Chapter 2: Selection on Observables

Peter Hull

Applied Econometrics II
Brown University
Spring 2024

Outline

1. Addressing Selection Bias with Controls

2. Matching, Weighting, and Regression

3. Application: Kline and Moretti (2014)

4. Bonus Track: Coefficient Stability

Selection Bias, Formalized

As we've seen, identifying causal effects is easy in an experiment, where $D_i \perp (Y_i(0), Y_i(1))$. Regressing Y_i on D_i is one simple way to go

- But most "treatments" we care about cannot be easily randomized

To make the "selection bias" challenge as clear as possible, we start with a constant effects causal model: $Y_i = \beta D_i + \varepsilon_i$ where β is the causal effect

- E.g., for binary D_i, we have $\beta = Y_i(1) - Y_i(0)$ and $\varepsilon_i = Y_i(0)$

$Y_i = \beta D_i + \varepsilon_i$ looks like a regression, but it very well could not be!

$$\frac{Cov(D_i, Y_i)}{Var(D_i)} = \frac{Cov(D_i, \beta D_i + \varepsilon_i)}{Var(D_i)} = \beta + \frac{Cov(D_i, \varepsilon_i)}{Var(D_i)}$$

which equals β if and only if $Cov(D_i, \varepsilon_i) = 0$ (sound familiar?)

- For binary D_i, this is
  $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \beta + E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0]$
- Easy to tell stories for $Cov(D_i, \varepsilon_i) \neq 0$, biasing regression up/down

Selection on Observables with Constant Effects

Our usual first line of defense against selection bias is to add controls

- As it turns out, figuring out when/how this "works" is harder than it seems! So let's start simple, with the constant-effects model

Selection on Observables (take 1): $E[\varepsilon_i \mid D_i, X_i] = E[\varepsilon_i \mid X_i]$ for some X_i

- I.e. untreated potential outcomes are mean-independent of treatment given a vector of observed covariates X_i

To use this assumption, consider the CEF

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

If we further assume $E[\varepsilon_i \mid X_i]$ is linear, we have a linear CEF with D_i carrying a coefficient of $\beta \implies$ *reg y d x, r* estimates β!

- Linearity is not so restrictive since X_i can be "flexible" (eg. saturated)

Illustration: Dale and Krueger (2002)

Dale and Krueger are interested in the effect of attending a more selective college on adult earnings

- They have data on earnings, schooling, and some information on application / admissions from the College and Beyond Survey

The stylized (MHE) version of DK'02 assumes selection-on-observables

- Here Y_i = earnings, D_i = private college, and X_i = group indicators for which colleges student i applied and was admitted to
- The idea is that most of the "important" selection (i.e. relevant to ε_i) into private school is captured by the application/admissions processes

Taking this at face value for now, the results are pretty striking...

Stylized Dale and Krueger

	No Selection Controls			Selection Controls		
	(1)	(2)	(3)	(4)	(5)	(6)
Private School	0.135	0.095	0.086	0.007	0.003	0.013
	(0.055)	(0.052)	(0.034)	(0.038)	(0.039)	(0.025)
Own SAT score/100		0.048	0.016		0.033	0.001
		(0.009)	(0.007)		(0.007)	(0.007)
Predicted log(Parental Income)			0.219			0.190
			(0.022)			(0.023)
Female			-0.403			-0.395
			(0.018)			(0.021)
Black			0.005			-0.040
			(0.041)			(0.042)
Hispanic			0.062			0.032
			(0.072)			(0.070)
Asian			0.170			0.145
			(0.074)			(0.068)
Other/Missing Race			-0.074			-0.079
			(0.157)			(0.156)
High School Top 10 Percent			0.095			0.082
			(0.027)			(0.028)
High School Rank Missing			0.019			0.015
			(0.033)			(0.037)
Athlete			0.123			0.115
			(0.025)			(0.027)
Selection Controls	N	N	N	Y	Y	Y

Notes: Columns (1)-(3) include no selection controls. Columns (4)-(6) include a dummy for each group formed by matching students according to schools at which they were accepted or rejected. Each model is estimated using only observations with Barron's matches for which different students attended both private and public schools. The sample size is 5,583. Standard errors are shown in parentheses.

Two Practical Takeaways

When making selection-on-observable arguments, it is useful to have some "held out" covariates W_i to validate ("check for balance")

- Here once X_i is controlled for, adding own SAT score/parental income/demographics doesn't change the estimate
- We could also directly check whether *reg w d x, r* gives a zero on *d*

The OVB formula can help diagnose concerns w/selection-on-observables

- We think private school students are "positively selected" (higher ε_i) relative to others, as evidenced by cols 1-3
- If this is true given X_i, the true effect can only be more negative!
- We'll see more sophisticated versions of this argument soon...

Outline

1. Addressing Selection Bias with Controls✓
2. Matching, Weighting, and Regression
3. Application: Kline and Moretti (2014)
4. Bonus Track: Coefficient Stability

Relaxing the Model

Constant effects is a strong assumption; we'd like to avoid it when possible

- Seems likely that effects vary across both observables & unobservables
- For binary Y_i (or other limited support), constant effects is impossible

Assuming binary D_i, let's return to the general potential outcomes model:

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

Selection on Observables (take 2): $(Y_i(0), Y_i(1)) \perp D_i \mid X_i$ for some X_i

- I.e. treatment is "as-good-as-randomly assigned," given X_i
- Implies the "take 1" version (independence $\rightarrow$ mean-independence)

What can we do with this?

CATEs and Matching

A conditional version of the RCT identification result we've seen tells us:

$$E[Y_i \mid D_i = 1, X_i] - E[Y_i \mid D_i = 0, X_i] = E[Y_i(1) - Y_i(0) \mid X_i] \equiv CATE(X_i)$$

for the *conditional average treatment effect* function, $CATE(x)$

The LIE further tells us:

$$ATE = E[Y_i(1) - Y_i(0)] = E[E[Y_i(1) - Y_i(0) \mid X_i]] = E[CATE(X_i)]$$

So to estimate the ATE under selection on observables (take 2), we can:

1. Match treated and control observations with the same value of X_i
2. Estimate $E[Y_i \mid D_i = 1, X_i = x] - E[Y_i \mid D_i = 0, X_i = x]$ for each x
3. Average these estimates together by the marginal distribution of X_i

Simple, right? Not always...

Challenge 1: The Curse of Dimensionality

Matching can be tricky when X_i takes on many values / has many rows

- Unfortunate, b/c richer X_i can make identification more plausible

The *Propensity Score Theorem* of Rosenbaum and Rubin offers a solution:

- Rather than matching on X_i, it's enough to match on the scalar propensity score $p(X_i) = Pr(D_i = 1 \mid X_i)$

Prop: $(Y_i(0), Y_i(1)) \perp D_i \mid X_i$ implies $(Y_i(0), Y_i(1)) \perp D_i \mid p(X_i)$

Proof: Using the LIE,

$$
\begin{aligned}
&Pr(D_i = 1 \mid p(X_i), Y_i(0), Y_i(1)) = E[D_i \mid p(X_i), Y_i(0), Y_i(1)] \\
&= E[E[D_i \mid X_i, p(X_i), Y_i(0), Y_i(1)] \mid p(X_i), Y_i(0), Y_i(1)] \\
&= E[E[D_i \mid X_i] \mid p(X_i), Y_i(0), Y_i(1)] \\
&= E[p(X_i) \mid p(X_i), Y_i(0), Y_i(1)] \\
&= p(X_i) = Pr(D_i = 1 \mid p(X_i))
\end{aligned}
$$

Using P-Scores

If $p(\cdot)$ is known (e.g. in a stratified RCT) we can:

1. Match treated and control observations with the same value of $p(X_i)$
2. Estimate $E[Y_i \mid D_i = 1, p(X_i) = p] - E[Y_i \mid D_i = 0, p(X_i) = p]$ for all p
3. Average these estimates together by the marginal distribution of $p(X_i)$

We can also weight inversely by $p(X_i)$; helpful with many values of p!

$$E\left[\frac{Y_i D_i}{p(X_i)}\right] = E\left[\frac{Y_i(1) D_i}{p(X_i)}\right] = E\left[E\left[\frac{Y_i(1) D_i}{p(X_i)} \mid X_i\right]\right] = E\left[\frac{E\left[Y_i(1) D_i \mid X_i\right]}{p(X_i)}\right]$$
$$= E\left[\frac{E\left[Y_i(1) \mid D_i = 1 X_i\right] Pr(D_i = 1 \mid X_i)}{p(X_i)}\right] = E[E[Y_i(1) \mid X_i]] = E[Y_i(1)]$$

Following the same steps, $E\left[\frac{Y_i(1-D_i)}{1-p(X_i)}\right] = E[Y_i(0)]$. Thus:

$$E[\omega_i Y_i] = ATE \text{ for } \omega_i = \frac{D_i}{p(X_i)} - \frac{1 - D_i}{1 - p(X_i)} = \frac{D_i - p(X_i)}{p(X_i)(1 - p(X_i))}$$

Using Estimated P-Scores

In most settings we must estimate $p(X_i) = E[D_i \mid X_i]$, e.g. by OLS/probit

- This mostly works as you'd expect, under appropriate regularity conditions, though there are some subtleties (outside this class' scope)
- See e.g. Abadie and Imbens (2016) for matching and Hirano, Imbens, and Ridder (2003) for weighting (also a large/growing ML literature)

Note that we can not only estimate the ATE with p-scores, but also any weighted average of $CATE(X_i)$. E.g.:

- ATT (aka TOT): $E[Y_i(1) - Y_i(0) \mid D_i = 1] = E[CATE(X_i) \mid D_i = 1]$
- ATU (aka TNT): $E[Y_i(1) - Y_i(0) \mid D_i = 0] = E[CATE(X_i) \mid D_i = 0]$

these only differ from ATE under (observable) "selection-on-gains": i.e. when $p(X_i)$ and $CATE(X_i)$ are correlated:

- E.g. $E[CATE(X_i) \mid D_i = 1] = \frac{E[CATE(X_i)D_i]}{E[D_i]} = E\left[\frac{p(X_i)}{E[p(X_i)]} CATE(X_i)\right]$

Challenge 2: Limited Overlap

The ATE is only identified when $p(X_i)$ is bounded away from zero and one

- Intuitively, can't identify effects at X_i where $D_i = 0$ or $D_i = 1$ always

ATE estimators are likely to be very noisy of $p(X_i)$ is ever near zero or one

- Intuitively, need a lot of data to estimate effects at such X_i

The finite-sample performance of ATE estimators under limited overlap can be improved by "trimming" propensity scores near 0 and 1

- Trimming in large samples changes the estimand, from ATE to a weighted-average $CATE(X_i)$ among X_i with non-trimmed $p(X_i)$
- Is this "Moving the Goal Posts"? (Crump et al. 2009)

Challenge 3: Just Kinda Annoying?

P-score matching/weighting can be a little involved, esp. if you want SEs

- Some packages exist (e.g. *teffects* in Stata) but they're a bit finicky
- ATE estimates are often quite noisy, even if overlap is decent

The previous ("take 1") version of selection-on-observables suggests regression might provide a tractable alternative ...

- *Q:* What does *reg y d x, r* identify when $(Y_i(1), Y_i(0)) \perp D_i \mid X_i$?
- *A:* A convex weighted average of $CATE(X_i)$, as long as $p(X_i)$ is linear!

Angrist (1998): Regression Meets Matching

Prop.: Suppose $(Y_i(1), Y_i(0)) \perp D_i \mid X_i$ for a vector of group indicators X_i. The coefficient on D_i in a regression of Y_i on D_i and X_i identifies:

$$\beta = E\left[\omega(X_i) CATE(X_i)\right]$$

where $\omega(x) = \frac{Var(D_i|X_i=x)}{E[Var(D_i|X_i)]} \geq 0$, with $E[\omega(X_i)] = 1$

- Regression weights TEs by the conditional variance of treatment

If $CATE(X_i)$ or $\omega(X_i)$ are constant, weighting is irrelevant: $\beta = ATE$

- More generally, $\beta = ATE$ if $CATE(X_i)$ and $\omega(X_i)$ are uncorrelated

Since $Var(D_i \mid X_i) = p(X_i)(1 - p(X_i))$, more weight on X_i w/ $p(X_i) \approx 0.5$

- Intuitively, this is where we have the most information about TEs $\implies$ β is likely more precise than ATE
- No weight on groups where overlap fails (no variation in treatment)

Angrist '98 Proof

By the FWL theorem, β is the coefficient from regressing Y_i on the residuals $\tilde{D}_i$, obtained from regressing D_i on group dummies X_i

Auxiliary regression is saturated, so it gives the p-score $p(X_i) = E[D_i \mid X_i]$. Residuals are group-demeaned $\tilde{D}_i = D_i - E[D_i \mid X_i]$, with $E[\tilde{D}_i \mid X_i] = 0$

Thus, by the LIE,

$$\beta = \frac{E\left[\tilde{D}_i Y_i\right]}{E\left[\tilde{D}_i^2\right]} = \frac{E\left[E\left[\tilde{D}_i(Y_i(0) + (Y_i(1) - Y_i(0))D_i) \mid X_i\right]\right]}{E\left[\left[\tilde{D}_i^2 \mid X_i\right]\right]}$$

$$= \frac{E[E[\tilde{D}_i D_i(Y_i(1) - Y_i(0)) \mid X_i]]}{E[Var(D_i \mid X_i)]} = E[\omega(X_i) CATE(X_i)]$$

(skipping some steps... see MHE for a careful derivation)

Regression vs. Matching in Angrist '98

TABLE 3.3.1
Uncontrolled, matching, and regression estimates of the effects of voluntary military service on earnings

Race	Average Earnings in 1988–1991 (1)	Differences in Means by Veteran Status (2)	Matching Estimates (3)	Regression Estimates (4)	Regression Minus Matching (5)
Whites	14,537	1,233.4	−197.2	−88.8	108.4
		(60.3)	(70.5)	(62.5)	(28.5)
Non-whites	11,664	2,449.1	839.7	1,074.4	234.7
		(47.4)	(62.7)	(50.7)	(32.5)

Notes: Adapted from Angrist (1998, tables II and V). Standard errors are reported in parentheses. The table shows estimates of the effect of voluntary military service on the 1988–91 Social Security–taxable earnings of men who applied to enter the armed forces between 1979 and 1982. The matching and regression estimates control for applicants' year of birth, education at the time of application, and AFQT score. There are 128,968 whites and 175,262 nonwhites in the sample.

Takeaways

Angrist (1998) says that regression identifies a sensible convex average of heterogeneous treatment effects if we have a binary treatment D_i which is as-good-as-randomly assigned given group dummies X_i

- Key to the proof is that *reg d x, r* identifies $p(X_i) = E[D_i \mid X_i]$, so the result generalizes for "flexible" controls which linearly span $p(X_i)$
- Of course $p(X_i)$ could be given by outside knowledge or estimated, then controlled for directly (more on this later)

Angrist and Krueger (1999) generalize this to non-binary treatments:

- If $(Y_i(s))_{s \in Supp(S_i)} \perp S_i \mid X_i$ and $E[S_i \mid X_i]$ is linear, the coefficient on S_i in a regression of Y_i on S_i and X_i identifies a variance-weighted average of potential outcome function derivatives $\frac{\partial}{\partial s} Y_i(s)$

Bottom line: OLS plays well with such selection-on-observable arguments

- At least with a single treatment ... more on multiple treatments soon

Outline

1. Addressing Selection Bias with Controls✓
2. Matching, Weighting, and Regression✓
3. Application: Kline and Moretti (2014)
4. Bonus Track: Coefficient Stability

Motivation: Do "Big Push" Development Strategies Work?

Vast income disparities across regions in the U.S. (and other countries)

- Can place-based economic development programs reduce inequality?
- "Zero sum game" concern: growth in (e.g.) manufacturing in one part of the country shifts economic production from elsewhere

KM study one of the most ambitious place-based economic development policies in U.S. history: the *Tennessee Valley Authority* (TVA)

- Series of large-scale infrastructure investments over 1930-1960: electric dams, new roads, canals, flood control systems, etc...
- At peak, annual federal subsidy was roughly 10% of household income

Did these huge federal investments change the economic trajectory of TV?

- And did any gains come at the cost of reduced growth elsewhere?

TVA Program Background

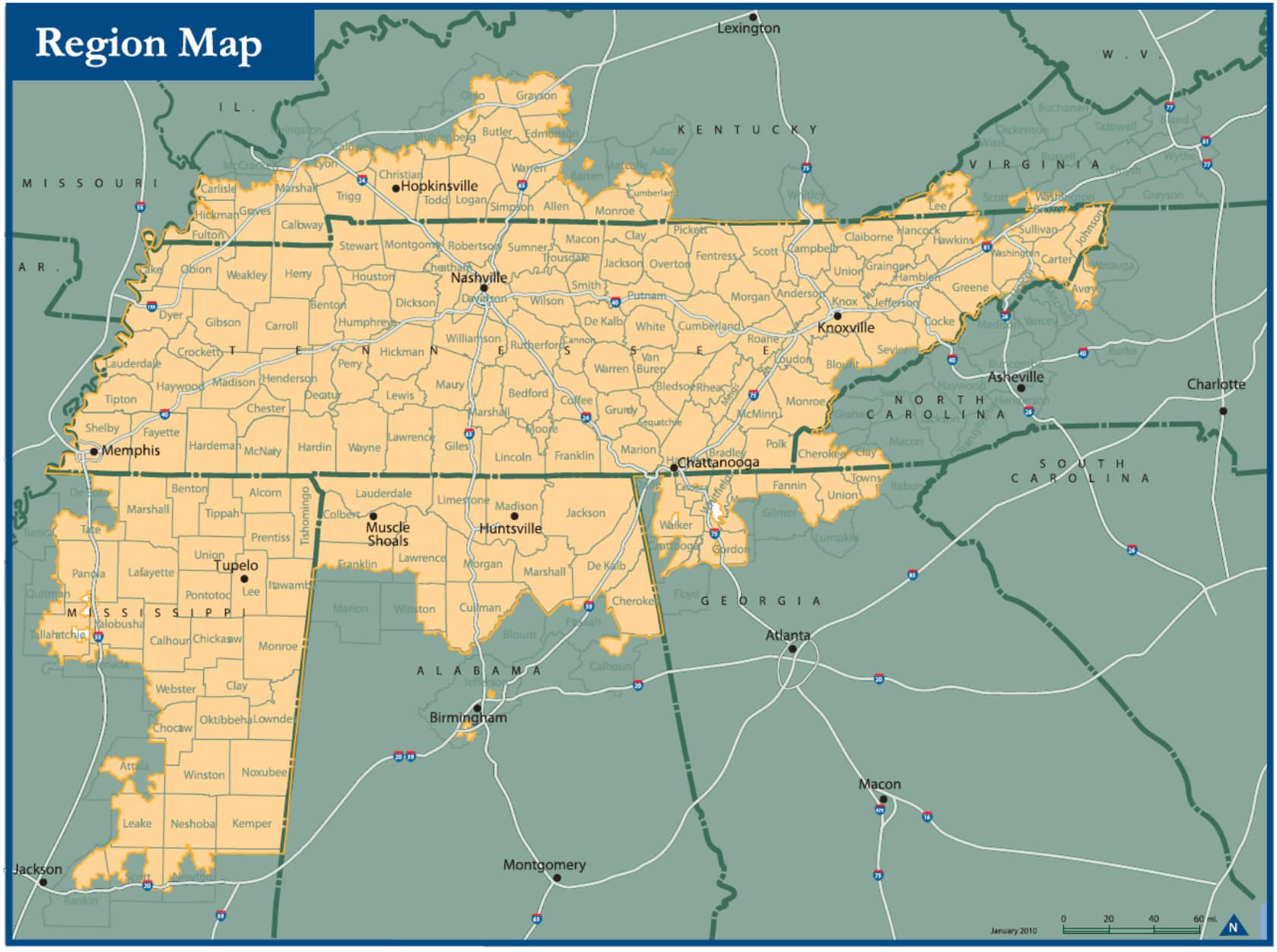
The TVA service area included 163 counties covering basically all of Tennessee, as well as in Kentucky, Alabama, and Mississippi

- Big focus was to increase electricity generation/availability, in part to attract manufacturing to the largely agricultural region
- Around $20 billion in 2000 dollars was appropriated between 1934 and 2000, with around 73% spent over 1940-1958

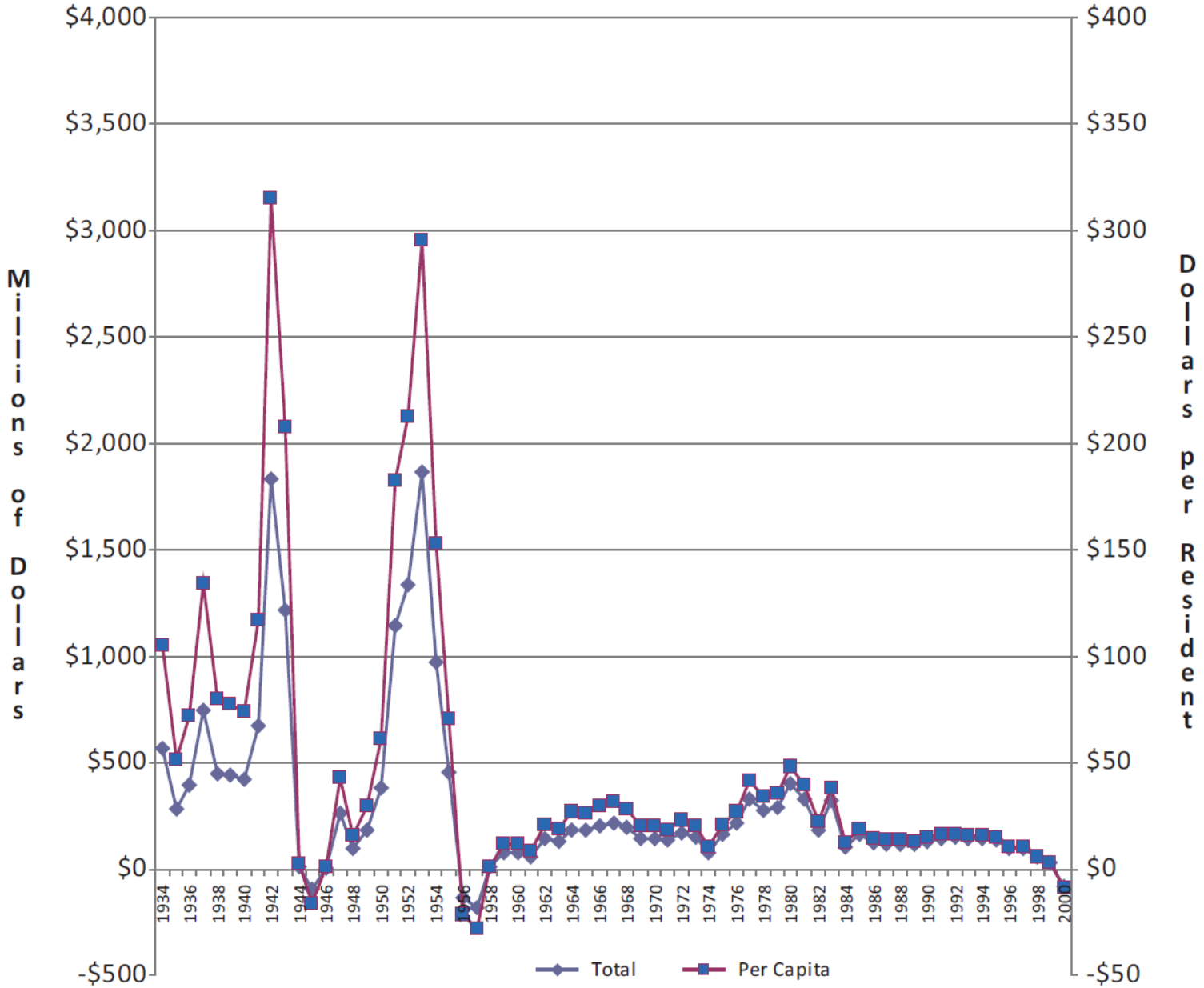
Counties were prioritized for TVA if they were:

- Heavily rural, required electric power, lacked public facilities
- Experienced severe flooding, had large deficits
- Were willing/able to work with TVA authorities
- Were reasonably close to power plants

TVA, Mapped


Region Map
Lexington
Hopkinsville
Nashville
Knoxville
Asheville
Charlotte
Memphis
Chattanooga
Muscle Shoals
Huntsville
Tupelo
Atlanta
Birmingham
Macon
Montgomery
Jackson
KENTUCKY
VIRGINIA
W. V
TENNESSEE
NORTH CAROLINA
SOUTH CAROLINA
GEORGIA
ALABAMA
MISSISSIPPI
MISSOURI
ILL.
AR.
January 2010
0 20 40 60 mi.
N

TVA Spending Over Time



Research Design: Planned-but-Abandoned Programs

TVA was meant to be the first of many regional authorities

- A 1937 senate bill proposed seven new authorities, across U.S. regions
- But political infighting / the start of WW2 halted progress past TVA

KM use the geographic scope of these proposed authorities to define counterfactual counties, similar to TVA but never receiving similar funding

- Six authorities: Atlantic Seabord, Great Lakes-Ohio Valley, Missouri Valley, Arkansas Valley, Columbia, and Western
- Total of 828 counterfactual counties in 25 states

Unsurprisingly, TVA looks different from the rest of the country on observable characteristics and trends

- But things seem more balanced when comparing to non-TVA South or non-TVA proposed authorities

Summary Statistics

	(1) TVA	(2) Non-TVA	(3) Non-TVA South	(4) Non-TVA proposed authorities
1930 characteristics				
Log population	9.991	9.977	9.989	9.940
Log employment	8.942	8.967	8.959	8.908
Log # of houses	8.445	8.508	8.455	8.466
Log average manufacturing wage	1.406	1.802	1.545	1.685
Manufacturing employment share	0.075	0.090	0.080	0.077
Agricultural employment share	0.617	0.455	0.541	0.510
% White	0.813	0.885	0.722	0.830
% Urbanized	0.153	0.280	0.233	0.216
% Illiterate	0.088	0.045	0.092	0.060
% of Whites foreign born	0.002	0.059	0.013	0.020
Log average farm value	5.252	5.646	5.386	5.552
Log median housing value	9.271	9.581	9.360	9.452
Log median contract rent	8.574	9.030	8.679	8.834
% Own radio	0.079	0.296	0.114	0.210
Max elevation (meters)	1,576.190	2,364.531	1,068.943	1,758.893
Elevation range (max–min)	1,127.761	1,521.322	712.336	1,083.293
% Counties in South	1.000	0.342	1.000	0.554

Summary Statistics (Cont.)

	(1) TVA	(2) Non-TVA	(3) Non-TVA South	(4) Non-TVA proposed authorities
Changes 1920–1930				
Log population	0.051	0.049	0.067	0.004
Log employment	0.082	0.096	0.111	0.045
Log # of houses	0.078	0.092	0.108	0.046
Log average manufacturing wage	0.117	0.217	0.108	0.172
Manufacturing employment share	−0.010	−0.035	−0.018	−0.018
Agricultural employment share	−0.047	−0.036	−0.047	−0.046
% White	0.012	−0.011	−0.010	0.000
% Urbanized	0.047	0.064	0.080	0.042
% Illiterate	−0.030	−0.014	−0.029	−0.019
% of Whites foreign born	−0.001	−0.023	−0.016	−0.012
Log average farm value	−0.013	−0.076	0.025	−0.182
# of Observations	163	2,326	795	828
# of States	6	46	14	25

Implementing Selection-on-Observables

KM use an Oaxaca-Blinder regression estimator to estimate ATTs:

1. In non-TVA counties j, estimate by OLS $Y_{jt} - Y_{j,t-1} = \alpha + X_j'\beta + \varepsilon_{jt}$
2. Estimate over TVA counties $i = 1, \ldots, N$:

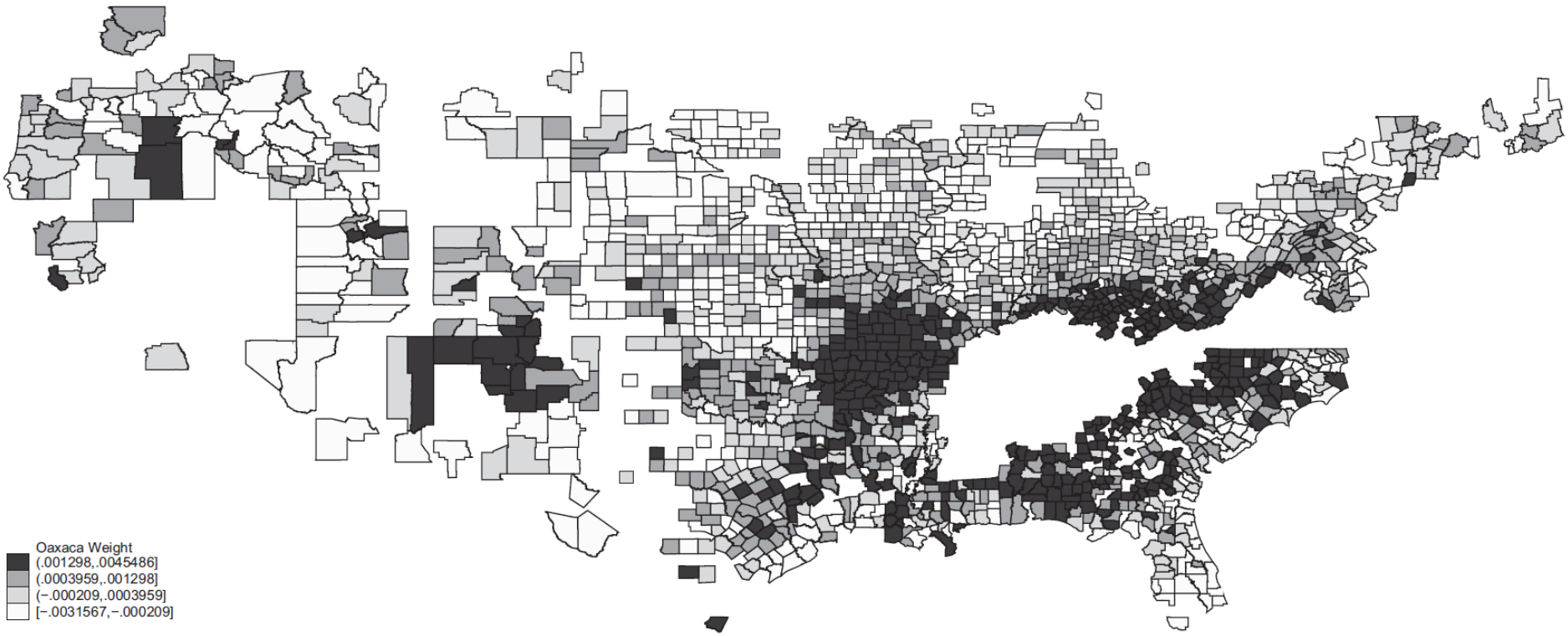
$$\widehat{\tau}_{ATT} = \underbrace{\frac{1}{N}\sum_i (Y_{it} - Y_{i,t-1})}_{\text{Observed treated outcome}} - \underbrace{\left(\hat{\alpha} + \frac{1}{N}\sum_i X_i'\hat{\beta}\right)}_{\text{Imputed counterfactual outcome}}$$

X_i here includes a flexible set of 38 economic, social, demographic, and geographical variables from 1920 and 1930

Kline (2011) shows $\widehat{\tau}_{ATT}$ can be interpreted as a propensity score weighting estimator: counties more similar to TVA ones get more weight

- Plotting the weights can help understand identifying variation
- Conventional regression-adjustment yield very similar results
- We'll see this sort of "imputation estimator" again soon...

Implicit Weight Given to Control Counties



Balance Tests

Decadalized Growth Rates in Tva Region versus Conterfactual Regions, 1900–1940

Outcome	(1) Point estimate (unadjusted)	(2) Point estimate (controls)	(3) N
Panel A: TVA region versus rest of U.S.			
Population	0.007	0.010	1,776
Total employment	−0.009	0.005	1,776
Housing units	−0.006	0.007	1,776
Average manufacturing wage	0.009	0.010	1,428
Manufacturing share	0.007*	0.005	1,776
Agricultural share	−0.007*	−0.001	1,776
Average agricultural land value	0.078***	0.025	1,746
Panel B: TVA region versus U.S. South			
Population	−0.018	0.003	850
Total employment	−0.028	0.001	850
Housing units	−0.025	0.005	850
Average manufacturing wage	0.001	0.001	687
Manufacturing share	0.005	0.005	850
Agricultural share	0.003	−0.002	850
Average agricultural land value	−0.009	−0.007	839
Panel C: TVA region versus proposed authorities			
Population	0.026	0.011	926
Total employment	−0.012	0.006	926
Housing units	−0.014	0.006	926
Average manufacturing wage	0.012	0.008	734
Manufacturing share	0.007	0.005	926
Agricultural share	−0.005	0.004	926
Average agricultural land value	0.080***	0.017	908

Effect Estimates

DECADALIZED IMPACT OF TVA ON GROWTH RATE OF OUTCOMES (1940–2000)

Outcome	(1) Point estimate (unadjusted)	(2) Point estimate (controls)	(3) N
Panel A: TVA region versus rest of U.S.			
Population	0.004	0.007	1,907
Average manufacturing wage	0.027***	0.005	1,172
Agricultural employment	−0.130***	−0.056**	1,907
Manufacturing employment	0.076***	0.059***	1,907
Value of farm production	−0.028	0.002	1,903
Median family income (1950–2000 only)	0.072***	0.021	1,905
Average agricultural land value	0.066***	−0.002	1,906
Median housing value	0.040**	0.005	1,906
Panel B: TVA region versus U.S. South			
Population	−0.007	0.014	942
Average manufacturing wage	0.003	0.001	610
Agricultural employment	−0.097***	−0.051*	942
Manufacturing employment	0.079***	0.063***	942
Value of farm production	−0.005	−0.006	939
Median family income (1950–2000 only)	0.041***	0.024**	942
Average agricultural land value	0.031*	−0.003	942
Median housing value	0.019	0.007	942
Panel C: TVA region versus proposed authorities			
Population	0.011	0.001	991
Average manufacturing wage	0.018***	0.005	618
Agricultural employment	−0.101***	−0.071***	991
Manufacturing employment	0.066***	0.053**	991
Value of farm production	0.002	0.011	989
Median family income (1950–2000 only)	0.060***	0.025**	991
Average agricultural land value	0.060***	−0.003	991
Median housing value	0.033**	0.009	991

Short- and Long-Run Effects

Decadalized Impact of Tva on Growth Rate of Outcomes over Two Subperiods

	(1)	(2)	(3)	(4)	(5)	(6)
	Entire U.S.		South		Proposed authorities	
Outcome	1940–1960	1960–2000	1940–1960	1960–2000	1940–1960	1960–2000
Population	0.037	−0.008	0.042	−0.000	0.028	−0.013
Average manufacturing wage	−0.005	0.014*	−0.003	0.010	0.007	0.012
Agricultural employment	0.106***	−0.134***	0.106***	−0.130***	0.119***	−0.166***
Manufacturing employment	0.114***	0.033**	0.116***	0.035*	0.097**	0.032**
Value of farm production	0.076*	−0.030	0.081**	−0.044	0.118**	−0.033
Median family income	N/A	0.017	N/A	0.016	N/A	0.019*
Average agricultural land value	0.027	−0.017	0.018	−0.015	0.029	−0.021
Median housing value	0.019	−0.003	0.010	0.005	0.020	0.003

Recall: TVA funding ran primarily from 1940-1960

Towards General Equilibrium Effects

We still don't know whether TVA increased *national* welfare, or just shifted gains to one part of the country from other

- Intuitive that we can't learn this from (simple) regression alone ... there is no "national control group"

In the second part of the paper, KM develop a tractable equilibrium model where key elasticities govern the extent of agglomeration externalities

- Identified by dynamic panel IV methods (a bit outside the scope of this class, at least for now)
- Key point: model structure permits extrapolation of identifying variation we have to policy counterfactuals we don't directly see

Summary

Selection on observables is a highly tractable identification strategy, with many different ways to implement

- I'd always recommend starting with regression (are you surprised?) but fancier methods are useful if you're after a particular estimand

A compelling selection-on-observables argument starts with a clear story for how treatment was chosen — institutional details are key!

- Ideally, some controls fall out of the story as "necessary" for identification while others can be used to test balance
- Good to *align* your validation exercise with your estimation procedure

Adding more controls need not bring you closer to identification!

- See MHE discussion of "bad controls," or more sophisticated discussions of "collider bias" from the DAG literature

Outline

1. Addressing Selection Bias with Controls✓
2. Matching, Weighting, and Regression✓
3. Application: Kline and Moretti (2014)✓
4. Bonus Track: Coefficient Stability

Altonji, Elder, and Taber (2005) and Oster (2019)

As we saw in DK'02, stability of an OLS coefficient across different control specifications can be compelling for a selection on observables story

- But what if the coeff changes? Can we learn something from that?

AET'05 and Oster'19 consider scenarios where unobserved selection is "similar to" observed selection

- They approach the meaning of "similar" slightly differently, and end up with slightly different approaches
- Since then, others have proposed other approaches too: e.g. Cinelli and Hazlett (2020) and Masten, Diegert and Poirier (2022)

My take: these techniques can be good to know, but are probably not ready for "primetime" (i.e. being your main identification strategy)

- Referees may well ask you to run them as robustness checks

AET Motivation: Catholic HS Graduation Effects

	All Students		Catholic Elementary	
	No Controls	**w/ Controls**	**No Controls**	**w/ Controls**
Probit coefficient	0.97	0.41	0.99	1.27
S.E.	(0.17)	(0.21)	(0.24)	(0.29)
Marginal effects	[0.123]	[0.052]	[0.11]	[0.088]
Pseudo R^2	0.01	0.34	0.11	0.58

Source: Table 3 in Altonji, Elder, Taber (2005)

AET Assumption: "Equal Selection" on Obs. / Unobs.

Returning to a constant-effects causal model $Y_i = \beta D_i + \nu_i$, imagine regressing ν_i on a vector of observables X_i (w/ a constant): $\nu_i = X_i'\gamma + \varepsilon_i$,

$$Y_i = \beta D_i + X_i'\gamma + \varepsilon_i \tag{1}$$

$Cov(X_i, \varepsilon_i) = 0$ by construction, but need not have $Cov(D_i, \varepsilon_i) = 0$

Now consider the regression of D_i on X_i and ε_i:

$$D_i = \phi_0 + \phi_{X'\gamma} X_i'\gamma + \phi_\varepsilon \varepsilon_i + \eta_i$$

Note that when ϕ_ε=0, $Cov(D_i, \varepsilon_i) = 0$ so eq (1) is a regression

AET make a weaker assumption: $\phi_\varepsilon = \phi_{X'\gamma}$. Equivalent to:

$$\frac{E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0]}{Var(\varepsilon_i)} = \frac{E[X_i'\gamma \mid D_i = 1] - E[X_i'\gamma \mid D_i = 0]}{Var(X_i'\gamma)}$$

Using AET

The OVB formula tells us the OLS estimator $\hat{\beta}$ satisfies

$$\begin{aligned} p\lim(\hat{\beta}) &= \beta + \frac{Cov(\tilde{D}_i \varepsilon_i)}{Var(\tilde{D}_i)} = \beta + \frac{Cov(D_i \varepsilon_i)}{Var(\tilde{D}_i)} \\ &= \beta + \frac{Var(D_i)}{Var(\tilde{D}_i)} (E[\varepsilon_i \mid D_i = 1] - E[\varepsilon_i \mid D_i = 0]) \end{aligned}$$

Under the AET assumption, this means

$$p\lim(\hat{\beta}) = \beta + \frac{Var(D_i)}{Var(\tilde{D}_i)} \frac{Var(\varepsilon_i)}{Var(X_i'\gamma)} \left(E[X_i'\gamma \mid D_i = 1] - E[X_i'\gamma \mid D_i = 0]\right)$$

The only thing we don't observe / can't estimate here is $Var(\varepsilon_i)$, so given a range of plausible $Var(\varepsilon_i)$ we can bound β!

AET Discussion

The AET assumption seems reasonable, but can be hard to think through.

AET show it is satisfied when:

1. Observed elements of X_i are chosen randomly from some set of potential "relevant covariates"
2. There are a large number of observed elements (so a "LLN" kicks in)
3. An additional (apparently hard to state) assumption holds: *"The regression of D_i^* on $Y_i - \beta D_i$ is equal to the regression of the part of D_i^* that is orthogonal to X_i on the corresponding part of $Y_i - \beta D_i$"* where D_i^* is an unobserved latent variable determining D_i

The AET bounds can be wide, depending on what restrictions we think are reasonable on $Var(\varepsilon_i)$ (how do we think about this?)

Oster (2019): Making the Link to Coefficient Stability

Starting from a similar setup as AET, Oster defines

$$\delta = \frac{Cov(\varepsilon_i, D_i)}{Var(\varepsilon_i)} \Big/ \frac{Cov(X_i'\gamma, D_i)}{Var(X_i'\gamma)}$$

She then shows that for $\delta \approx 1$ ("equal selection"),

$$\beta \approx \tilde{\beta} + \delta(\tilde{\beta} - \beta^*)\frac{R_{max} - \tilde{R}}{\tilde{R} - R^*}$$

where β^* and R^* are the coefficient and R^2 from regressing Y_i on D_i, $\tilde{\beta}$ and $\tilde{R}$ are the same with controls, and R_{max} is the maximum achievable R^2

- Intuition: adjust $\tilde{\beta}$ in the direction of $\tilde{\beta} - \beta^*$, with more adjustment if there are a lot of potential relevant unobservables
- Oster also gives a more complicated formula for $\delta \not\approx 1$

Oster Application: Maternal Behavior and Child Outcomes

Panel A: Child IQ, standardized (NLSY) ($R_{max} = 0.61$)

Treatment variable	(1) Baseline effect (Std. error), [R^2]	(2) Controlled effect (Std. error), [R^2]	(3) Null reject? (extrnl. evid.)	(4) Sibling FE estimate	(5) Identified set	(6) $\tilde{\delta}$ for $\beta = 0$ given R_{max}
Breastfeed (Months)	0.045*** (0.003) [0.045]	0.017***(0.002) [0.256]	No	−0.007 (0.005)	[−0.033,0.017]	0.37
Drink in Preg. (Any)	0.176***(0.026) [0.008]	0.050**(0.023) [0.249]	No	0.026 (0.036)	[−0.146,0.050]	0.26
LBW + Preterm	−0.188***(0.057) [0.004]	−0.125***(0.050) [0.251]	Yes	−0.111 (0.070)	[−0.124,−0.033]†	1.37

Panel B: Birth weight in grams (NLSY) ($R_{max} = 0.53$)

Treatment variable	Baseline effect (Std. error), [R^2]	Controlled effect (Std. error), [R^2]	Null reject? (extrnl. evid.)	Sibling FE estimate	Identified set	$\tilde{\delta}$ for $\beta = 0$ given R_{max}
Smoking in Preg	−183.1*** (12.9) [0.31]	−172.5***(13.3) [0.35]	Yes	−94.3***(27.6)	[−172.5,-30.3]†	1.08
Drink in Preg. (Amt)	−16.7***(5.15) [0.30]	−14.1*** (5.06) [0.34]	No	−1.53 (7.48)	[−14.1,0.49]	0.96

NOTES: This table shows the validation results for the analysis of the impact of maternal behavior on child birth weight and IQ. Baseline effects include only controls for child sex and (1) age dummies in the case of IQ and (2) gestation week in the case of birth weight. Full controls: race, age, education, income, marital status. Sibling fixed effects estimates come from NLSY in all panels. The identified set in Column (5) is bounded below by $\tilde{\beta}$ and above by β^* calculated based on R_{max} given in the top row of each panel and $\tilde{\delta} = 1$. Column (6) shows the value of $\tilde{\delta}$ which would produce $\beta = 0$ given the values of R_{max} reported in the title of each Panel. * significant at 10% level, ** significant at 5% level, *** significant at 1% level. †identified set excludes zero.

Practical Takeaways

With Oster's *psacalc* Stata package, you can:

- Estimate β for a choice of δ and R_{max}
- Find the "breakdown" δ which makes $\beta = 0$, for a set of R_{max}
- Important: Don't just use the $\delta \approx 1$ approximation "by hand"

If you are writing a paper with a selection-on-observables argument, referees are likely to ask you to do some or all of these things

- Not obvious to me the best way to do this (what's R_{max}?)
- Some discussion of possibly better alternatives, e.g. Masten et al '22

Probably not the best idea to base your main estimates on $\delta \approx 1$!