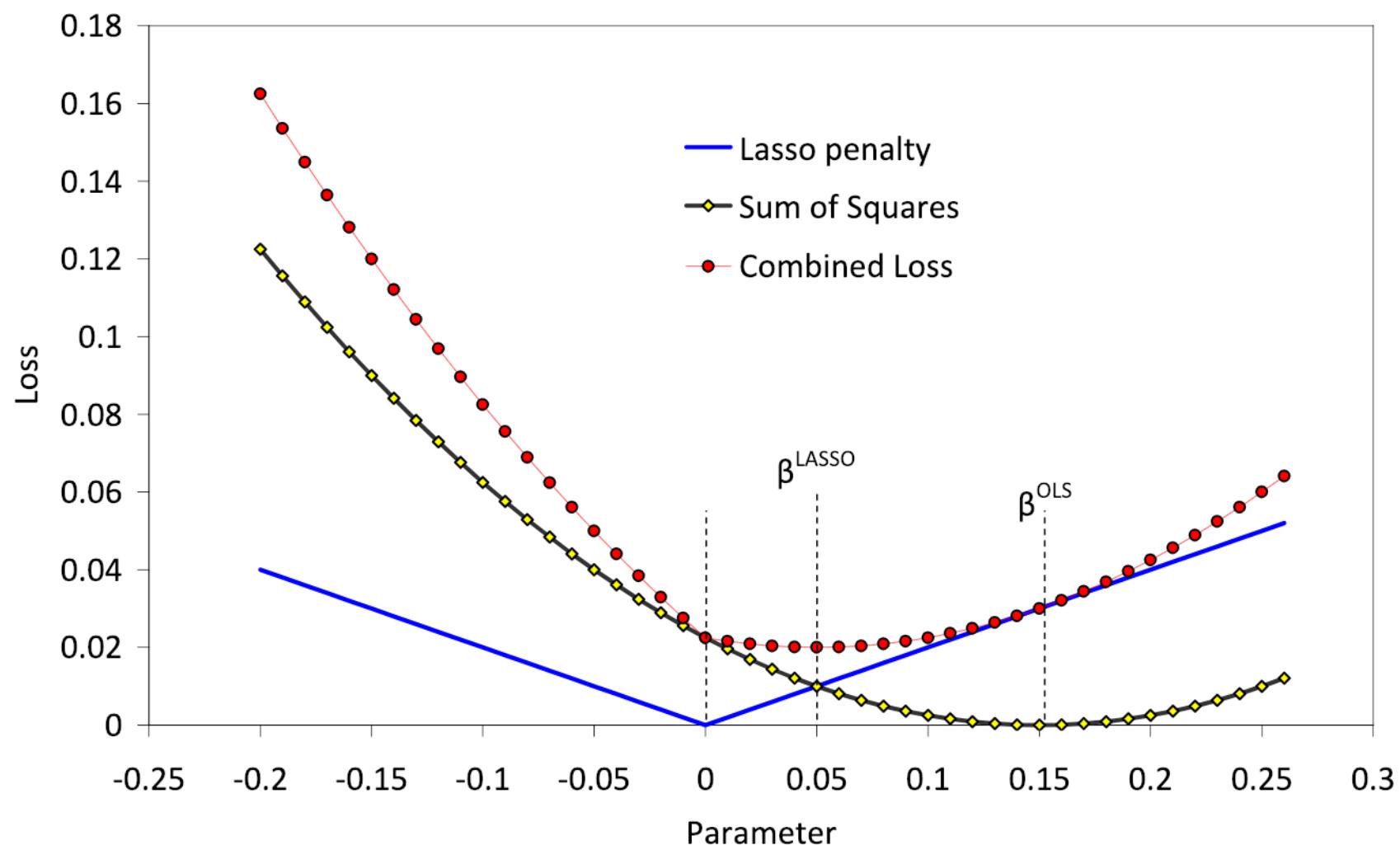
$$\frac{\partial L_j^*}{\partial \beta_j} = 2Nv_j(\beta_j - \beta_j^{ols}) + \lambda$$

If we have an interior solution with $\beta_j > 0$ then we must have

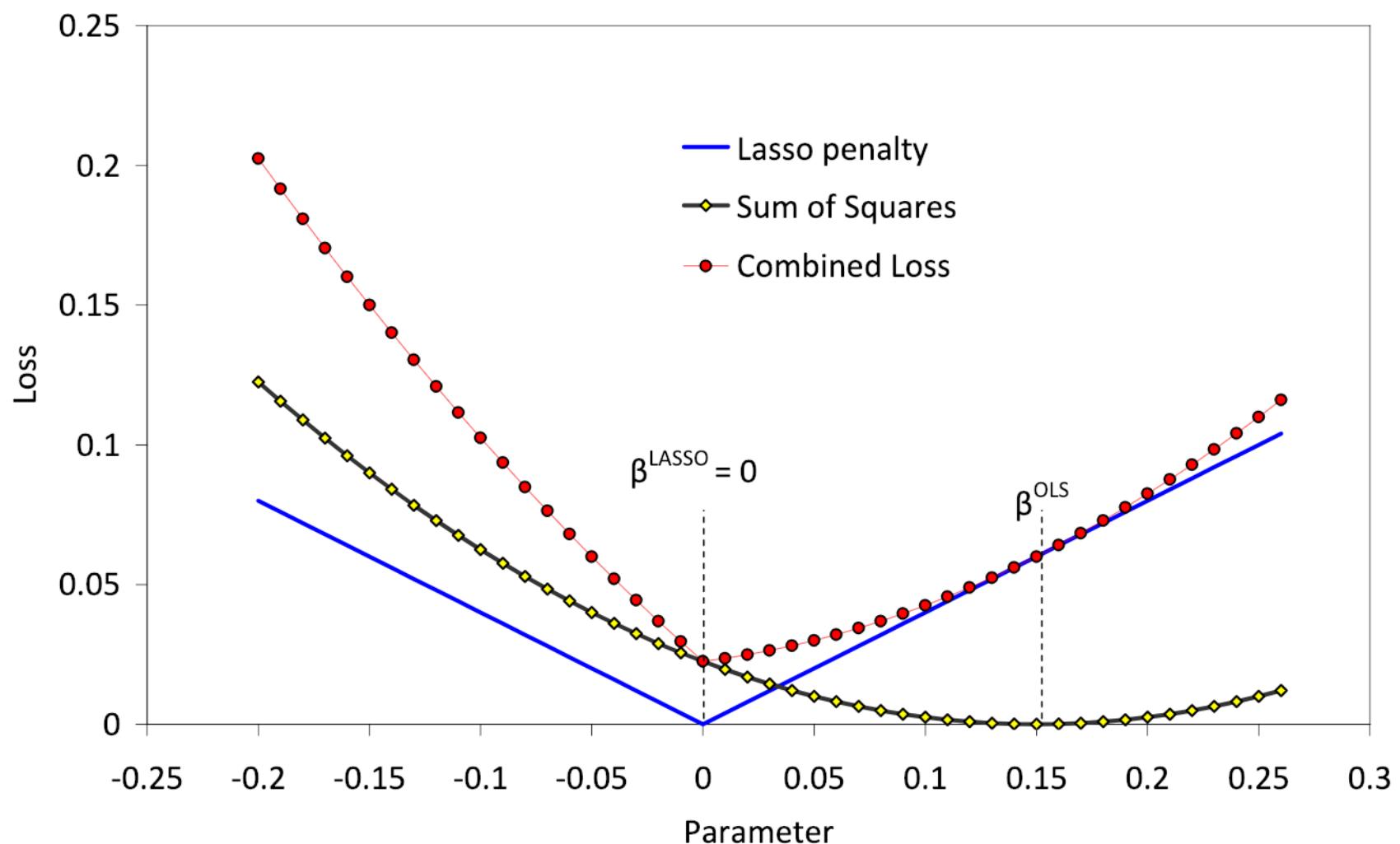
$$\beta_j = \beta_j^{ols} - \lambda/(2Nv_j)$$

But if $\beta_j^{ols} - \lambda/(2Nv_j) < 0$ then we have $\beta_j = 0$!

Lasso Loss Function



Lasso Loss Function



Let's consider a special case where the x 's are dummy variables for a set of mutually exclusive categories, so we know that the x 's are orthogonal to each other. Suppose we have an outcome y_{ij} and let's imagine we first transform the dependent variable to be $y_{ij} - \bar{y}$. In this case the "grand mean" of the transformed outcome is 0. Let's consider OLS, Ridge, and Lasso.

a) OLS gives the mean of the transformed variable for each category, so:

$$\beta_j^{ols} = \bar{y}_j - \bar{y}.$$

b) For a given λ , ridge regression gives (see lecture 21):

$$\beta_j^{ridge}(\lambda) = \frac{\bar{y}_j - \bar{y}}{1 + \lambda/N_j} = \frac{\bar{y}_j - \bar{y}}{1 + \lambda N/\bar{p}_j}$$

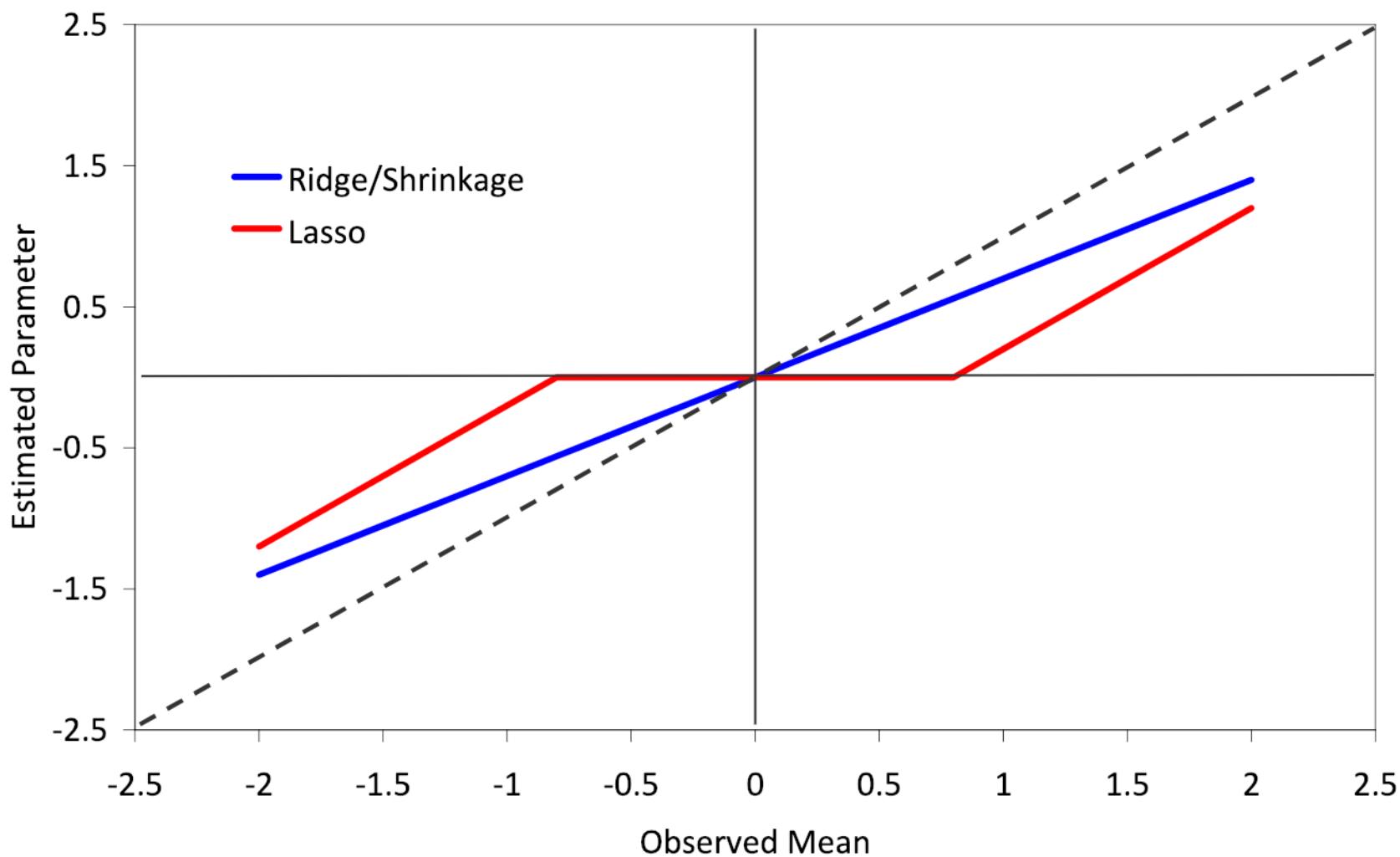
where \bar{p}_j is the fraction of observations in group j .

c) Lasso gives:

$$\begin{aligned}\beta_j^{lasso}(\lambda) &= \bar{y}_j - \bar{y} - \delta_j \text{ if } \beta_j^{ols} > \delta_j \\ &= \bar{y}_j - \bar{y} + \delta_j \text{ if } \beta_j^{ols} < -\delta_j \\ &= 0 \text{ otherwise}\end{aligned}$$

where $\delta_j = \lambda/(2Nv_j) = \lambda/(2N\bar{p}_j(1 - \bar{p}_j))$, since $v_j = \bar{p}_j(1 - \bar{p}_j)$. Notice that unlike Ridge/shrinkage, Lasso “kicks out” the dummies for the categories that get a small OLS estimate.

Shrinkage: Ridge vs Lasso



There are also other algorithms for kicking (or selecting) out regressors. In all cases you have to decide how to select a “best model”. Some of the criteria:

- a) within-sample R^2 – overfitting!
- b) cross-validation (k-fold, etc)
- c) penalized within-sample statistics

Examples of (c):

- adjusted R^2
- AIC, BIC

From lecture 6, recall the adjusted R^2 is

$$\bar{R}^2 = 1 - \frac{\frac{1}{N-J} \sum_{i=1}^N \hat{u}_i^2}{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2} = 1 - \frac{RSS/(N - J)}{\hat{\sigma}_y^2}$$

This penalizes bigger models (larger J).

Two other criteria (sample size N , number parameters p , $SSE = RSS$):

AIC (Akaike information criteria):

$$AIC = N(1 + \log(2\pi)) + N \log\left(\frac{SSE}{N}\right) + 2(p + 1)$$

BIC (Bayesian information criteria or “Schwartz” criterion):

$$BIC = N(1 + \log(2\pi)) + N \log\left(\frac{SSE}{N}\right) + \log N \times (p + 1)$$

Note BIC penalizes $\log N \times (p + 1)$ whereas AIC penalizes $2(p + 1)$.

Now that we've got comparison criteria (either MSE from CV, or AIC/BIC...) how do we proceed?

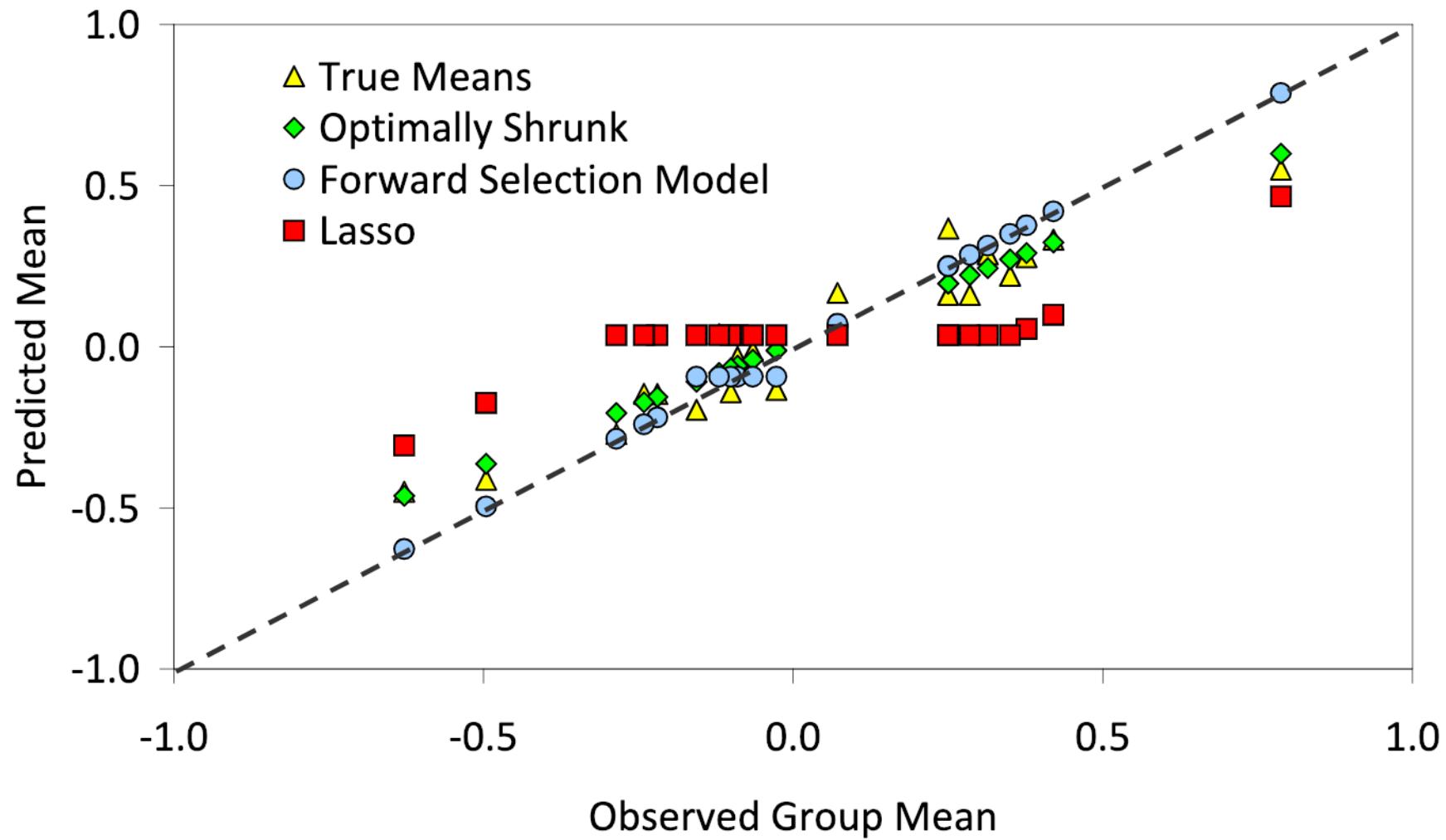
1. Best subset – evaluate all possible combinations of the X 's using chosen criteria. Not feasible unless $K = \max \# \text{ of predictors}$ is small.
2. Forward selection. Start with just a constant. Then add 1 regressor at at time, in each case adding the regressor from among those not selected that gives smallest SSE. Select among possible model sizes (0, 1, 2... included regressors) using criterion.
3. Backward selection. Start with the full model. Then subtract 1 regressor at at time, in each case removing the regressor that contributes the least to SSE. Select among possible model sizes using criterion.

Simple simulation example: I constructed a data set with 20 groups, each with 50 observations per group. I simulated data:

$$y_{iv} = \beta_v + \varepsilon_{iv}$$

where $\beta_v \sim N(0, \sigma_v^2)$ and $\varepsilon_{iv} \sim N(0, \sigma_\varepsilon^2)$. For this example we can calculate the “optimal” θ_j for Ridge. We can also run the data through Lasso, using cross validation to find a “best λ ” factor. And I tried forward selection. Results....

Predicted Group Means from Three Models
Simulation of 20 Groups of Size 50



Non-linear models

Chapter 7 of ISLR

So far we have assumed the set of regressors is known in advance. But in lots of situations, we suspect there may be a nonlinear relationship between x_j and y .

The usual approach is to posit a polynomial approximation. In general, any smooth function $f(x)$ can be approximated by a polynomial function:

$$f(x) \simeq b_0 + b_1x + b_2x^2 + \dots + b_px^p + r^p(x)$$

where $r^p(x)$ is the remainder (or approximation error). This leads to the most common approach to handling non-linearity: enter polynomials of x_j .

There are some major concerns about polynomials:

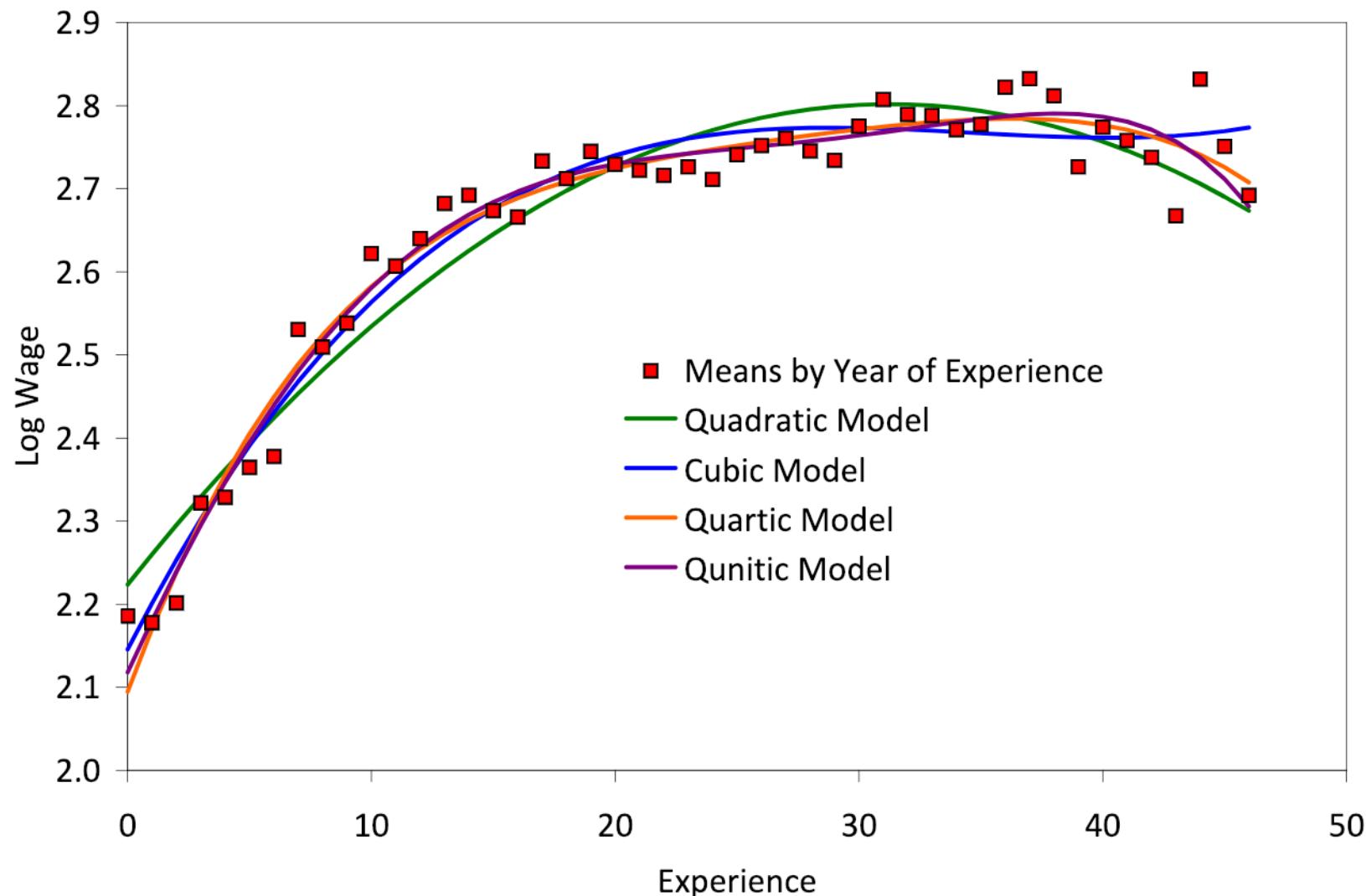
- 1) over-fitting
- 2) bad predictive performance at the extremes of x_j
- 3) “shape” of the fitted function is driven by “global” shape of the data; polynomials are not “local”

Example: wage data from March CPS, 2012 and 2013 (244,000 adults)

Look first at men with 12 years of education - the “experience” profile

Compare polynomials of order 2-3-4-5

Experience Profile: Males with 12 Years of Education



How can we get a more “locally adaptive” model? SPLINES!

- splines are functions that are fit locally and “strung together” to approximate the function of interest
- idea: break range of x_j into intervals. Fit a polynomial in each range, but force the resulting function to be “smooth” at each boundary or “knot”.
- classic case: cubic spline.
 - cubic polynomial in each interval
 - fitted function is continuous, and has continuous 1st and 2nd derivatives

Formula for cubic spline in x

- divide the data into intervals (typically equal fractions in each interval)
- call the k th knot: ξ_k
- define the functions $h(x, k) = \mathbf{1}[x > \xi_k] \times (x - \xi_k)^3$
- fit the model:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \sum_k a_k h(x, k) + u$$

- notice that at the first knot ($x = \xi_1$)

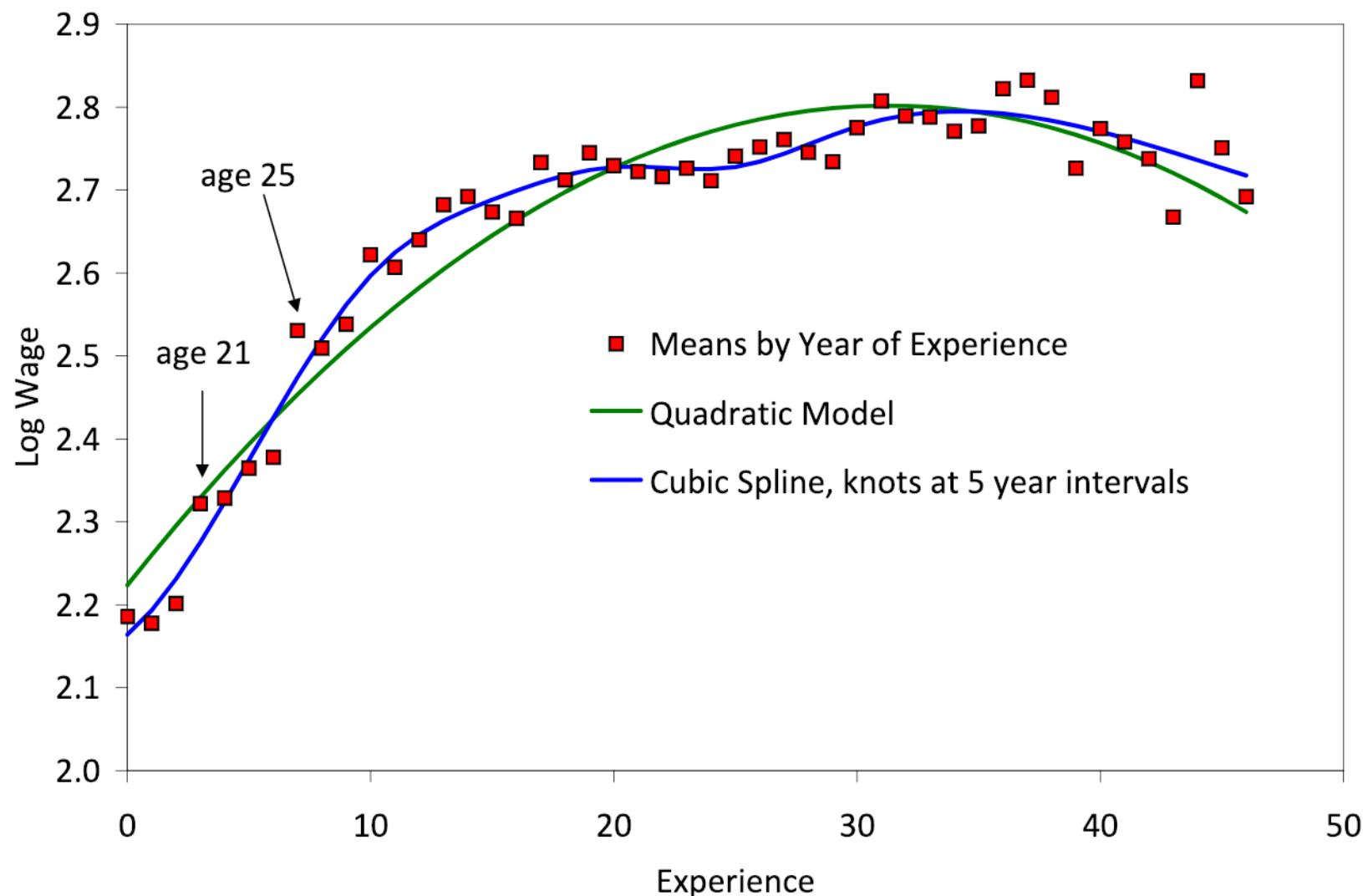
- from the left $\frac{dy}{dx} = b_1 + 2b_2\xi_1 + 3b_3\xi_1^2$, $\frac{d^2y}{dx^2} = 2b_2 + 6b_3\xi_1$,
- from the right at $x = \xi_1$, $\frac{dh(x,1)}{dx} = 3a_1(x - \xi_1)^2 = 0$, $\frac{d^2h(x,1)}{dx^2} = 6a_1(x - \xi_1) = 0$

How does this perform?

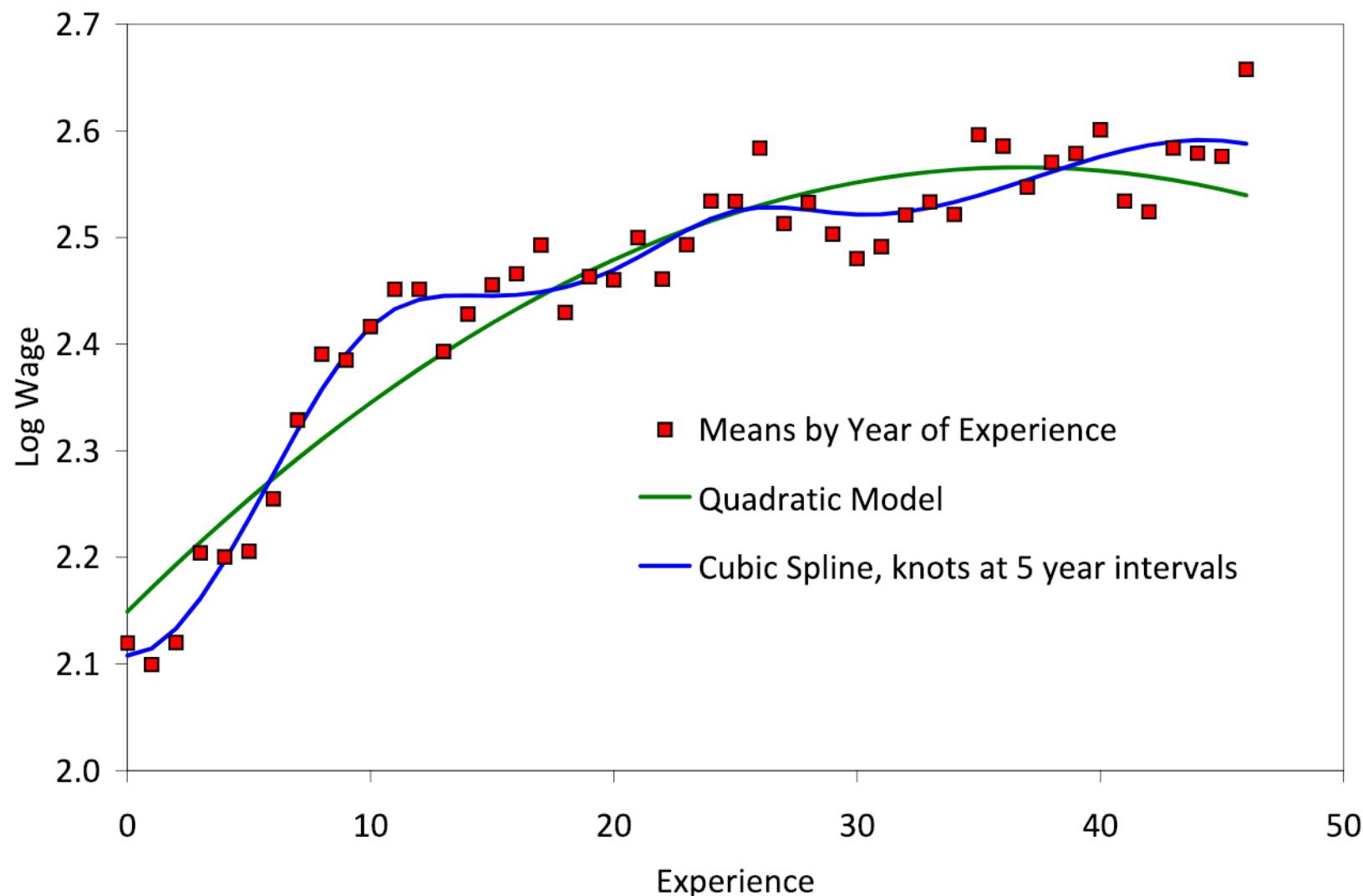
take the wage data, fit knots at experience=5,10.....

compare men, women with 12, 16 years of education

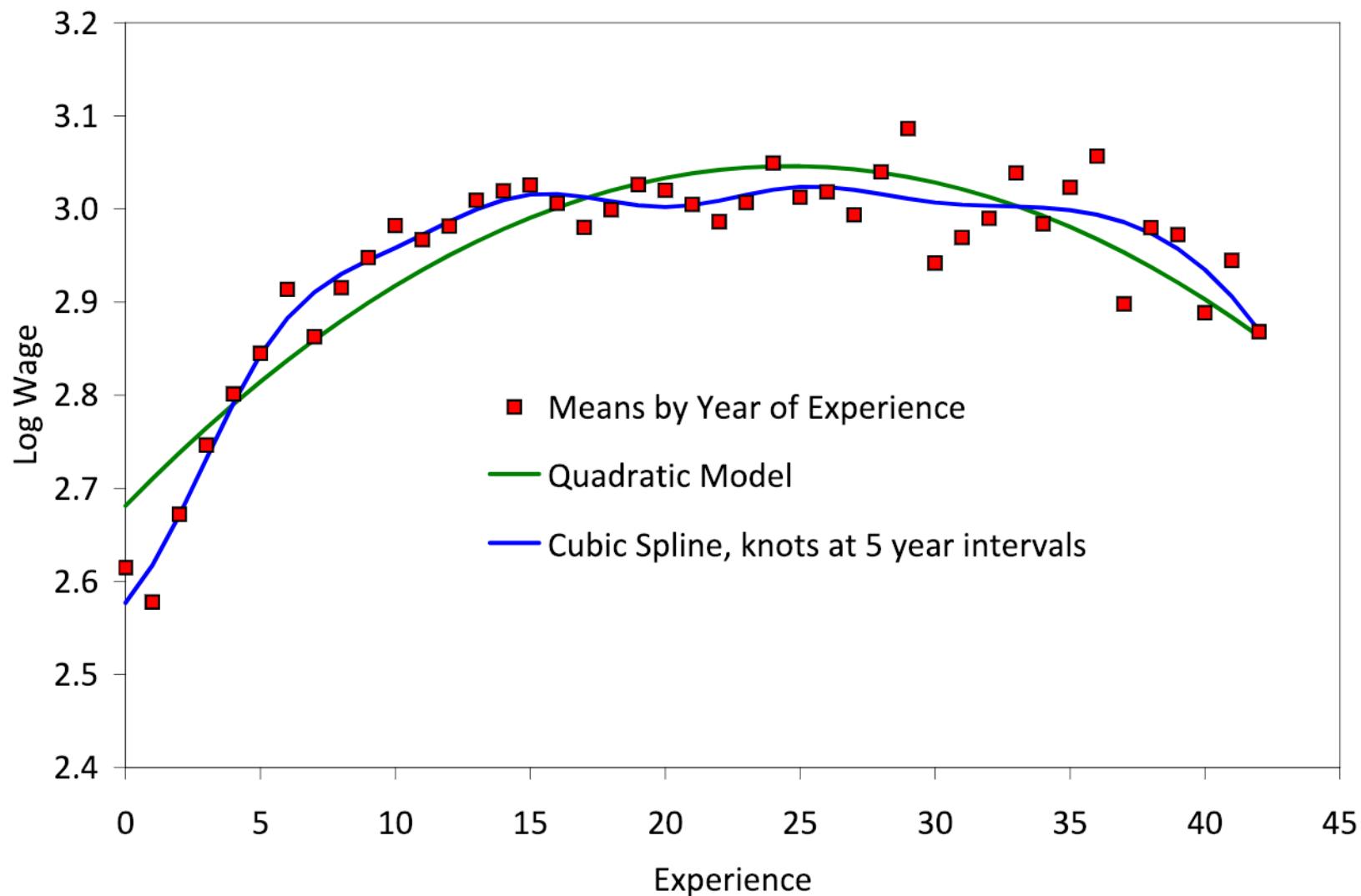
Experience Profile: Males with 12 Years of Education



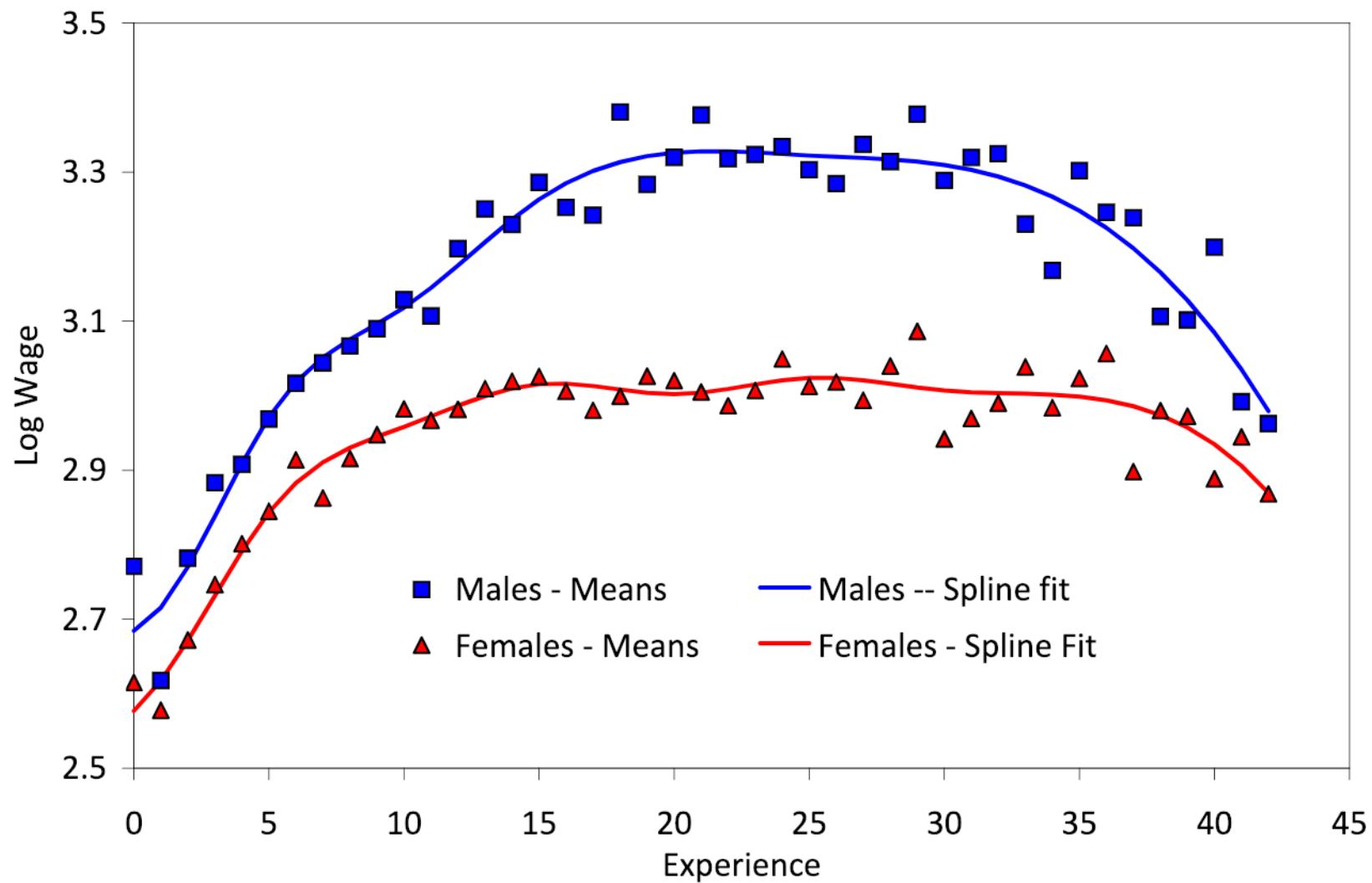
Experience Profile: Females with 12 Years of Education



Experience Profile: Females with 16 Years of Education



Experience Profile: Workers with 16 Years of Education



The functions $(1, x, x^2, x^3 \dots)$ are an example of a set of “basis functions”. There are many other possible basis functions:

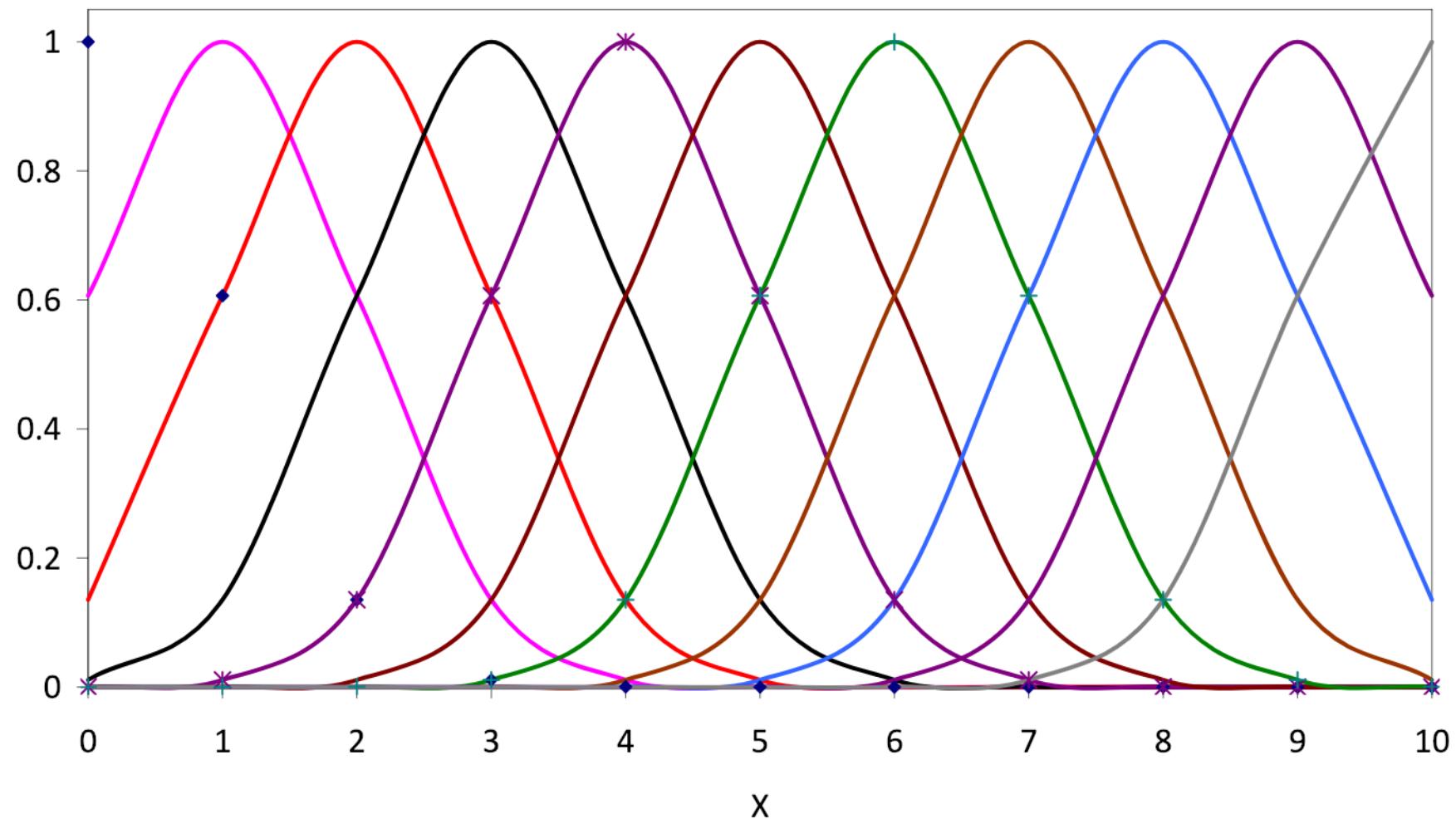
- sin functions (Fourier series)
- Legendre polynomials $(1, x, \frac{1}{2}(3x^2 - 1), \dots)$

For “very local” fitting, a popular choice is a set of translated normal density functions (the Gaussian basis functions). Divide the range of x into intervals of length s , let $\mu_k = x^{\min} + ks$ be the k^{th} knot, and let the k^{th} function be

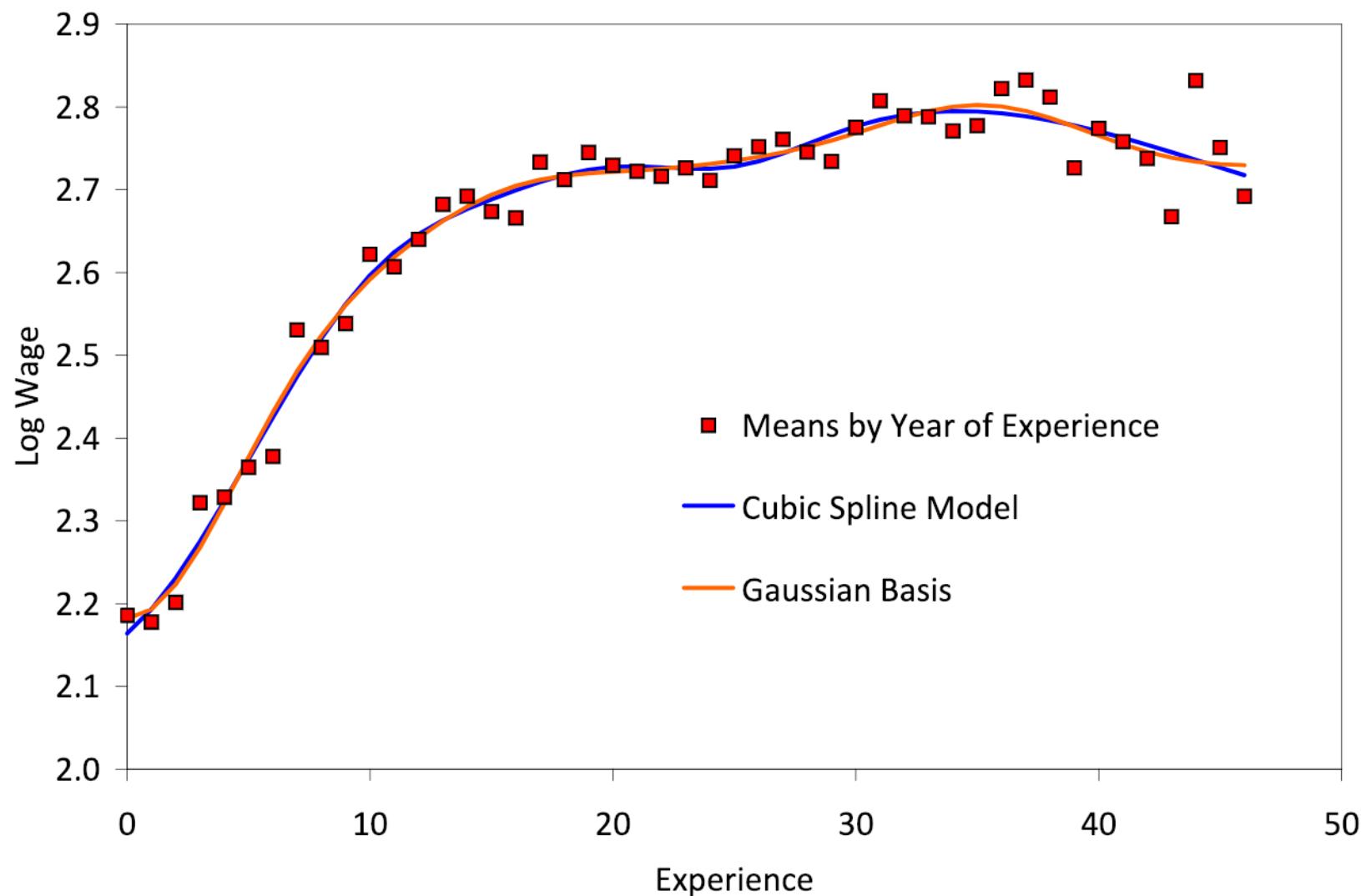
$$\phi_k(x) = \exp\left(-\frac{1}{2}(x - \mu_k)^2/s^2\right)$$

As we can see, this fits about as well as the cubic spline!

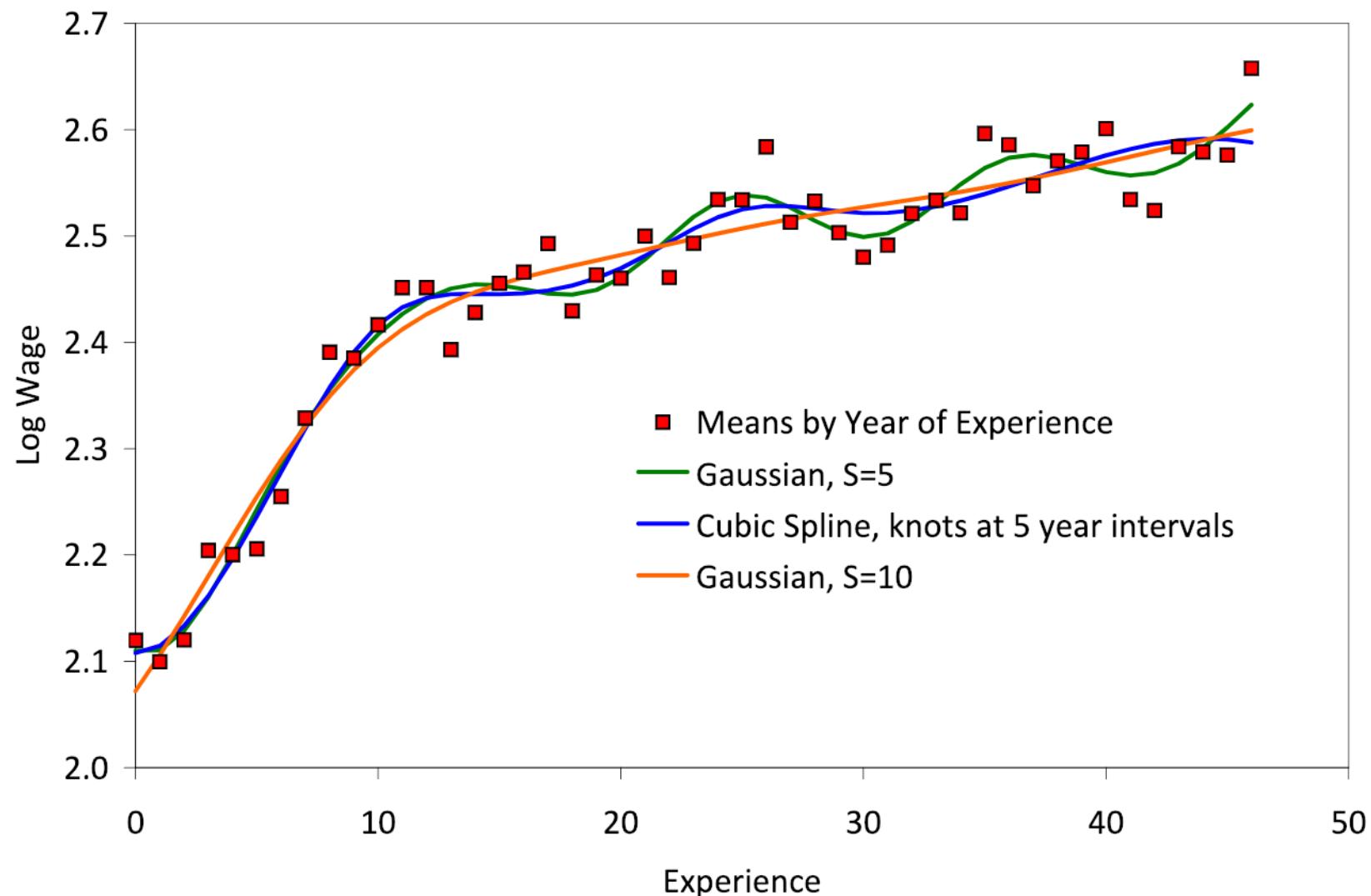
Gaussian Basis Functions



Experience Profile: Males with 12 Years of Education



Experience Profile: Females with 12 Years of Education



Lecture 23 – Classification Problems (ISLR, ch. 4)

1. classification problems
2. logit
3. decision boundaries
4. example: identifying females
5. tree models - introduction

Classification problems

$y \in \{y^1, y^2, \dots, y^G\}$ e.g. choice of car models

model designed to assign some value \hat{y} given predictors x

e.g.

- 1) have to “target” an ad: need to decide if a query is from M/F.
- 2) have to assign an occupation code to a person.

Best estimate of correct class:

$$\hat{y}_i = y^g \text{ if } P(y_i = y^g | x_i) = \max_h P(y_i = y^h | x_i)$$

the “Bayes classifier”. If we knew $P(y_i = y^g | x_i)$ for each outcome g we would assign the highest probability outcome. This is the analogue of $E[y|x]$ for the 1-dimensional outcome case.

We don’t know $P(y_i = y^g | x_i)$, so we have to estimate it given the data.

once we have predicted the classes for each observation, we can calculate mis-classification rates. In the case of 2 groups, we have:

among true 1's: "false negative rate" = fraction predicted 0's

among true 0's: "false positive rate" = fraction predicted 1's

if $y_i = \text{true}$ and $\hat{y}_i = \text{predicted}$

$$FNR = \frac{\sum_i (1 - \hat{y}_i) y_i}{\sum_i y_i}$$
$$FPR = \frac{\sum_i \hat{y}_i (1 - y_i)}{\sum_i (1 - y_i)}$$

		Predicted Status	
		Positive	Negative
Actual Status:	Positive	true pos.	false neg.
	Negative	false pos.	true neg.

False pos. rate = #false pos./#actual neg.

False neg. rate = #false neg./#actual pos.

The Logit Model

let's consider 2-group problems (i.e., $y_i \in \{0, 1\}$). Logit model assumes:

$$P(y_i = 1|x_i) = \Lambda(x_i\beta) \equiv \frac{e^{x_i\beta}}{1 + e^{x_i\beta}}$$

where $x_i = (1, x_{1i}, \dots x_{Ki})$ is a set of observed covariates (normally, including a constant). So to implement a logit classifier, we have to choose a set of x 's (and the right functional form) and estimate β . Usually, β is estimated using maximum likelihood, which maximizes:

$$\frac{1}{N} \sum_i \log P(y_i|x_i, \beta)$$

the log of the probability of observing the actual values of y_i in the sample, given x_i and the assumed “logit” form for $P(y_i|x_i, \beta)$.

Recall from Lecture 9 that $P(y_i|x_i, \beta) = \Lambda(x_i' \beta)^{y_i} (1 - \Lambda(x_i' \beta))^{1-y_i}$ so the objective is to maximize

$$L = \frac{1}{N} \sum_i y_i \log(\Lambda(x_i' \beta)) + (1 - y_i) \log(1 - \Lambda(x_i' \beta))$$

and we showed that the FOC for β can be written as:

$$\frac{1}{N} \sum_i x_i (y_i - \Lambda(x_i' \hat{\beta})) = 0$$

which is $K+1$ equations in the $K+1$ parameters of $\hat{\beta}$, implying that the prediction error $y_i - \hat{p}_i$ is orthogonal to x_i – just like the FOC for OLS!

Notice that after we've got $\hat{\beta}$ we can calculate

$$\begin{aligned}\hat{P}(y_i = 1|x_i) &= \frac{e^{x_i\hat{\beta}}}{1 + e^{x_i\hat{\beta}}} \\ \hat{P}(y_i = 0|x_i) &= \frac{1}{1 + e^{x_i\hat{\beta}}}\end{aligned}$$

so given x_i we classify i as 0 or 1 depending on whether $e^{x_i\hat{\beta}} \leq 1$ or

$$\log \left(\frac{\hat{P}(y_i = 1|x_i)}{\hat{P}(y_i = 0|x_i)} \right) = x_i\hat{\beta} \leq 0$$

Let's suppose that we have 2 covariates and a constant. So the decision depends on whether:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \leq 0$$

Given the logit parameter estimates, we classify depending on whether:

$$\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} \leq 0$$

so in (x_1, x_2) space there is a line:

$$x_2 = -(\hat{\beta}_0 / \hat{\beta}_2) - (\hat{\beta}_1 / \hat{\beta}_2)x_1.$$

If $\hat{\beta}_2 > 0$, points above this line get classified as 1's; points below get classified as 0's. This line is called the "Bayesian decision boundary".

We can increase the flexibility of the classifier boundary by allowing non-linearities in the x 's (polynomials, splines, basis functions).

We could also consider a linear probability model: $P(y_i = 1|x_i) = x_i\theta$. Notice that we can estimate θ by fitting a linear regression, since $E[y_i = 1|x_i] = P(y_i = 1|x_i)$ is (by assumption) linear in x :

$$y_i = x_i\theta + u_i$$

From this model, $\hat{P}(y_i = 1|x_i) = x_i\hat{\theta}$, and $\hat{P}(y_i = 0|x_i) = 1 - x_i\hat{\theta}$. We classify $i = 0$ or 1 depending on whether

$$x_i\hat{\theta} \leq 1 - x_i\hat{\theta}$$

$$x_i(\hat{\theta}) \leq 1/2$$

which also gives a linear Bayesian decision boundary.

In the 2-variable case LP case, we classify as a “1” iff:

$$\hat{\theta}_0 + \hat{\theta}_1 x_{1i} + \hat{\theta}_2 x_{2i} \leq 1/2$$

so in (x_1, x_2) space there is a line:

$$x_2 = 1/2 - (\hat{\theta}_0/\hat{\theta}_2) - (\hat{\theta}_1/\hat{\theta}_2)x_1$$

and points above the line are classified as 1's. Note that the slope of the logit and LP boundaries will be the same if

$$(\hat{\beta}_1/\hat{\beta}_2) = (\hat{\theta}_1/\hat{\theta}_2)$$

Example: identifying females

Suppose we have information on education and wages, and want to decide if a person is male or female. We can estimate a logit, estimate an LP model, or try to identify a Bayesian decision boundary “non-parametrically”.

Data set = 10% sample of March 2012/13 CPS.

Mean Characteristics of Females and Males

	Females	Males
Log wage	2.761 (0.723)	2.955 (0.792)
Education (yrs)	14.157 (2.532)	13.817 (2.697)

Sample contains 7,289 females and 7,846 males from 2012 and 2013 March CPS (age 26-60 only).

Estimates from Logit and Linear Probability Models

	Logit	Linear Prob.
Constant	-0.179	0.456
Log wage	-0.502 (0.008)	-0.120 (0.002)
Education (yrs)	0.110 (0.002)	0.026 (0.001)
Slope of BD boundary:	4.564	4.615

Sample contains 72,700 females and 78,893 males from 2012 and 2013 March CPS (age 26-60 only).

What's wrong with Logit/LP?

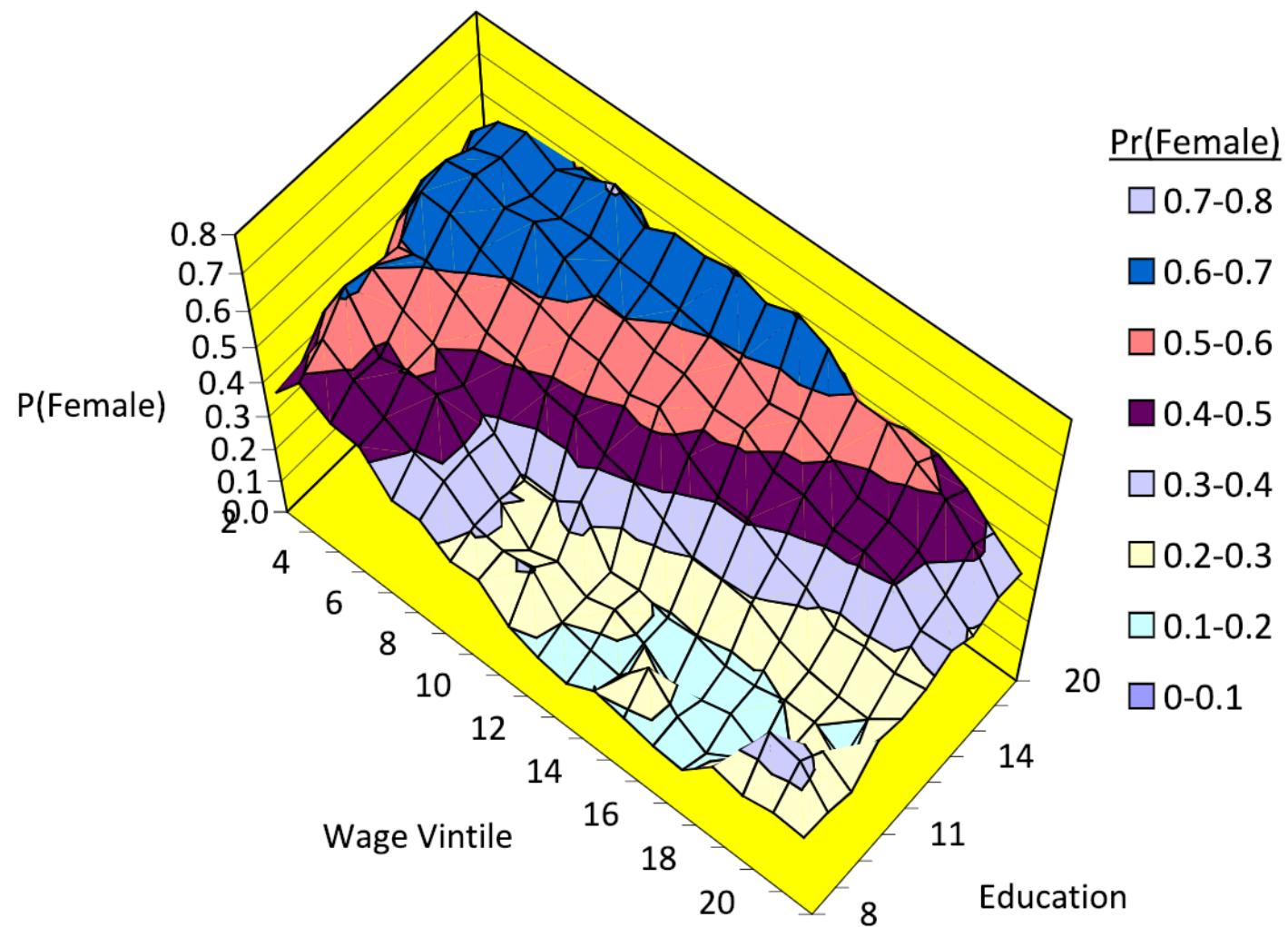
- sometimes, we think that the classification probabilities depend on the characteristics x in a model complex way
- example: assigning a “risk score” for the probability of an adverse health event (death, etc)
- features like age, weight, blood pressure, presence of diseases (diabetes) may interact in a complex way
- ideally, we'd like to take all the x 's, divide into discrete values, and fit a “fully interacted model”
- this will work for discrete variables like age, and dummies

- but how do we divide up continuous variables like the wage?
- one idea: “grid” wages into quartiles/deciles/vingtiles
- e.g.: education in 9 cells, log wage in 20 vingtiles
- get mean fraction female in 9×20 buckets

Issues:

- 1) how fine to make the buckets
- 2) should we smooth over buckets in some way?

Probability of Female Gender, Given Wage and Education



As with linear regression models, we can use cross-validation (e.g., k-fold) to evaluate alternative “smoothing choices”. For example, we could compare choices of the number of bins for wages (5,10,15,20...) assuming we use 9 bins for education. We can evaluate alternative models using the out-of-sample misclassification rate.

We can also use “penalized” logistic regression:

$$\max_{\beta} \sum_i y_i \log(\Lambda(x'_i \beta)) + (1 - y_i) \log(1 - \Lambda(x'_i \beta)) - \lambda \sum_{j=1}^J |\beta_j|$$

where no penalty is applied to the constant. As in the standard lasso, for larger values of λ this will set some coefficients to 0. We can select λ using (say) 5-fold CV to minimize out-of-sample misclassification.

Tree models

- the idea of a tree model is to divide the predictor space into a partition of “retangles” $R_1, R_2 \dots R_J$ then use as a predictor for any observation that falls into retangle R_j the most common outcome (in the simple discrete case, either 0 or 1)
- the rectangles are defined by *recursive binary splitting*, following a “tree” process:

first split by whether x_3 (say) is above or below some cutoff

continue splitting branches until some criterion is met

- notice that recursive binary splitting will always generate partitions that are (generalized) rectangles e.g.

split 1: $x_1 \geq k_1$;

split 2 (for $x_1 > k_1$): $x_2 \geq k_2$

split 3 (for $x_1 < k_1$): $x_2 \geq k'_2$

- in contrast, a logit model of the form

$$P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

will lead to classification regions that are not rectangles!

How do we choose a split point? Suppose there are N observations to be split, and the fraction of 1's is p . If we split into 2 groups, we'll have group size/proportions of 1's = $\{N_1, p_1\}$ and $\{N_2, p_2\}$. Define the Gini index of "impurity" in the subgroups as

$$\begin{aligned}G_1 &= 2N_1p_1(1 - p_1) \\G_2 &= 2N_2p_2(1 - p_2)\end{aligned}$$

A standard approach is to find the split that minimizes $G_1 + G_2$. Another measure that is sometimes used is "entropy":

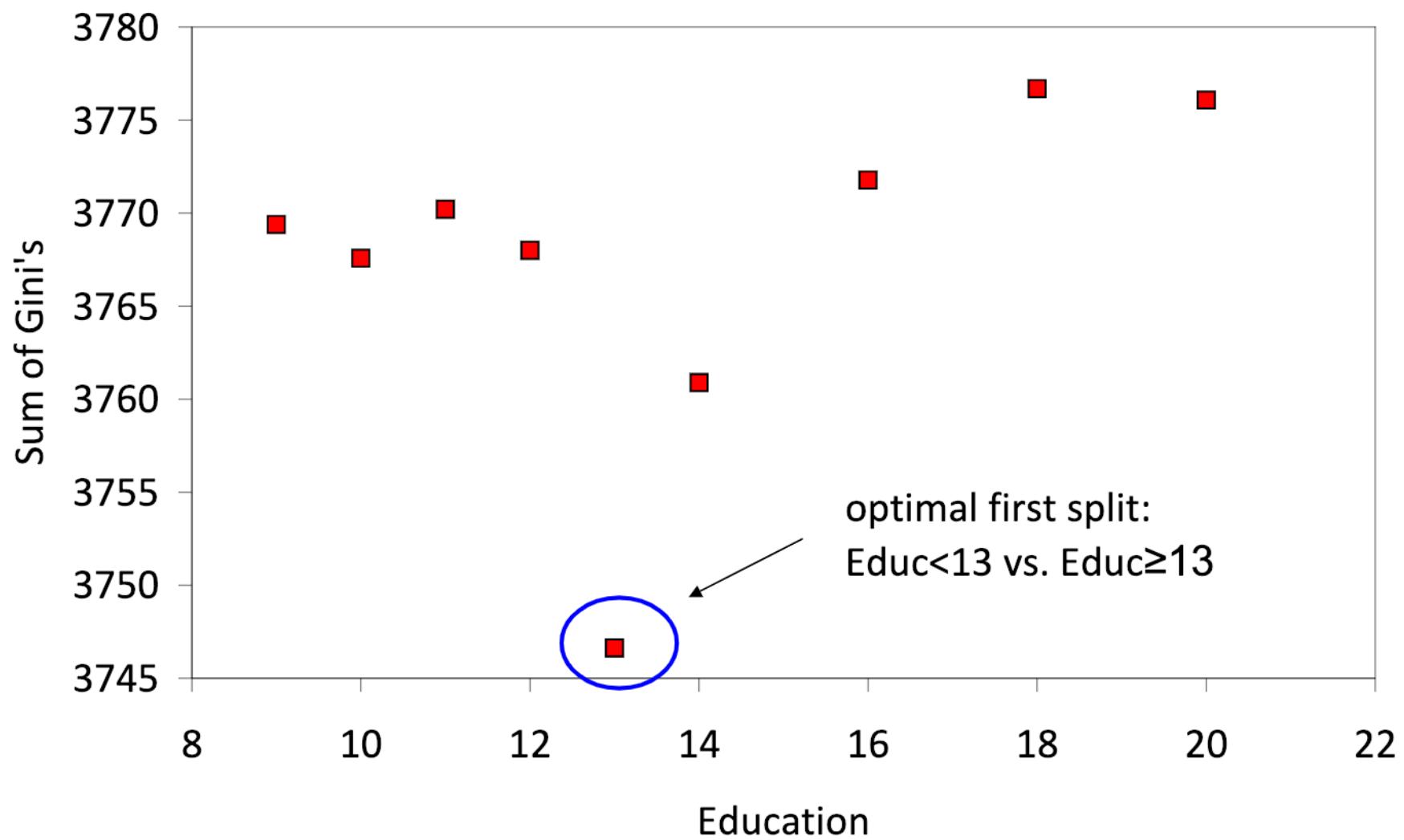
$$\begin{aligned}Ent_1 &= N_1[p_1\log(p_1) + (1 - p_1)\log(1 - p_1)] \\Ent_2 &= N_2[p_2\log(p_2) + (1 - p_2)\log(1 - p_2)]\end{aligned}$$

Notice that if the tree ends up perfectly classifying cases in subnode j , then $\hat{p}_j = 0$ or $\hat{p}_j = 1$ and both Gini and Entropy are 0 for that node.

Choice of Split Point: Predicting Female Gender Using Education

Education	Fraction Female	Fraction Female		Gini Index		Sum of Ginis
		Below Cut	At or Above Cut	Below Cut	At or Above Cut	
8	34.9					
9	38.4	34.9	48.6	114.5	3654.9	3769.4
10	49.2	35.8	48.7	156.5	3611.1	3767.6
11	40.5	38.7	48.7	206.9	3563.3	3770.2
12	42.7	39.3	49.0	297.0	3471.0	3768.0
13	51.5	41.9	51.5	1288.7	2457.9	3746.6
14	54.0	44.8	51.7	1929.0	1831.9	3760.9
16	50.9	46.5	50.9	2386.0	1385.8	3771.8
18	54.6	47.7	51.1	3282.9	493.8	3776.7
20	41.4	48.4	41.4	3647.1	129.0	3776.1

Sum of Gini's for Alternative Split Points



- the “tree” algorithm:
 1. find x_j and cutpoint s that leads to the best “split” of the observations, based on Gini/Entropy
 2. for the subset of observations on each branch: repeat step 1.
Select the next split that leads to the biggest reduction in Gini
 3. repeat step 2 until a minimum fraction of the data is in each node

Example: Titanic data set (Hal Varian JEP)

n=1309 passengers

outcome = 1 if survived (average rate = 38.2%)

predictors=age (missing in some cases), sex, cabin class

Survival rate by group:

-1st class cabin 61.9%

-2nd class cabin 43.0%

-3rd class cabin 25.5%

Survival rates by group:

-young children (≤ 10) 58.1%; other ages 36.8%

-females 72.8%; males 19.1%

-BUT:

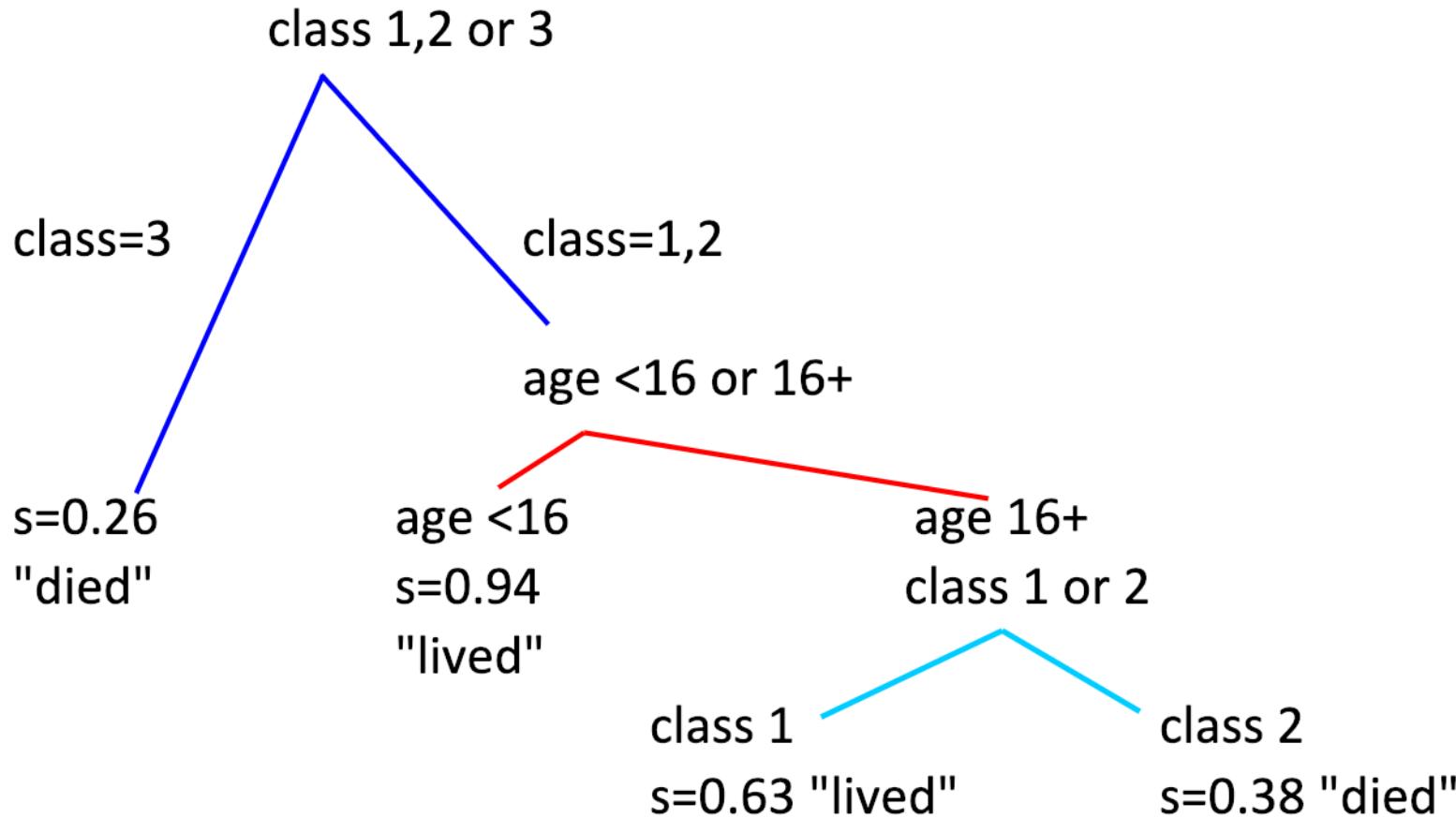
97% of females in 1st class survived

89% of females in 2nd class survived

49% of females in 3rd class survived

Hal presents a decision tree using only age and passenger class.

Titanic Survivor Decision Model Based on Age and Cabin Class



Extensions:

- 1) Pruning. In many situations the tree will be “over-fit” (just like a regression with lots of interactions). Pruning is a way to penalize the number of branches. So you minimize $Gini + \alpha T$ where T is # terminal nodes, and select α by CV.
- 2) Bagging (bootstrap averaging). We select a series of “bootstrap samples”, selected by drawing with replacement from the original sample. On average, about $2/3$ of the observations are in a given BS. The idea then is to select sample b , fit a tree (no pruning), get predicted classification for the OoB obs in this replication, ... After B replications there are approximately $B/3$ predicted classifications for each obs. Take the “majority vote”.
- 3) Random forest. Same as bagging, but in each replication, also randomly select \sqrt{J} of the x 's to be used to build the tree.

Lecture 24 – Comparing Tree Models and Logits

1. tree models - quick review
2. random forest
3. C-section data from California

Tree models

- we are interested in predicting a response y_i given a vector of predictors x_i
- the idea of a tree model is to divide the predictor space into a partition of “retangles” $R_1, R_2 \dots R_J$ then use as a predictor for any observation that falls into retangle R_j either the mean of the y 's (for a continuous outcome) or the most common outcome (for a discrete outcome)
- the rectangles are defined by *recursive binary splitting*, following a “tree” process:
 - first split by whether x_3 (say) is above or below some cutoff
 - continue splitting branches until some criterion is met

When the outcome of interest is discrete, “tree” tries to minimize “impurity” at each node. At node j

$$Gini(j) = 2\hat{p}_j(1 - \hat{p}_j)$$

where \hat{p}_j is the fraction of cases at node j that are 1's. Notice that if the tree ends up perfectly classifying cases, then $\hat{p}_j = 0$ or $\hat{p}_j = 1$ and Gini is 0.

- the “tree” function in R uses the following algorithm to “build” a classification tree
 1. find x_j and cutpoint s that leads to a 2-group model with lowest Gini at the 2 nodes ($N_1Gini(1) + N_2Gini(2)$)
 2. find a split of one of the branches that leads to the greatest reduction in Gini
 3. repeat step 2 until a minimum fraction of the data is in each region

Bagging (bootstrap averaging). We select a series of “bootstrap samples”, selected by drawing with replacement from the original sample. On average, about $2/3$ of the observations are in a given BS. The idea then is to select sample b , fit a tree (no pruning), get predicted classification for the out of the bag (OoB) obs in this replication, ... After B replications there are approximately $B/3$ predicted classifications for each obs. Take the “majority vote” for each observation.

Random forest.

This is an extension of bagging: in each replication, also randomly select \sqrt{J} of the x 's to be used to build the tree.

The advantage is that the models on the replication samples are forced to be different – so we don't end up with too much similarity in the trees constructed in each replication.

As in bagging, the tree model fit on each replication sample is then applied to the OoB observations and each is assigned a classification. Then we take a vote across replications.

Recall: once we have predicted the classes for each observation, we can calculate mis-classification rates. In the 2-group case we have

among true 1's: "false negative rate" = fraction predicted 0's

among true 0's: "false positive rate" = fraction predicted 1's

		Predicted Status	
		Positive	Negative
Actual Status:	Positive	true pos.	false neg.
	Negative	false pos.	true neg.

False pos. rate = #false pos./#actual neg.

False neg. rate = #false neg./#actual pos.

Data: 78,416 births in CA, 2007-2011

C-section rate = 31.9%

Data on:

Mom's age mean=31, range 16-45 (trim)

Mom's weight at delivery: mean =173, range 121-239 (trim)

Mom's height mean=64 inches

single/multiple birtht (mult=3%)

breech (2.9%)

1st birth (39%)

2nd+ no prev CS (46%),

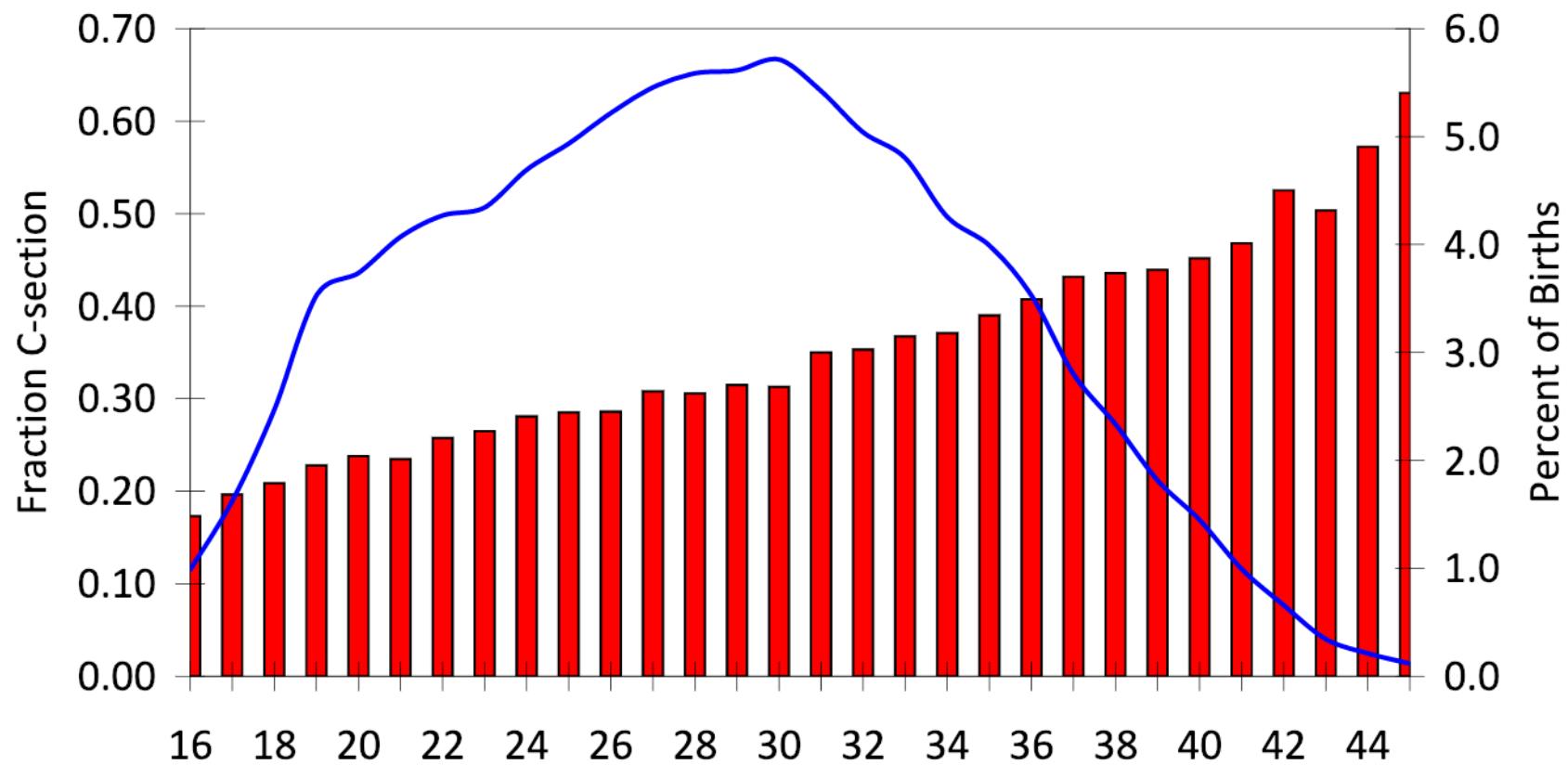
2nd+ prev CS (15%)

education, race, presence of father....

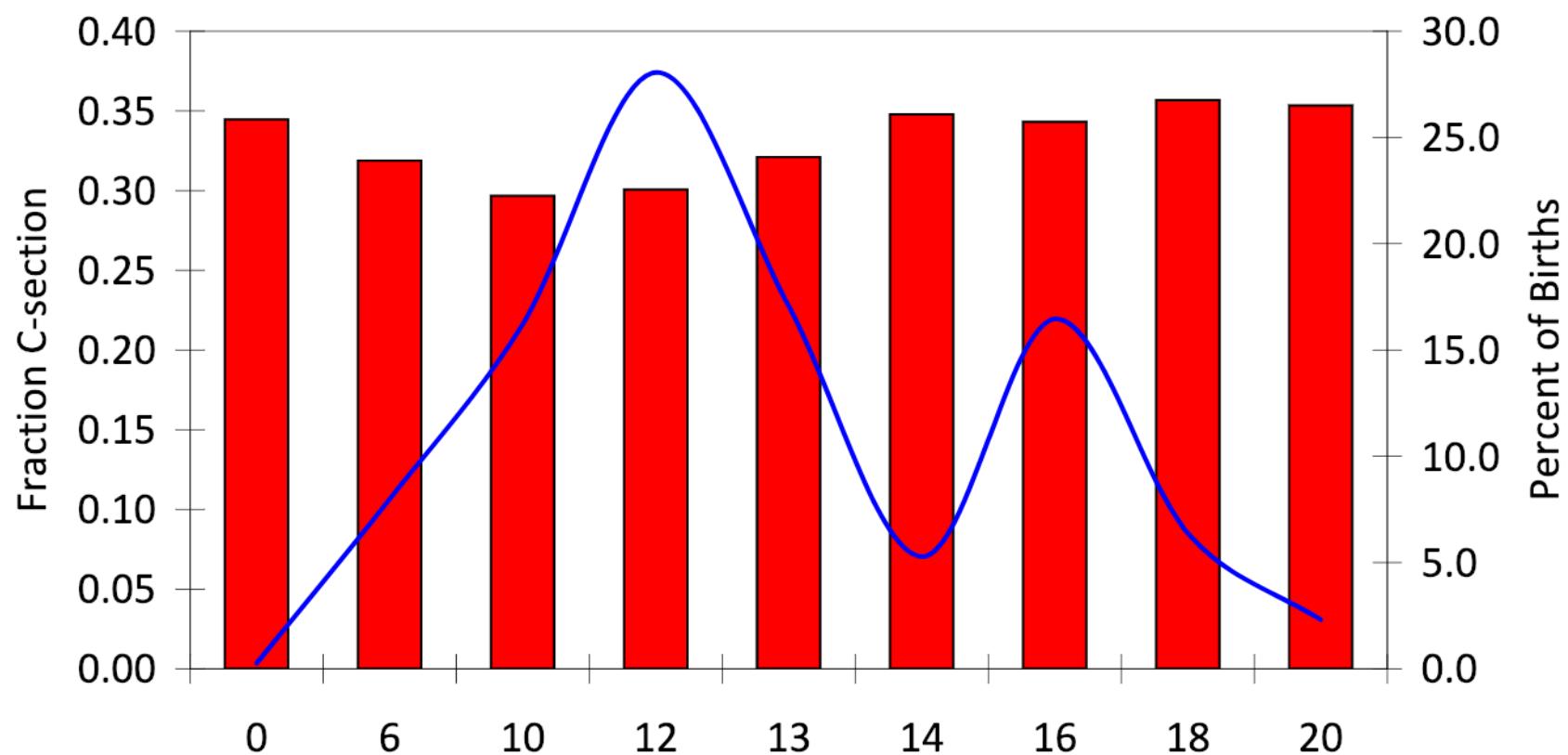
day of the week

mom's conditions (diabetes, hypertension...)

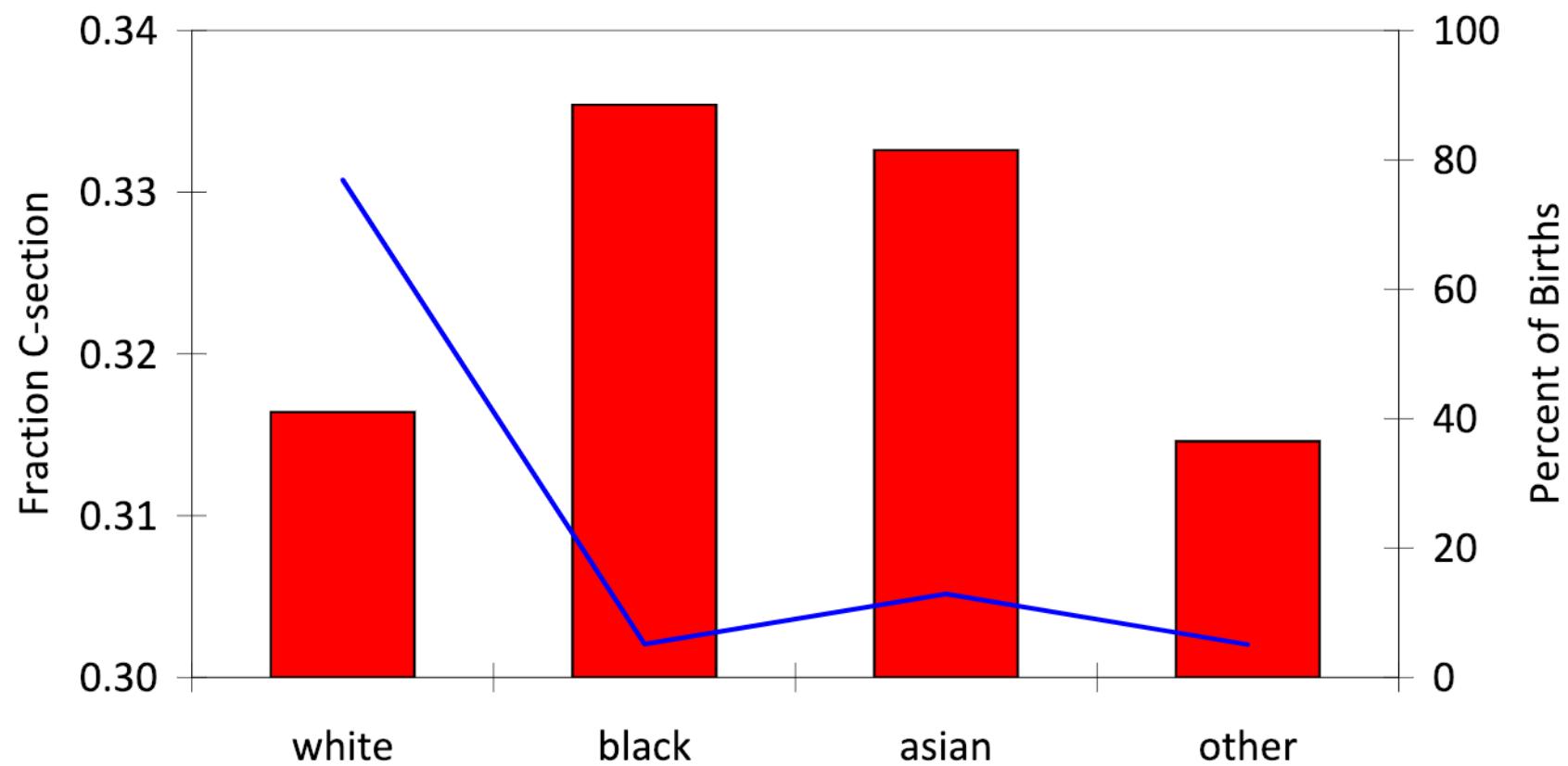
C-Section Rates by Mother's Age



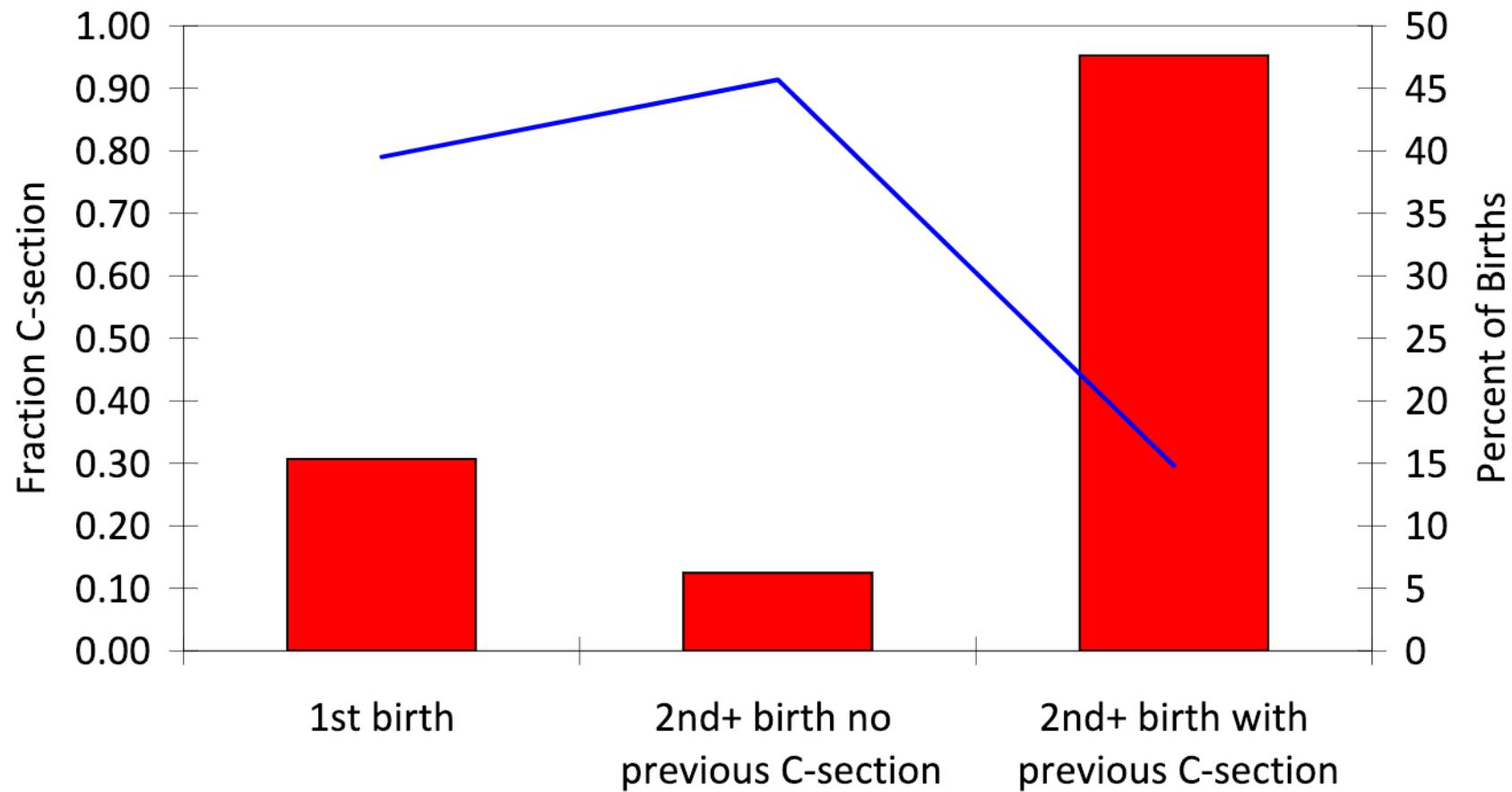
C-Section Rates by Mother's Education



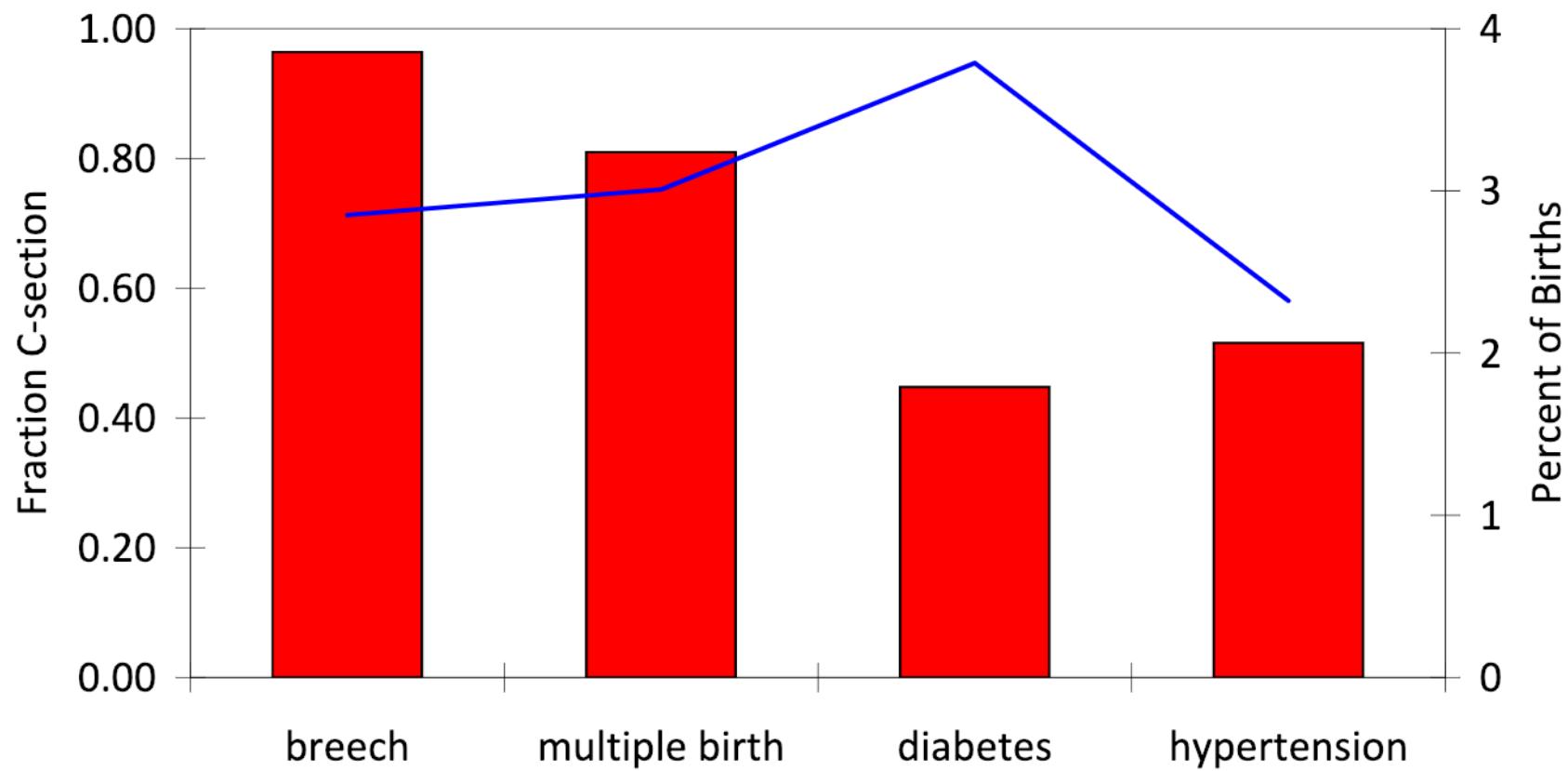
C-Section Rates by Mother's Race



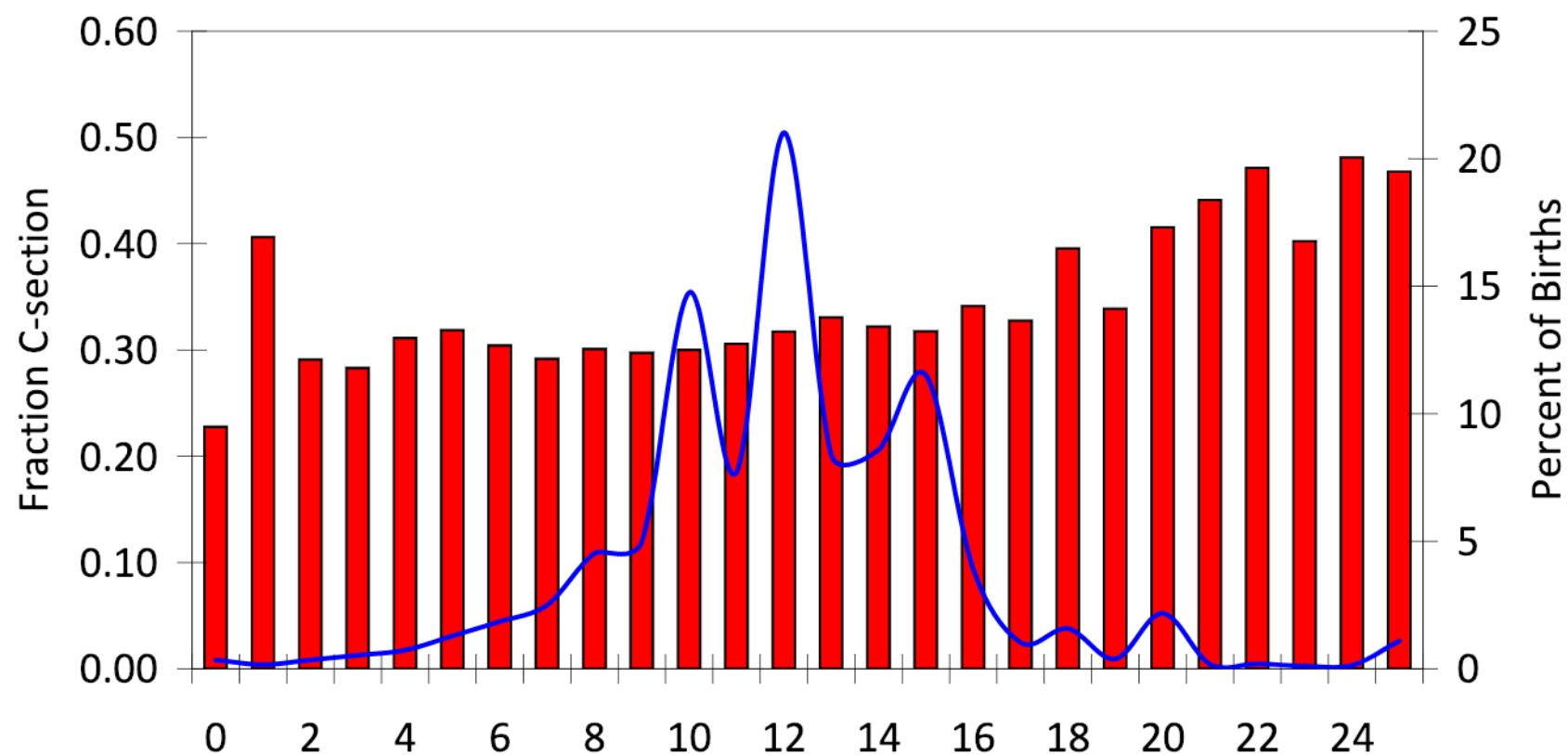
C-Section Rates by Parity/History



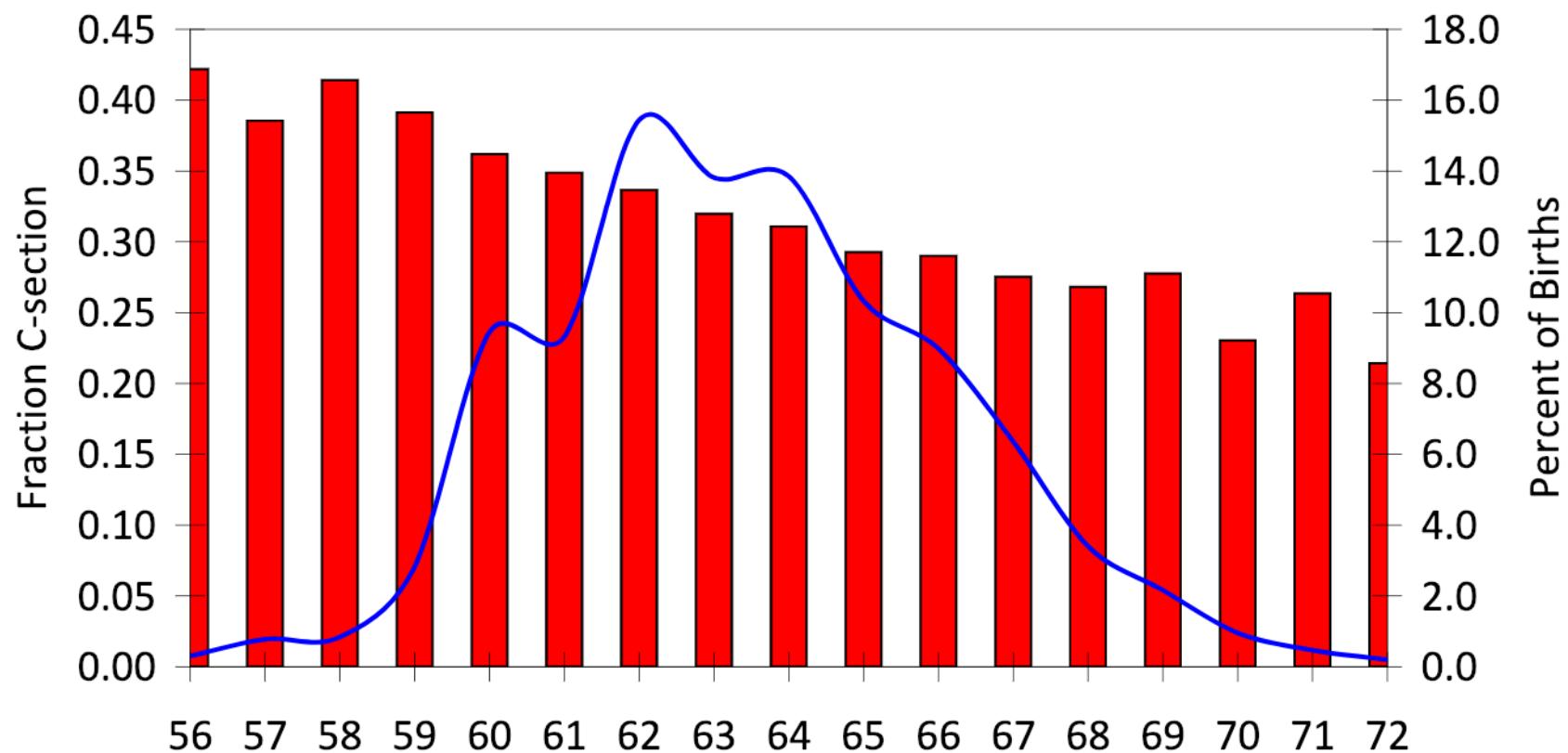
C-Section Rates for Major Complications



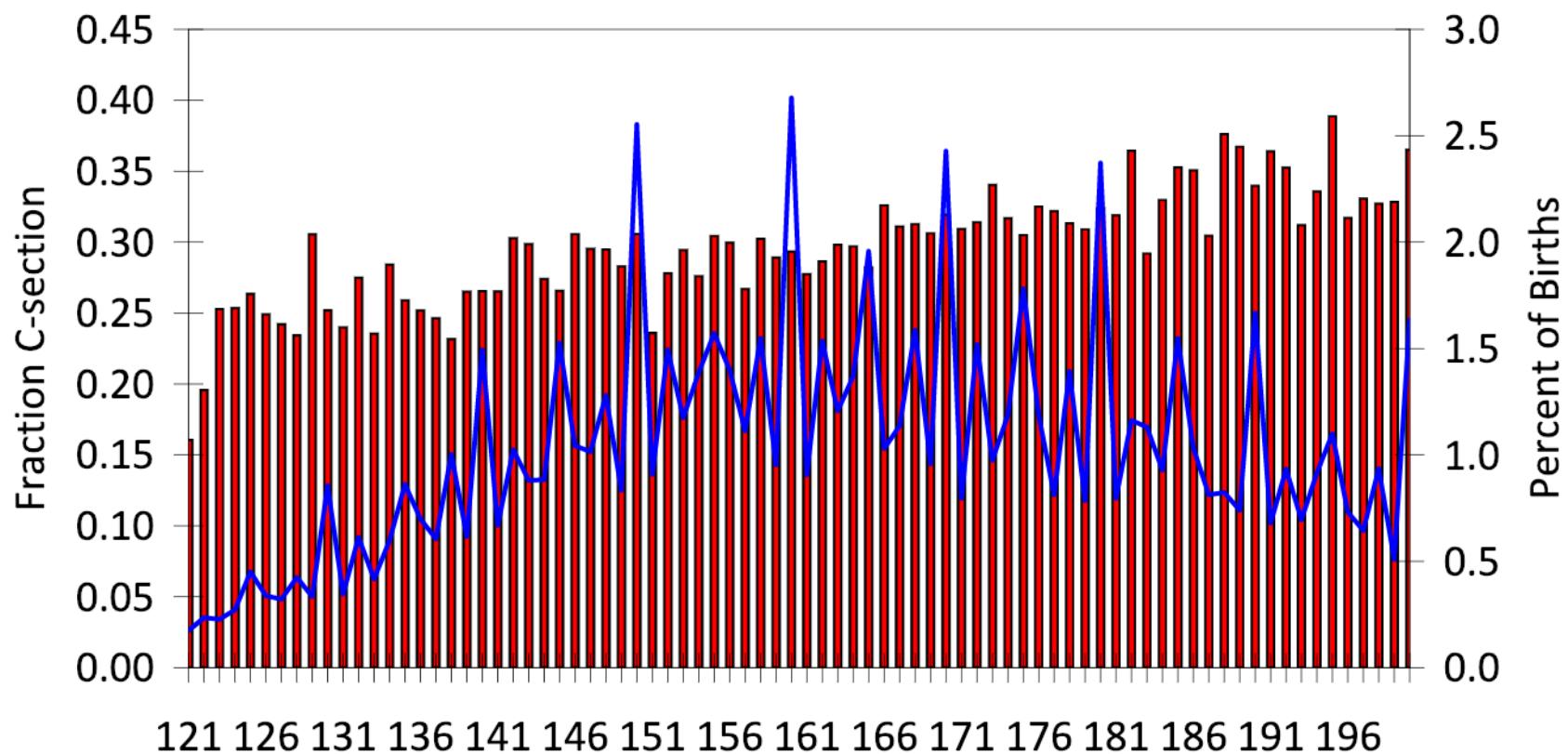
C-Section Rates by Number of Pre-natal Visits



C-Section Rates by Mother's Height (inches)



C-Section Rates by Mother's Weight (pounds)

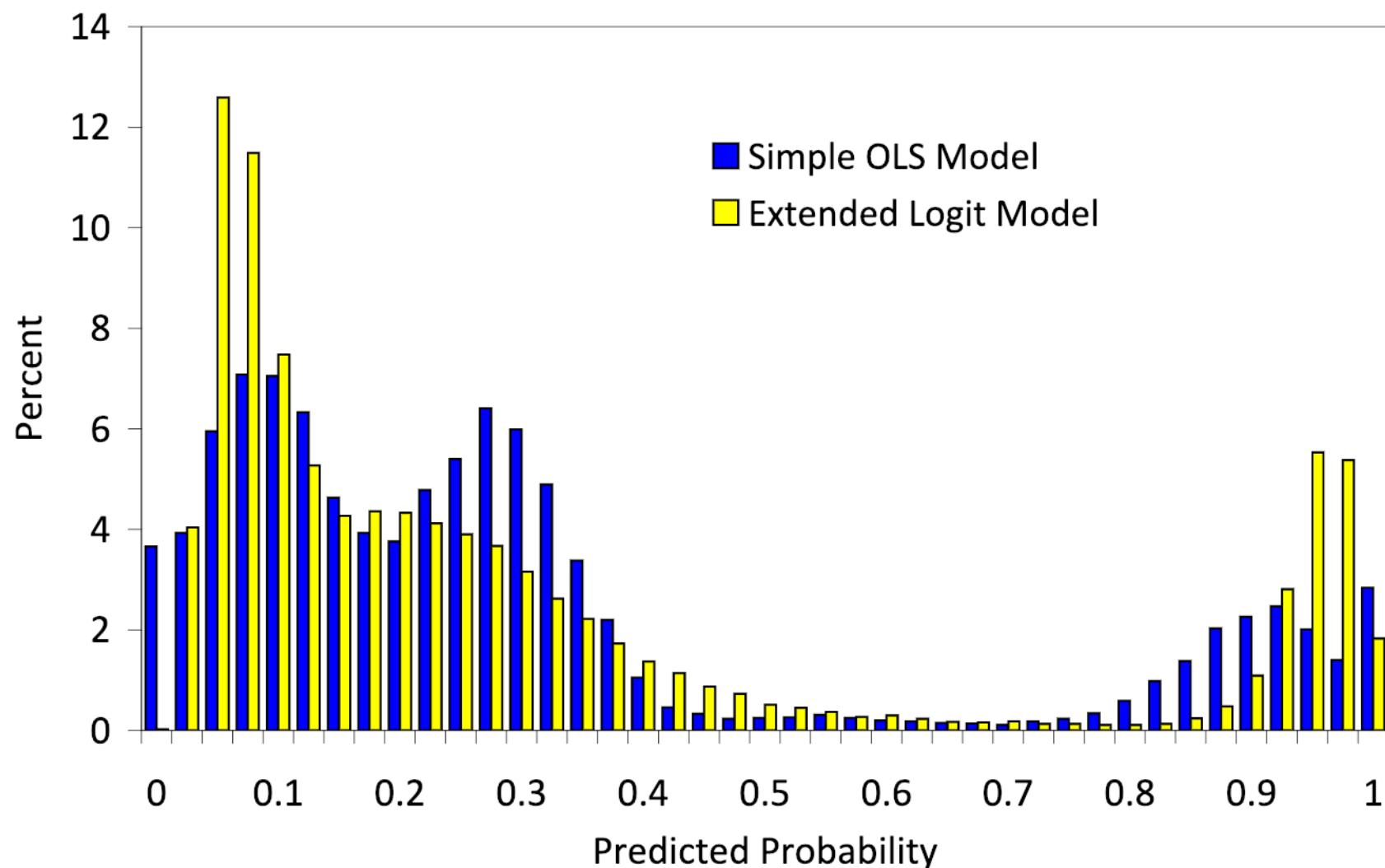


Comparison of Simple and Complex OLS and Logit Models for C-Section

	OLS Models		Logit Models		Ratio logit/OLS
Second+ Birth	-0.228 (0.004)	-0.232 (0.004)	-1.770 (0.035)	-1.828 (0.038)	7.88
No prev C-S					
Second+ Birth	0.577 (0.006)	0.569 (0.006)	3.570 (0.065)	3.541 (0.066)	6.22
Previous C-S					
Breech	0.525 (0.011)	0.523 (0.011)	4.778 (0.181)	4.790 (0.181)	9.16
Multiple Births	0.421 (0.011)	0.408 (0.011)	3.061 (0.087)	3.029 (0.089)	7.42
Log Height	-0.843 (0.043)	-0.779 (0.046)	-7.250 (0.365)	-6.549 (0.393)	8.41
Log Weight	0.239 (0.012)	0.195 (0.013)	2.068 (0.104)	1.735 (0.112)	8.90
Mother Age/100	0.755 (0.030)	--	6.090 (0.247)	--	
Other Stuff*	no	yes	no	yes	
R-squared	0.446	0.452			

Other stuff=age dummies, education dummies, controls for race, insurance, weight gain, diabetes, hypertension
 number of pre-natal visits, day of week of birth, presence of father, dummies for month of pre-natal care initiation

Histogram of Predicted Probabilities from OLS and Logit Models



Comparison of Fit: OLS, Logit

	OLS Models		Logit Models	
Mean Predicted Probability	0.317	0.317	0.317	0.317
Fraction of 1's	0.186	0.187	0.199	0.206
correl(y,y-hat)	0.634	0.634	0.639	0.639
Fraction of 0's correctly predicted	0.983	0.982	0.976	0.971
Fraction of 1's correctly predicted	0.546	0.548	0.572	0.584
<u>Misclassification Fractions:</u>				
y=1, predicted 0	0.145	0.145	0.137	0.133
y=0, predicted 1	0.012	0.012	0.017	0.020

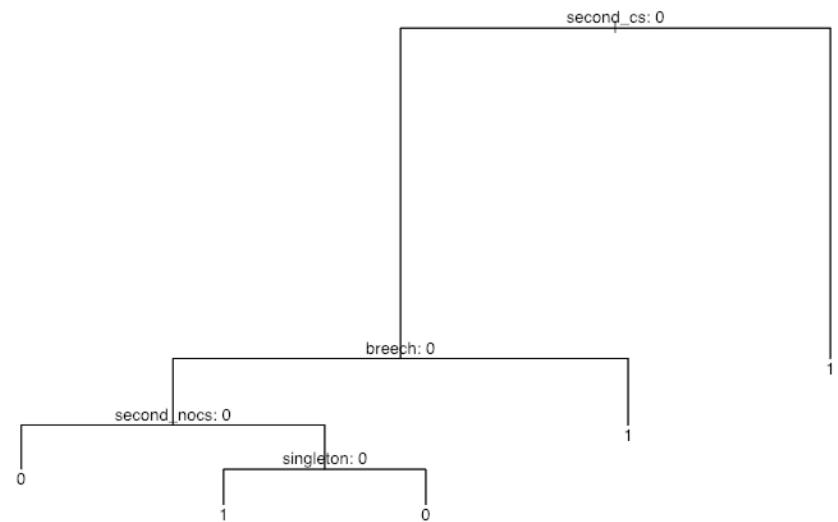
Simple Tree

Load up the basic ISLR tree package

```
library(tree);

set.seed(1);
train = sample(1:nrow(data), nrow(data)/2);
tree.csection.simple <- tree(c_section ~ second_nocs + sec-
                             breech + singleton +
                             mage + logmheight +
                             logmweight_pre + logmweight.
                             gain,
                             data = data[train,]);
```

Simple Tree

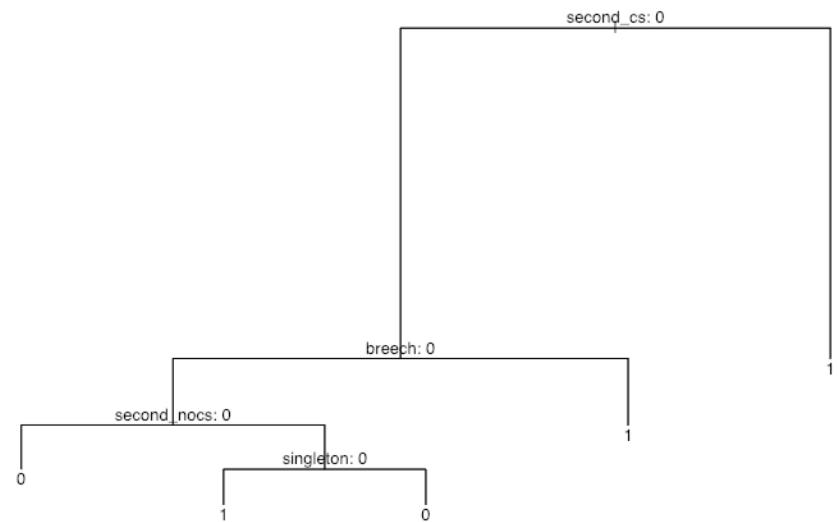


Complex Tree

```
tree.csection <- tree(c_section ~ .,
                      data = data[train,]);
summary(tree.csection);

##  
## Classification tree:  
## tree(formula = c_section ~ ., data = data[train, ])  
## Variables actually used in tree construction:  
## [1] "second_cs"    "breech"        "second_nocs" "singleto  
## Number of terminal nodes: 5  
## Residual mean deviance: 0.7836 = 30720 / 39200  
## Misclassification error rate: 0.156 = 6116 / 39206
```

Complex Tree



Comparison of Fit: OLS, Logit, Tree

	OLS Models	Logit Models	Tree Model	
Mean Predicted Probability	0.317	0.317	0.317	0.317
Fraction of 1's	0.186	0.187	0.199	0.206
correl(y,y-hat)	0.634	0.634	0.639	0.639
Fraction of 0's correctly predicted	0.983	0.982	0.976	0.971
Fraction of 1's correctly predicted	0.546	0.548	0.572	0.584
<u>Misclassification Fractions:</u>				
y=1, predicted 0	0.145	0.145	0.137	0.133
y=0, predicted 1	0.012	0.012	0.017	0.020

Random Forests

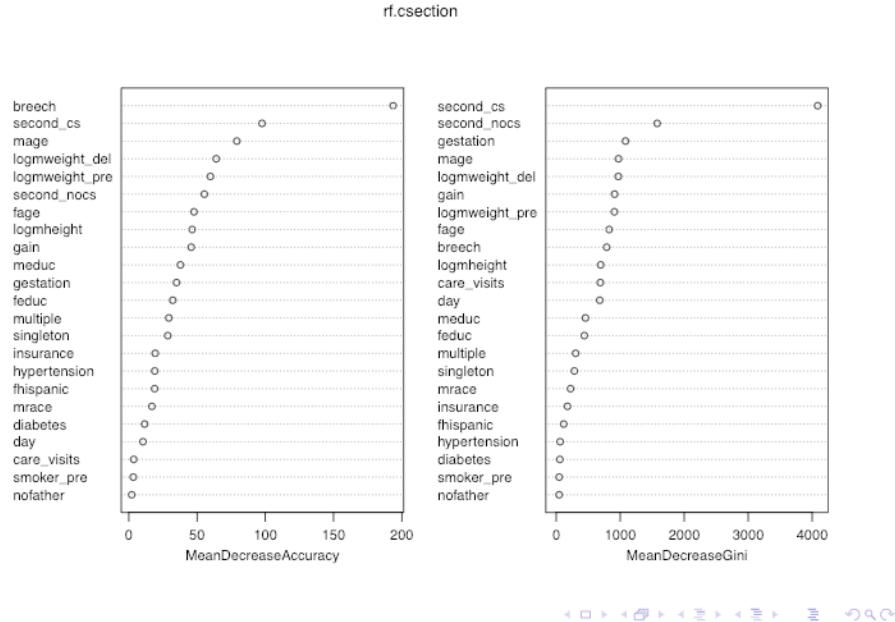
Initialise Train and Test Sets

```
library(randomForest);

## randomForest 4.6-10
## Type rfNews() to see new features/changes/bug fixes.

set.seed(1);
rf.csection = randomForest(c_section ~ .,
                            data,
                            subset = train,
                            mtry = 4,
                            importance = TRUE);
```

Variable Importance



Comparison of Fit: OLS, Logit, Tree, and Random Forest

	OLS Models	Logit Models		Tree Model	Random Forest
Mean Predicted Probability	0.317	0.317	0.317	0.317	--
Fraction of 1's	0.186	0.187	0.199	0.206	0.186
correl(y,y-hat)	0.634	0.634	0.639	0.639	0.630
Fraction of 0's correctly predicted	0.983	0.982	0.976	0.971	0.982
Fraction of 1's correctly predicted	0.546	0.548	0.572	0.584	0.544
<u>Misclassification Fractions:</u>					
y=1, predicted 0	0.145	0.145	0.137	0.133	0.146
y=0, predicted 1	0.012	0.012	0.017	0.020	0.012

0.263

0.821

0.958

0.790

0.067

0.010

Classification Rates: Logit vs RF

	True C-Section Status	
	0	1
<u>Classified by Extended Logit</u>		
Classified as 0	51,831 (66.1)	10,427 (13.3)
Classified as 1	1,541 (2.0)	14,613 (18.6)
<u>Classified by Random Forest</u>		
Classified as 0	52,571 (67.0)	5,250 (6.7)
Classified as 1	801 (1.0)	19,790 (25.2)

Characteristics of Cases Reclassified by Random Forest

	All Births	C-S=1 Misclass.	C-S=1 Misclass. By Logit	
		By Logit	RF=0	RF=1
C-Section	0.319	1.000	1.000	1.000
P-hat (logit)	0.319	0.251	0.244	0.258
Mean RF Pred	0.262	0.514	0.000	1.000
Breech	0.029	0.000	0.000	0.000
Second+ Birth w/ CS	0.148	0.000	0.000	0.000
Second+ Birth no CS	0.457	0.280	0.285	0.275
Multiple Births	0.030	0.007	0.000	0.014
Mother's Age	28.360	28.150	28.070	28.230
Log Height	4.149	4.148	4.147	4.147
Mother's Education	12.670	13.040	13.050	13.020
Diabetes	0.038	0.040	0.043	0.038
Number	78,412	10,427	5,068	5,359