Controlling for Observables

Alexander Torgovitsky

ECON 31720: Applied Microeconometrics
University of Chicago, Fall 2020

1. **Overview**
2. Selection on Observables
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. Other Approaches
10. Summary

Randomized controlled trials

- Form the cornerstone of causal inference — the "ideal experiment"
- In economics especially, often do not capture exactly what we want
- Compliance, ethical, and cost issues when dealing with human subjects
- Forces us to mentally extrapolate, or else turn to observational data

Selection on observables

- A generalization of the assumption behind randomized controlled trials
- Conditional on covariates, treatment is "as good as randomly assigned"

Implementing selection on observables

- Giant literature due to biostatistics, focused on nonparametrics
- Propensity score sufficiency a key concept in implementation

→ Increasingly gets used elsewhere too e.g. DID designs

Outline

1. Overview
2. **Selection on Observables**
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. Other Approaches
10. Summary

Notation

- Focus on the case of a binary treatment $D \in \{0, 1\}$
- Potential outcomes $Y(0)$ and $Y(1)$ with $Y = DY(1) + (1 - D)Y(0)$
- Other observable variables X

Workhorse example

- D is enrolling in a job-training program
- $Y(0)$, $Y(1)$ and Y are potential and actual future earnings
- X are sociodemographics, work history, etc.
- Impact of (federally-funded) programs on labor market outcomes?
- Big topic in the 1980s–1990s, and still important (*massive* literature)
- Methodological proving grounds due to LaLonde (1986) critique

$\rightarrow$ Heckman & Hotz (1989), Dehejia & Wahba (2002), Smith & Todd (2006)

Definition

- There is **selection** into the treatment state D if

$$\underbrace{Y(d)|D=1}_{\text{observable}} \quad \text{is distributed differently from} \quad \underbrace{Y(d)|D=0}_{\text{unobserved}} \quad \text{for } d \in \{0,1\}$$

- Expected to occur if agents choose D with knowledge of $(Y(0), Y(1))$

Selection is a common concern

- Particularly concerning for neoclassical economists
- Agents choose job training $D \in \{0,1\}$ to max utility
- Utility will incorporate expected future earnings $Y(0), Y(1)$
- Agents who choose $D = 1$ might do so because of low $Y(0)$
- Data commonly supports this story — "Ashenfelter's (1978) dip"

The random assignment assumption

- **Random assignment:** $(Y(0), Y(1)) \perp\!\!\!\perp D$

$\rightarrow$ Treatment state and potential outcomes are independent

- Random assignment implies that there is no selection

Identification under random assignment

- RA implies the (marginal) distributions of $Y(0), Y(1)$ are identified:

$$F_{Y(d)}(y) \equiv \mathbb{P}[Y(d) \leq y] \underbrace{=}_{\text{random assignment}} \mathbb{P}[Y(d) \leq y | D = d] = \mathbb{P}[Y \leq y | D = d]$$

- Any parameter that is a function of $F_{Y(0)}, F_{Y(1)}$ is also point identified

$\rightarrow$ e.g. ATE $= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y|D = 1] - \mathbb{E}[Y|D = 0]$

- Treated/untreated subgroups identical $\Rightarrow$ ATE $=$ ATT $=$ ATU
- X not needed, but often used for balance tests and variance reduction

"The fundamental problem of causal inference"

- Even with random assignment, joint distributions aren't (point) id'd

$\Rightarrow$ For example, quantiles of $Y(1) - Y(0)$

- Sometimes called the **fundamental problem of causal inference**
- Intuitive: We never see both $Y(0)$ and $Y(1)$ for anyone
- Still, random assignment is better than no random assignment!

Random assignment is hard to get

- Randomized controlled experiments are the leading (only?) case
- Common in biostatistics, e.g. drug trials
- Lab/field experiments widely used in economics too, but have limitations

$\rightarrow$ "**external validity**" — to be discussed more later

- Random assignment rarely compelling with observational data

$\rightarrow$ When agents can control D, we typically expect selection

Definition

- Consider the **treatment/control contrast**: $\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0]$
- *Without* random selection this is contaminated with **selection bias**:

$$\begin{aligned}&\mathbb{E}[Y|D=1] - \mathbb{E}[Y|D=0] \\ &\quad = \underbrace{(\mathbb{E}[Y(1)|D=1] - \mathbb{E}[Y(0)|D=1])}_{\text{ATT}} + \underbrace{(\mathbb{E}[Y(0)|D=1] - \mathbb{E}[Y(0)|D=0])}_{\text{selection bias}}\end{aligned}$$

- First term is the causal effect for those who were treated
- Second term is how the treated would have been different anyway

Significance

- The mean contrast no longer represents the effect of D on Y

$\rightarrow$ It is **confounded** with other differences between treated/untreated

- Circumventing selection bias is the main challenge of causal inference
- When random assignment doesn't hold, we turn to other tools

Definition

- **Selection on observables** is the assumption that

$$(Y(0), Y(1)) \perp\!\!\!\perp D | X$$

- AKA: **unconfoundedness** and **ignorable treatment assignment**

$\rightarrow$ *Conditional on* X, treatment is *as-good-as* randomly assigned

- Random assignment the special case of $X = 1$

Thought experiment: a randomized controlled trial given $X = x$

1. Fix an $X = x$
2. **Match** treated ($D = 1$) and untreated agents ($D = 0$) with $X = x$

$\rightarrow$ Requires the **overlap condition**: $0 < \mathbb{P}[D = 0 | X = x] < 1$

3. Compare outcomes of the treated and untreated *within* $X = x$
4. Aggregate across different values of $X = x$

Outline

1. Overview
2. Selection on Observables
3. **Imputation Estimators**
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. Other Approaches
10. Summary

Argument

- Conditional version of random assignment:

$$F_{Y(d)}(y|x) \equiv \mathbb{P}[Y(d) \leq y|X = x]$$
$$= \mathbb{P}[Y(d) \leq y|D = d, X = x] = \mathbb{P}[Y \leq y|D = d, X = x]$$

- Second equality requires the **overlap condition**: $\mathbb{P}[D = d|X = x] > 0$
- Aggregating by averaging over x identifies the marginals:

$$F_{Y(d)}(y) \equiv \mathbb{P}[Y(d) \leq y] = \mathbb{E}\left(\mathbb{P}[Y(d) \leq y|X]\right) = \mathbb{E}\left(\mathbb{P}[Y \leq y|D = d, X]\right)$$

Implication for specific parameters

- $\text{ATE} = \mathbb{E}\left[\mathbb{E}[Y|D = 1, X]\right] - \mathbb{E}\left[\mathbb{E}[Y|D = 0, X]\right]$
- $\text{ATT} = \mathbb{E}[Y|D = 1] - \mathbb{E}\left[\mathbb{E}[Y|D = 0, X]|D = 1\right]$ — first term is easy
- $\text{ATU} = \mathbb{E}\left[\mathbb{E}[Y|D = 1, X]|D = 0\right] - \mathbb{E}[Y|D = 0]$ — second term is easy

ATE

- Let $\mu_d(x) \equiv \mathbb{E}[Y|D=d, X=x]$ for $d \in \{0,1\}$
- Previous expressions involve averaging over $\mu_0(X)$ and/or $\mu_1(X)$, e.g.

$$\text{ATE} = \underbrace{\mathbb{E}}_{\text{over } X}\Big[\mathbb{E}[Y|D=1,X] - \mathbb{E}[Y|D=0,X]\Big] \equiv \underbrace{\mathbb{E}}_{\text{over } X}\Big[\mu_1(X) - \mu_0(X)\Big]$$

- An **imputation estimator** of the ATE based on data $\{(Y_i, D_i, X_i)\}_{i=1}^N$ is

$$\widehat{\text{ATE}} \equiv \frac{1}{N}\sum_{i=1}^{N} \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i) \quad \text{where } \hat{\mu}_d(x) \text{ is an estimator of } \mu_d(x)$$

- Estimate conditional means, then take the sample analog

ATT/ATU are similar, but require less estimation

- ATT/ATU only need $\mu_0(X)/\mu_1(X)$ — $\mathbb{E}[Y|D=d]$ estimated directly
- Sample average should be conditional on $D=1$ or $D=0$

Estimating conditional means

- Need to choose estimators $\hat{\mu}_0$ and $\hat{\mu}_1$

→ Many nonparametric options — see previous lecture

- Curse of dimensionality will typically kick in quickly
- Most common are linear regression and matching

Imputation with linear regression

- Easiest: regress Y on X and D, take coefficient on D
- Better: regress Y on X among $D = d$: $\mu_d(x) = \alpha_d + \beta_d' x$ then impute:

$$\widehat{\text{ATE}} \underbrace{=}_{\text{see supplement!}} \underbrace{\overline{Y}_1 - \overline{Y}_0}_{\text{naive contrast}} + \underbrace{\left(\frac{N_1}{N} \widehat{\beta}_0 + \frac{N_0}{N} \widehat{\beta}_1 \right)' (\overline{X}_0 - \overline{X}_1)}_{\text{regression adjustment}}$$

- Concerns about functional forms driving results via extrapolation

→ The usual concern when using a parametric estimator

Linear regression imputation as a weighted average (scalar X_i)

$$\widehat{\mathbb{E}}[Y(0)|D=1] = \frac{1}{N_0}\sum_{i:D_i=0} Y_i \bar{W}_{i0} \quad \text{where} \quad \bar{W}_{i0} \equiv 1 - (X_i - \bar{X}_0)\left(\frac{\bar{X}_0 - \bar{X}_1}{\overline{X_0^2} - \bar{X}_0^2}\right)$$

Example due to Imbens (2015)

- LaLonde (1986) data, target parameter is ATT, just need $\mathbb{E}[Y(0)|D=1]$
- Control group taken from Current Population Survey

$\rightarrow$ General population, so has *much* higher past earnings

- If X_i is earnings before the program, then weighting above becomes

$$W_i = 2.8091 - .0949 \times X_i$$

- So $X_i = 100K \Rightarrow W_i \approx -6.67$ — high earners negatively weighted
- Would probably prefer weighting them 0 — why should they matter?

Define distance and matches

- Mahalonobis: $\text{dist}_{ij} \equiv (X_i - X_j)'\hat{V}^{-1}(X_i - X_j)$ $\quad \hat{V}$ var-cov matrix

$\rightarrow$ Gives all X's the same scale — can also just use the diagonal

- For each i, find Kth smallest element of $\{\text{dist}_{ij} : D_j \neq D_i\}_{j=1}^{n}$ — $\text{dist}_i^\star$
- Let $\mathcal{J}_i = \{j : D_j \neq D_i \quad \text{and} \quad \text{dist}_{ij} \leq \text{dist}_i^\star\}$

$\rightarrow$ Could have more than K elements if there are ties

Impute

- Apply the general formula with

$$\hat{\mu}_d(X_i) = \mathbb{1}[D_i = d]Y_i + \mathbb{1}[D_i \neq d]\frac{1}{|\mathcal{J}_i|}\sum_{j \in \mathcal{J}_i} Y_j$$

- Use Y_i to impute $Y_i(D_i)$ (could also do this with any other method!)
- Use average Y_i for $\approx K$ closest matches to impute $Y_i(d), d \neq D_i$

Outline

1. Overview
2. Selection on Observables
3. Imputation Estimators
4. **Empirical Example #1: Angrist (1998)**
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. Other Approaches
10. Summary

Empirical question

- US military plays an important role in labor market for young men
- What is the effect of *volunteer* service on future labor market outcomes?

Empirical challenge

- Voluntarily serving in the military is not random
- Influenced by other job options, as well as health and screening
- Veterans more likely to be healthier, but had worse job prospects

Data

- US military admin. data linked with Social Security earnings
- Only covers those who have applied and taken preliminary tests
- $\approx$ 300,000 observations with some demographics and test scores

Variables

- Y is earnings in a given year (both pre- and post- service)
- D is whether an applicant ended up serving in the military
- X is race, application year, schooling, AFQT, age

Selection on observables assumption

- "Conditional on X and applying, serving in military is independent of potential labor market outcomes"

→ Nothing that *we can't see* matters for both serving and outcomes

- Job prospects, interview, psych. eval., firm handshake, face tattoo ... ?

Implementation

- Splits by race, then discretizes X into roughly 450 cells
- Binning/saturated regression imputation estimator of ATT
- Also, "semi-saturated" regression of Y on D *and* X ...

Result from Angrist (1998)

- Regress Y on D and fully saturated X (but no interactions)
- *Not* a saturated regression (no interactions) — "semi-saturated"?
- The coefficient on D converges to β_{ssat}:

$$\beta_{\text{ssat}} \equiv \mathbb{E}\left[\frac{\text{Var}(D|X)}{\mathbb{E}[\text{Var}(D|X)]}(\mu_1(X) - \mu_0(X))\right] \quad \text{(see supplement)}$$

Discussion

- Shows a particular LR is a positively-weighted average of $\mu_1(x) - \mu_0(x)$

→ Nonparametric imputation has positive weights, but LR can be negative

- Positive weights necessary but not sufficient for an interesting parameter

→ But $\beta_{\text{ssat}} \neq$ ATE, ATT, ATU — how to interpret?

- Our first example of **reverse engineering** — why do this?

		Whites		
Year	Mean (1)	Difference in Means[c] (2)	Controlled Contrast (3)	Regression Estimates (4)
A. *Earnings*[a]				
74	182.7	−26.1	−14.0	−13.0
		(7.0)	(9.2)	(9.4)
75	237.9	−41.4	−14.2	−12.0
		(6.3)	(7.6)	(7.8)
76	473.4	−47.9	−14.8	−12.7
		(8.1)	(9.0)	(9.3)
77	1012.9	−7.1	−8.6	−9.4
		(11.3)	(12.3)	(12.2)
78	2147.1	40.3	−23.5	−22.4
		(16.7)	(18.1)	(17.2)
79	3560.7	188.0	−8.4	−11.2
		(21.0)	(23.2)	(21.6)
80	4709.0	572.9	178.0	175.9
		(23.4)	(27.2)	(24.6)
81	6226.0	855.5	249.5	249.9
		(27.2)	(32.4)	(29.1)
82	7200.6	1508.5	783.3	782.4
		(30.3)	(36.4)	(32.5)
83	8398.1	1390.5	588.8	601.5
		(34.4)	(41.1)	(36.6)
84	9874.2	652.8	−235.7	−198.5
		(39.5)	(46.9)	(41.7)
85	10972.7	469.8	−521.3	−459.6
		(44.6)	(52.6)	(46.8)
86	12004.5	543.7	−557.3	−491.7
		(50.4)	(59.0)	(52.5)
87	13045.7	663.9	−548.0	−464.3
		(54.6)	(63.9)	(56.8)
88	14136.1	904.3	−415.5	−311.7
		(58.3)	(68.2)	(60.6)
89	14716.1	1169.1	−248.6	−136.3
		(61.0)	(71.2)	(63.2)
90	14886.1	1300.8	−154.5	−53.2
		(63.0)	(73.6)	(65.2)
91	14407.9	1559.6	29.8	146.2
		(64.6)	(75.6)	(66.9)

- Naive contrast (2) small before application but grows larger

$\rightarrow$ Same sign in the short-run, but flips in the long-run

- Nonparametric (3) and semi-saturated (4) somewhat similar (“cosmetic”)
- But still important differences and estimate *different parameters*

1. Overview
2. Selection on Observables
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. **Imputation with the Propensity Score**
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. Other Approaches
10. Summary

Definition

- Binary treatment case, $D \in \{0, 1\}$
- $p(x) \equiv \mathbb{P}[D = 1|X = x]$ is called the **propensity score**
- Let $P \equiv p(X)$ be the random variable $\mathbb{P}[D = 1|X]$

Rosenbaum and Rubin (1983) sufficiency argument

- Selection on observables implies $(Y(0), Y(1)) \perp\!\!\!\perp D|P$. **Proof:**

$$\begin{aligned}
\mathbb{P}\left[D = 1 \mid Y(0), Y(1), P\right] &= \mathbb{E}\left(\mathbb{P}\left[D = 1 \mid Y(0), Y(1), P, X\right] \Big| Y(0), Y(1), P\right) \\
&= \mathbb{E}\left(\mathbb{P}\left[D = 1 \mid Y(0), Y(1), X\right] \Big| Y(0), Y(1), P\right) \\
&= \mathbb{E}\left(\mathbb{P}\left[D = 1 \mid X\right] \Big| Y_0, Y_1, P\right) \\
&\equiv \mathbb{E}\left(P \mid Y(0), Y(1), P\right) = P \qquad \text{Q.E.D.}
\end{aligned}$$

- Implication is that we can condition on P instead of X
- Still need overlap, but now with P (scalar) instead of X (vector)

The propensity score and dimension reduction

- Sufficiency $\Rightarrow$ replace $\mu_d(x)$ with $\nu_d(p) \equiv \mathbb{E}[Y|D=d, P=p]$

→ Given estimates $\hat{P}_i \equiv \hat{p}(X_i)$, we can impute with $\hat{P}_i$ in place of X_i

- Appears to break the curse of dimensionality since P is scalar

→ But of course it doesn't — now we need to estimate $p(x)$

- Still, having to parameterize p is arguably better than both (μ_0, μ_1)
- In practice, usually see a logit estimator for $\hat{p}$

Estimators

- **Propensity score matching** is very popular in biostatistics
- Basically the same as matching — and no Mahanobis distance needed

→ Although there are dozens of variations (replacement? one-to-one? etc.)

- Imbens (2015) recommends blocking with a linear regression …

- One could use kernel or sieve methods for $\nu_d(p)$
- Subclassification (**blocking**) is a particular type of sieve

$\rightarrow$ Constant spline, also called a partitioning estimator

Blocking

- Divide $[0, 1]$ into $\{b_0, b_1, \ldots, b_J\}$ with $b_0 = 0$, $b_J = 1$
- Define $B_j = 1$ if $p(X) \in (b_{j-1}, b_j)$ as membership in block j
- If $b_j - b_{j-1}$ is small then roughly random assignment within block
- Estimate $\widehat{\text{ATE}}_j = \overline{Y}_{1,j} - \overline{Y}_{0,j}$ per block, i.e. conditional on $B_j = 1$
- Then average $\widehat{\text{ATE}}_j$ by block size into $\widehat{\text{ATE}}$

- Key question is how to construct the blocks
- Imbens (2015) suggests an algorithm based on testing $D \perp\!\!\!\perp X | \{B_j\}_{j=1}^J$

$\rightarrow$ $D \perp\!\!\!\perp X | P$ implied by selection on observables — so check within blocks

Combining two approaches

- Imbens (2015) suggests combining blocking with linear regression
- First construct the blocks
- Then *within each block*, run a linear regression Y on $1, D, X$
- Coefficient on D for each block, average up over blocks

Why?

- Intuitively, this could potentially reduce both bias and variance
- The variance part is clear if accounting for X reduces variation in Y
- The bias part is less clear (i.e. not necessarily true) ...
- Recall that linear regression extrapolates if $\overline{X}_1 \neq \overline{X}_0$

$\rightarrow$ However within each block $\overline{X}_1 \approx \overline{X}_0$ — little extrapolation

- Adjusting for X reduces remaining differences within blocks

$\rightarrow$ But presumably the remaining differences should be small anyway?

Outline

1. Overview
2. Selection on Observables
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. **Assessing Selection on Observables**
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. Other Approaches
10. Summary

Only predetermined observables

- For selection on observables to be plausible, X should be **predetermined**
- In particular, D should not have a causal effect on X
- Usually this really is a temporal issue (i.e. measured before vs. after D)
- Intuition is clear — we want to condition on selection *into* treatment

Simple but trivial example

- Suppose we accidentally included Y as part of X
- Then clearly we aren't going to have $(Y(0), Y(1)) \perp\!\!\!\perp D|X$

Less trivial examples in the context of job training

- Don't include earnings 1 year after the program in X
- Don't include employment after the program in X
- Don't include marital status after the program in X
- Ok to include sunspots after the program, but it won't help

Question

- Suppose $(Y(0), Y(1)) \perp\!\!\!\perp D | X_1, X_2, X_3, \ldots$
- All we have available is X_1
- Is it better to condition on X_1 instead of not conditioning on anything?

Answer

- Not necessarily

$\rightarrow$ Surprising? Conditioning on something should be better than nothing?

- Why: Selection bias conditional on X_1 could be worse than unconditional

$\rightarrow$ See the supplemental notes for a simple example

Implications

- Means we really need to have "the correct set of X"

$\Rightarrow$ Need to be careful with automated model selection (machine learning)

- Point is not well-appreciated but should be concerning

Idea and motivation

- Selection on observables is not directly testable — more next lecture
- Instead, auxiliary **placebo tests** are sometimes used as support
- Suppose there is another variable W known to be unaffected by D

→ Typical choice would be another pre-determined covariate not in X

- Suppose we treat W as Y and estimate the ATE
- If we reject the hypothesis that ATE $= 0$, then we should be concerned

→ Suggests unobservable differences in treated/untreated given X

Critique

- Can be difficult to see what would comprise a good W
- Needs to be something that is not otherwise included in X
- Otherwise you are changing the selection on observables assumption
- Also need power — not rejecting when W is a sunspot isn't helpful

Inherent unobservables

- Selection on observables can be difficult to believe in economics

$\rightarrow$ **Inherent unobservables:** preferences, private info, expectations, ...

- Observationally identical people behave differently due to ... a coin flip?

Controlling for more is not a solution

- Often argued that large X makes selection on observables "more likely"

$\rightarrow$ But remember the previous example — conditioning on more was *worse*

- Even if you buy this, still raises an uncomfortable friction with overlap

$\rightarrow$ If we could *perfectly* explain D with X then $\mathbb{P}[D = 1|X] \in \{0, 1\}$

Better methods for choosing observables will not solve this

- Selection on observables is seeing a resurgence with machine learning
- Fancier methods, but the identifying assumption is still the same
- Bias/variance trade-off is not the first-order issue here

1. Overview
2. Selection on Observables
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. **Empirical Example #2: List et al (2003)**
8. Statistical Inference
9. Other Approaches
10. Summary

Empirical question

- Local discretion in the US on stringency of environmental regulation
- What is the effect of this on where manufacturers locate?

$\rightarrow$ Is there a "race to the bottom"?

Empirical challenge

- Local governments do not randomly choose environmental enforcement
- Influenced by current federal attainment status and economic health
- Also potentially confounded by public attitudes/demographics

Data

- County-level yearly panel data 1980–1990 for New York state
- 62 counties observed in each of 11 years $\Rightarrow$ sample size 682
- Number of individual plant openings and closings
- Federal ozone attainment status and other county characteristics

TABLE A1.—DESCRIPTION OF VARIABLES

Variable	Mean	In-Attainment Mean	Out-of-Attainment Mean	Definition and Source
New pollution-intensive plants	0.41 (0.89)	0.31 (0.64)	0.70 (1.32)	Actual count of new plants from 1980 to 1990 labeled as having production activities that are pollution-intensive. Industrial Migration File, NYS DED.
New non-pollution-intensive plants	1.05 (2.09)	0.71 (1.25)	2.02 (3.36)	Actual count of new plants from 1980 to 1990 labeled as having production activities that are non-pollution-intensive. Industrial Migration File, NYS DED.
Attainment status	0.26 (0.44)	—	—	Intensity of county-level pollution regulations. Dichotomous variable = 1 if county is out of attainment of federal standards for ozone, 0 otherwise. Federal Register Title 40 CFR, Part 81.305.
ln(employment)	10.81 (1.33)	10.55 (1.15)	11.59 (1.53)	Natural logarithm of total employment in manufacturing. *County Business Patterns.*
ln(wage)	9.71 (0.23)	9.74 (0.22)	9.65 (0.25)	Natural logarithm of total annual manufacturing payroll divided by the number of employees by county, adjusted for inflation. *County Business Patterns.*
ln(population)	11.66 (1.25)	11.39 (1.07)	12.47 (1.38)	Natural logarithm of county population. *Current Population Reports,* U.S. Bureau of Census.
ln(property tax)	6.26 (0.34)	6.27 (0.35)	6.25 (0.28)	Natural logarithm of real property tax collected per capita. *Census of Governments.*

Data are for the 62 New York counties from 1980 to 1990. $N = 682$ (176 out of attainment).
Standard deviations in parentheses.

- Y is number (or net number) of plants that open in a year
- $D \in \{0, 1\}$ is federal attainment status ($D = 1$ is polluted)
- X are wages, existing plants, population, per capita income, etc.

The selection on observables assumption

- $Y(0), Y(1)$ are plants that *would have* opened under attainment status
- Assumption: Conditional on county-time observation characteristics, actual attainment status is independent of potential plant openings

One-to-one propensity score matching with caliper

1. Estimate propensity score by county-year — specification next slide
2. Match each treated observation to observation w/ closest P among:
   a. Untreated and in the same year (across counties)
   b. Untreated and in the same year/region (across counties)
   c. Untreated and in the same county (across years)

$\rightarrow$ Like matching on both P and certain components of X

3. Drop if difference in P is greater than .01 or .05 (**caliper matching**)
4. Drop untreated observations not matched to a treated observation (ATT)
5. Take simple difference in means across treated/untreated pairs

TABLE A2.—FIRST-STAGE LOGIT ESTIMATES OF THE DETERMINANTS OF ATTAINMENT STATUS

Independent Variable	Coefficient (SE) (1)		(2)	
Neighboring attainment status	2.85*	(0.33)	—	
Man. employment	1.99E−06	(1.29E−06)	—	
Property taxes	−1.85E−03*	(8.75E−04)	—	
Man. wages	−3.95E−06	(7.08E−05)	3.63E−03	(2.55E−03)
(Man. wages)^1			−2.23E−07	(1.41E−07)
(Man. wages)^2			4.27E−12	(2.74E−12)
Man. plants			1.40*	(0.58)
(Man. plants)^1			−0.09*	(0.05)
(Man. plants)^2			1.84E−03*	(1.04E−03)
Population	1.62E−06*	(5.09E−07)	−1.85E−06	(6.28E−06)
Population^1			7.37E−12	(6.12E−12)
Population^2			−3.14E−18*	(1.82E−18)
Per capita income			4.73E−03*	(1.25E−03)
(Per capita income)^1			−1.86E−07*	(9.64E−08)
(Per capita income)^2			2.63E−12*	(1.40E−12)
Man. wages × man. plants			−9.57E−06	(3.20E−05)
Man. wages × population			1.08E−09*	(4.53E−10)
Man. wages × per capita income			−1.61E−08	(6.61E−08)
Man. plants × population			−8.61E−07*	(3.54E−07)
Man. plants × per capita income			1.67E−05	(3.04E−05)
Population × per capita income			−8.88E−10*	(4.10E−10)
Time effects	Yes		Yes	
Log likelihood	−180.7		−145.8	
Pseudo R^1	0.54		0.63	
N	682		682	

Dependent variable is equal to 1 if county is out of attainment of federal ozone standards during the year, 0 otherwise. Neighboring attainment status is the percentage of western contiguous neighbors that out of attainment.

Time effects jointly significant at the 1% level.

[1] Standard errors are in parentheses beside the coefficient estimates and are adjusted for clustering within counties. * indicates significant at the 10% level using a two-sided alternative.

[2] Model (1) is used in the two-step FE Poisson estimation. Model (2) is used to generate the propensity score estimates.

Results

TABLE 1.—PROPENSITY SCORE ESTIMATES OF ATTAINMENT-STATUS EFFECT

Independent Variable	Matching Algorithm: Within Year Max. Difference (0.01)	(0.05)	Within Region & Year Max. Difference (0.01)	(0.05)	Within County Max. Difference (0.01)	(0.05)
Propensity score	−0.00 (0.99)	0.00 (0.97)	0.00 (0.98)	0.00 (0.98)	0.00 (1.00)	0.01 (0.97)
Man. wages ($1000s)	−0.73 (0.33)	−0.20 (0.66)	−0.06 (0.98)	−0.91 (0.44)	0.54 (0.60)	−0.01 (0.99)
Man. employment ($1000s)	−38.86 (0.27)	−52.88 (0.07)	29.94 (0.63)	4.05 (0.93)	3.11 (0.98)	2.53 (0.98)
Man. plants	−0.72 (0.52)	−0.76 (0.32)	−0.79 (0.82)	−2.37 (0.26)	0.59 (0.70)	0.48 (0.74)
Population (1000s)	−53.91 (0.50)	−40.74 (0.57)	59.49 (0.55)	4.61 (0.96)	−0.65 (1.00)	−0.31 (1.00)
Per capita income ($1000s)	−0.09 (0.89)	0.15 (0.72)	−0.66 (0.79)	−0.61 (0.66)	0.33 (0.84)	−0.20 (0.90)
Property tax	−31.38 (0.40)	7.85 (0.73)	−389.13 (0.06)	−186.81 (0.10)	1.22 (0.98)	1.00 (0.98)
High school graduates (%)	−1.10 (0.34)	−0.85 (0.29)	−3.61 (0.32)	−3.39 (0.12)	−1.09 (0.70)	−0.89 (0.71)
Highway expenditure	−0.01 (0.38)	0.01 (0.31)	−0.16 (0.09)	−0.07 (0.16)	−0.00 (0.97)	−0.00 (0.92)
Number of matched pairs	37	81	8	16	9	11
Number of unique controls	33	44	8	15	6	7

Entries represent mean difference between treatment counties (out of attainment) and control counties (in attainment). *p*-values in parentheses are for the tests that the mean difference across the treatment a controls groups are equal.
"Dirty" plants are those defined as pollution-intensive (see text); "clean" are all remaining manufacturing plants.
"Unique controls" reports the number of control counties that are matched with at least one treatment county.

- Three matches and two calipers each — their preferred columns
- Well-balanced on p-score (by construction) — on observables it varies
- Observables left out of p-score (e.g. property tax) seen as placebo

Results

TABLE 1.—PROPENSITY SCORE ESTIMATES OF ATTAINMENT-STATUS EFFECT

	Matching Algorithm					
	Within Year Max. Difference		Within Region & Year Max. Difference		Within County Max. Difference	
Independent Variable	(0.01)	(0.05)	(0.01)	(0.05)	(0.01)	(0.05)
New dirty plants ($\tau_{TT,\cdot}$)	−0.32	−0.69	0.38	−0.19	−1.33	−1.18
	(0.08)	(0.00)	(0.25)	(0.60)	(0.09)	(0.07)
New clean plants	0.03	−0.59	1.25	0.50	0.00	−0.18
	(0.95)	(0.08)	(0.07)	(0.36)	(1.00)	(0.84)
Net new plants ($\tau_{DID,\cdot}$)	−0.35	−0.10	−0.88	−0.69	−1.33	−1.00
	(0.27)	(0.68)	(0.12)	(0.05)	(0.03)	(0.08)
Lagged new dirty plants (1 year)	−0.07	−0.06	0.71	0.43	1.00	1.04
	(0.79)	(0.70)	(0.08)	(0.10)	(0.12)	(0.05)
Lagged net new plants (1 year)	0.53	0.71	0.50	0.44	0.00	−0.14
	(0.31)	(0.04)	(0.41)	(0.43)	(1.00)	(0.74)

- New dirty plants are the main outcome
- Also net (dirty - clean) new plants — argue differences out unobservables
- Lags and clean plants viewed as types of balance/placebo tests
- Preferred estimates are -.7 to -1.3 plants (off of a mean of .4)

1. Overview
2. Selection on Observables
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. **Statistical Inference**
9. Other Approaches
10. Summary

Bootstrap

- Variance calculations are complicated even for imputation estimators:

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mu}_1(X_i)-\hat{\mu}_0(X_i) \quad \text{complicated by extra} \quad \frac{1}{N}\sum_{i=1}^{N}$$

- Propensity score methods have two steps, make this even more annoying
- Possible to derive SE formulas, but why? Just use the bootstrap …

Caveat: matching estimators

- The bootstrap requires some underlying smoothness to work

→ It is conditional on data, parameter needs to change smoothly with data

- Matching estimators are not smooth (Abadie and Imbens, 2008)
- Abadie and Imbens (2006, 2016), Imbens (2015) provide SE formulas

Something to be alert about

- Some (mainly/only Imbens?) argue we should do statistical inference on

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[Y_i(1)-Y_i(0)|X_i]}_{\text{“conditional ATE” (CATE)}} \quad \textit{instead of} \quad \underbrace{\mathbb{E}[Y(1)-Y(0)]}_{\text{the usual population ATE}}$$

- The asymptotic variance of the CATE is always weakly lower

→ Intuitively, not taking into account variation in X_i

I recommend focusing on population treatment effects

- It is more standard, everyone will know what you are talking about
- I have not seen a compelling scientific argument for CATE

→ Here's one against: your parameter of interest now depends on your data

- Moving the goalposts to gain a few p–value points isn't worth it

1. Overview
2. Selection on Observables
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. **Other Approaches**
10. Summary

The ATE as a weighted average

- Selection-on-observables implies a weighting result using P:

$$\text{ATE} = \mathbb{E}\left[\frac{Y(D-P)}{P(1-P)}\right] \qquad \text{(see supplement)}$$

- Similar expressions can be derived for the ATT and ATU, e.g.

$$\text{ATT} = \mathbb{E}\left[\frac{Y(D-P)}{\mathbb{P}[D=1](1-P)}\right]$$

Implementation

- Estimate P, then take a simple average of weighted Y
- Another way of shifting the curse of dimensionality to $p(x)$
- Practical issues with P being close to 0 or 1 — trimming

$\rightarrow$ Probably why less popular (but see Busso et al 2014, Ma & Wang 2019)

Idea

- Imputation based on X requires modeling $\mu_d(x)$, but not $p(x)$
- Propensity score weighting requires modeling $p(x)$, but not $\mu_d(x)$
- Model both and combine into a **doubly robust** estimator

$\rightarrow$ Consistent estimator if *either* $\mu_d(x)$ or $p(x)$ is correctly specified

Form of the estimator

- Take propensity score weighting and add a correction term:

$$\mathbb{E}\left[\frac{DY}{p(X)} - \frac{(D - p(X))}{p(X)}\mu_1(X)\right] = \mathbb{E}[Y(1)] \qquad \leftarrow \text{see supplement}$$

- Equality holds if $p(x) = \mathbb{P}[D = 1|X = x]$ **or** $\mu_1(X) = \mathbb{E}[Y(1)|X = x]$
- Estimate by sample analog — replace $p(x)$ by $\hat{p}(x)$ and $\mu_1(x)$ by $\hat{\mu}_1(x)$
- Analogous argument holds for $\mathbb{E}[Y(0)]$

We haven't discussed multivalued treatments ...

- It *is* interesting — many counterfactual states are multivalued
- Selection on observables becomes: $\{Y(d)\}_{d\in\mathcal{D}} \perp\!\!\!\perp D|X$
- Not many interesting/relevant methodological differences

→ Some details regarding the (generalized) propensity score (problem set)

- The literature is overwhelmingly about $D \in \{0, 1\}$

Why the focus on binary treatments? (*Speculation*)

- The reason seems (to me) to be mostly sociological
- Nonparametric methods are highly valued by those in this literature
- With $D \in \{0, 1\}$ there is only nonparametric (at least, in D)
- If $D \in \{0, 1, 2\}$, then one needs to make a choice:

1. Make a (potentially wrong) functional form assumption
2. Remain nonparametric — basically reduces back to the binary case

- Community is against the first option, and second has low payoff

1. Overview
2. Selection on Observables
3. Imputation Estimators
4. Empirical Example #1: Angrist (1998)
5. Imputation with the Propensity Score
6. Assessing Selection on Observables
7. Empirical Example #2: List et al (2003)
8. Statistical Inference
9. Other Approaches
10. **Summary**

Key points

- Selection on observables a generalization of random assignment
- Many ways to implement — dimensions reduction with propensity score

→ Methods differ on details, not on the main idea

- Requires strong assumptions about role of unobservables

→ "Inherent unobservables" are crucially important in economics

- Not a satisfying choice model — given X, choices are … random?
- Requires conditioning on exactly the right set of X's

What next?

- Selection on observables is less widely used in economics now
- Researchers want to allow for selection on *unobservables*
- We will discuss methods that allow for this (to differing degrees)
- Alternatives come with other challenges (heterogeneity & extrapolation)