Randomization and Selection on Observables

Lukáš Lafférs

Matej Bel University, Dept. of Mathematics

We may be fortunate to run a randomized experiment.

This makes identification and estimation of causal effects easy.

But even a proper experiment may be ”broken” in many interesting ways.

In many other cases, this is not possible.

We rely on the fact that observable characteristics make the treatment ”as good as random”.

There are different ways how to do this. With different pros and cons.

Randomization

- N individuals
- $D_i \in \{0, 1\}$ treatment indicator
- $Y_i(D_i)$ potential outcomes
- $Y_i = Y_i(1)D_i + (1 - D_i)Y_i(0)$ observe variable
- $Y_i(.)$ is only a function of i-th treatment nad there are no interactions
- there are no hidden versions of the treatment, everyone receives 0 or 1

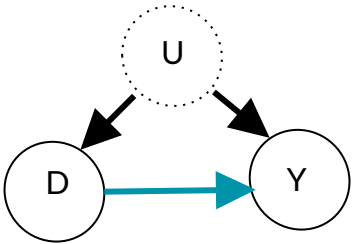
$$\delta_i = Y_i(1) - Y_i(0)$$

is individual treatment effect

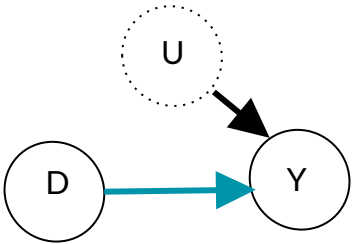
$Y(1)$ and $Y(0)$

- What are they really?
- $Pr(Y_i(1) = y) = Pr(Y_i = y|do(D = 1))$
- What if we cannot manipulate the treatment? What if it does not make sense?
- Is it enough if we can contemplate it?
- Sometimes we can manipulate the treatment.
- Sometimes nature can manipulate the treatment (e.g. gender).
- Missing data problem. You have to fix this.
- Somehow.

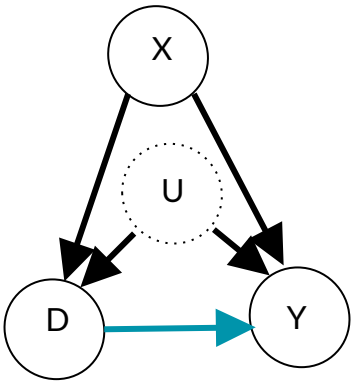
Observational data



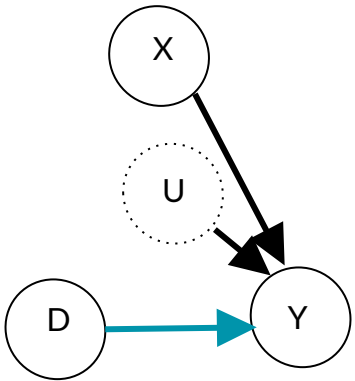
Randomized trial



Observational data



Randomized trial



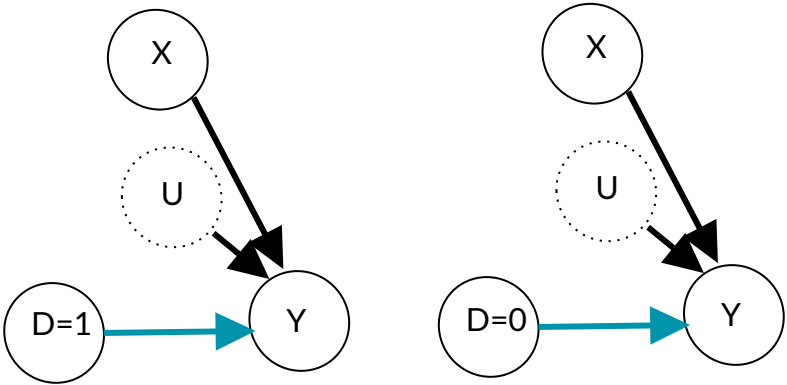
Treatment is randomized.

All the parents of D are removed.

There is no way how X or U have any influence on D.

Y is a ”collider” on the path between D and X and the path is therefore blocked.

$D \perp\!\!\!\perp X$ and $D \perp\!\!\!\perp U$



Randomization manipulated the treatment status of these people.

If randomization was successful, these two groups will not differ in terms of *X*

Randomization is the benchmark

If randomization worked, we should have:

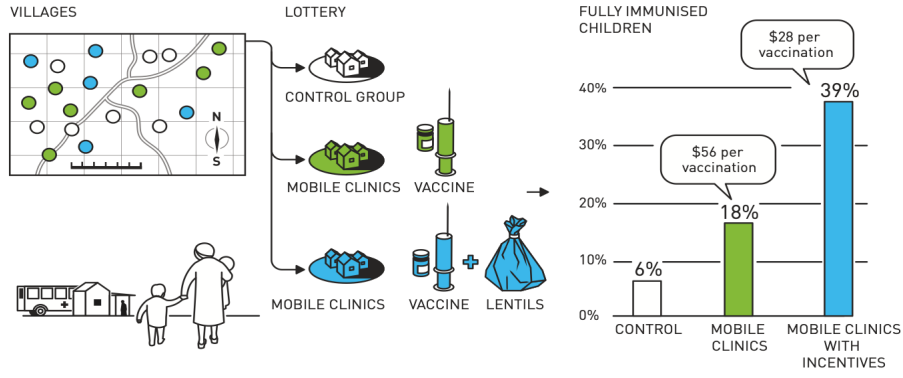
$$E[X|D=1] = E[X|D=0]$$

and this can be checked in the data.

The subjects should ideally only differ in terms of D.

Apples to apples.

Econ Nobel price 2019



Better service availability and stronger incentives improved vaccination rates.

Abhijit Banerjee, Esther Duflo and Michael Kremer for their experimental approach to alleviating global poverty

https://www.nobelprize.org/prizes/economic-sciences/2019/press-release/

We aim to have comparable units.

$E[Y(1) - Y(0)]$ - Average treatment effect

$$E[Y|do(D=1)] = E[Y(1)] = \underbrace{E[Y(1)|D=1]}_{\textit{observed}} = \underbrace{E[Y(1)|D=0]}_{\textit{unobserved}}$$

$$E[Y|do(D=0)] = E[Y(0)] = \underbrace{E[Y(0)|D=0]}_{\textit{observed}} = \underbrace{E[Y(0)|D=1]}_{\textit{unobserved}}$$

$$E[Y(1)] - E[Y(0)] = E[Y(1)|D=1] - E[Y(0)|D=0] = E[Y|D=1] - E[Y|D=0]$$

$E[Y(1) - Y(0)|D = 1]$ - Average treatment effect on the treated

$$\underbrace{E[Y(1)|D = 1]}_{\textit{observed}} = \underbrace{E[Y(1)|D = 0]}_{\textit{unobserved}}$$

$$E[Y(1) - Y(0)|D = 1] = \underbrace{E[Y(1)|D = 1]}_{\textit{observed}} - \underbrace{E[Y(0)|D = 1]}_{\textit{unobserved}} = E[Y|D = 1] - E[Y|D = 0]$$

Here, only one counterfactual is needed.

Decomposition

$$
\begin{aligned}
E[Y|D=1] - E[Y|D=0] &= \overbrace{\underbrace{E[Y(1)|D=1]}_{E[Y|D=1]} - \underbrace{E[Y(0)|D=1]}_{\textit{unobserved}}}^{ATT = E[Y(1)-Y(0)|D=1]} \\
&+ \underbrace{\underbrace{E[Y(0)|D=1]}_{\textit{unobserved}} - \underbrace{E[Y(0)|D=0]}_{E[Y|D=0]}}_{\textit{Selection bias}}
\end{aligned}
$$

Selection bias is zero under randomization.

Potential problems (not outcomes this time)

- Randomization itself
- Outcome attrition
- Knowing you are in an experiment
- Sample size (expensive)
- External validity
- Non-scalability
- Peer-effects, general equlibrium effects

Duflo, Esther, Rachel Glennerster, and Michael Kremer. "Using randomization in development economics research: A toolkit." Handbook of development economics 4 (2007): 3895-3962.

Some further tips

- Prospective trials often lead to surprises.
- Some programs fail. Beware of publication bias.
- Not only effects we are interested in, but also mechanisms, potential side effects.
- RCTs are costly, difficult, but feasible.
- Spillovers effects are real.

Kremer, Michael. "Randomized evaluations of educational programs in developing countries: Some lessons." American Economic Review 93.2 (2003): 102-106.

Implementation matters too

Important to have a partner company you can trust.

Feature » BMJ Investigation

Covid-19: Researcher blows the whistle on data integrity issues in Pfizer's vaccine trial

BMJ 2021 ; 375 doi: https://doi.org/10.1136/bmj.n2635 (Published 02 November 2021)
Cite this as: *BMJ* 2021;375:n2635

Read our latest coverage of the coronavirus pandemic

Article | Related content | Metrics | Responses

Paul D Thacker, investigative journalist

Author affiliations

Revelations of poor practices at a contract research company helping to carry out Pfizer's pivotal covid-19 vaccine trial raise questions about data integrity and regulatory oversight. **Paul D Thacker** reports

Example: Tennessee STAR experiment

- **S**tudent **T**eacher **A**chievment **R**atio
- Do smaller classes make sense?
- They are expensive.
- Cost $12mil and implemented on 11600 kids in kindergartens in 1985/86
- Long, expensive, logistically difficult
- Useful benchmark, but we might want to learn about the effects sooner

You can try to work with it on your own https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766

Example: Tennessee STAR experiment

Apples to apples?

Students who entered STAR in kindergarten

	Variable	Small	Regular	Regular/Aide	Joint P-value
1.	Free lunch	.47	.48	.50	.09
2.	White/Asian	.68	.67	.66	.26
3.	Age in 1985	5.44	5.43	5.42	.32
4.	Attrition rate	.49	.52	.53	.02
5.	Class size in kindergarten	15.10	22.40	22.80	.00
6.	Percentile score in kindergarten	54.70	48.90	50.00	.00

Table 2.2.1 of Angrist and Pischke (2009)

RCT and regression

$$Y = \underbrace{\alpha}_{E(Y(0))} + \underbrace{\rho}_{Y(1)-Y(0)} D + \underbrace{\eta}_{Y(0)-E(Y(0))}$$

$$\Longrightarrow$$

$$\begin{aligned} E[Y|D=1] &= \alpha + \rho + E[\eta|D=1] \\ E[Y|D=0] &= \alpha + E[\eta|D=0] \end{aligned}$$

$$E[Y|D=1] - E[Y|D=0] = \underbrace{\rho}_{\textit{treatment effect}} + \underbrace{E[\eta|D=1] - E[\eta|D=0]}_{\textit{selection bias}}$$

if we assume that ρ is non-random (homogenous treatment effects)

RCT and regression + covariates

- assignment was random only within schools - add schools specific intercept
- inclusion of covariates may improve the statistical precision of ρ estimate

$$Y = \alpha + \rho D + X^T \gamma + \eta$$

Note:

- we still assume homogenous treatment effects
- we now assume a specific linear form how X is connected to Y
- this may be thought of as an approximation

[Adjusting for X in RCT or not? See Negi and Wooldridge 2021.]

Example: Tennessee STAR experiment

Experimental estimates of the effect of class-size assignment on test scores

Explanatory variable	(1)	(2)	(3)	(4)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.26)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes)	–	–	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	–	–	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	–	–	-13.15	-13.07
			(.77)	(.77)
White teacher	–	–	–	-.57
				(2.10)
Teacher experience	–	–	–	.26
				(.10)
Master's degree	–	–	–	-0.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

reassuring

Table 2.2.2 of Angrist and Pischke (2009)

Selection on observables

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

- We rarely have the luxury of an RCT, especially in economics.
- Observational data may be useful in recovering causal relationship.
- This often requires modelling and deep institutional knowledge.
- Sometimes we have something that resembles RCT, we will discuss this later

Assume that the richness of X allows us to close all the backdoor paths from D to Y.

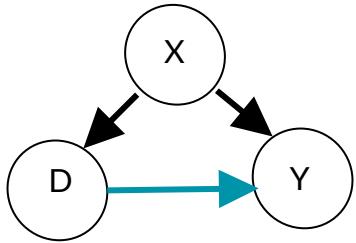
Selection on observables

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

It has various labels:

- Conditional independence assumption
- Unconfoundedness
- Ignorability

Selection on observables



How realistic is this model?

- Well, obviously: it depends.
- If you have rich set of information (many many variables X), it might be fine.
- But then it is tricky to model, you also need large data set.
- Within a large data set, units are very different and homogeneity makes rarely sense.

Selection on observables

- Identification is straightforward.
- There are, however, different statistical techniques how to estimate the effects.

We will cover these classes of estimation techniques:

- Regression
- Matching
- Propensity score weighting

What do they have in common?

- Estimated from observation data.
- There is no randomization, no quasi-randomization involved.

Regression

We know a lot about the mechanics of the linear regression, projections etc.

In the first part of the course we were silent about the causal interpretation. We have assumed that the model is correctly specified.

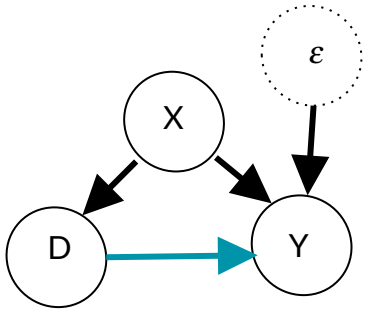
$$Y = \alpha + \rho D + X^T \gamma + \varepsilon_i$$

$$\begin{aligned} E[Y(1)|X] &= E[Y|X, D=1] = \alpha + \rho + X^T \gamma \\ E[Y(0)|X] &= E[Y|X, D=0] = \alpha + X^T \gamma \end{aligned}$$

For a simple linear model - no heterogeneity:

$$ATT = E[Y(1) - Y(0)|X] = \rho = E[Y(1) - Y(0)] = ATE$$

We made use of $E[\varepsilon|X, D] = 0$



Linearity?

$$Y = f(D, X) + \varepsilon_i$$

$$\begin{aligned} E[Y(1)|X, D=1] &= E[Y|X, D=1] = f(1, X) \\ E[Y(0)|X, D=1] &\underbrace{=}_{C.I.A.} E[Y|X, D=0] = f(0, X) \end{aligned}$$

$$\delta_X \equiv E[Y|X, D=1] - E[Y|X, D=0]$$

$$E[Y(1) - Y(0)|X, D=1] = f(1, X) - f(0, X) \underbrace{=}_{C.I.A.} \delta_X$$

$$\begin{aligned} E[Y(1) - Y(0)|D=1] &= E\big[E[Y(1) - Y(0)|X, D=1]\big] = \sum_x \delta_x Pr(X=x|D=1) \\ E[Y(1) - Y(0)] &= E\big[E[Y(1) - Y(0)|X]\big] = \sum_x \delta_x Pr(X=x) \end{aligned}$$

Matching

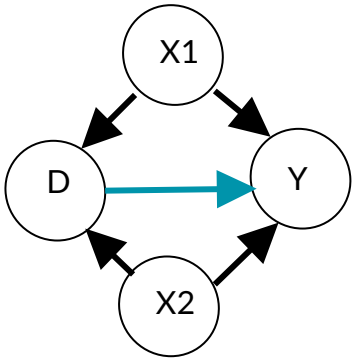
Matching is a class of statistical techniques that takes:

We aim to have comparable units.

very seriously.

Example: Matching - Titanic

- 700 out of 2200 on board survived
- did wealth affected survival probability?
- women and children were given priority, but they were also likely to be in the first class



- D - first class
- X_1 - gender
- X_2 - age (old/young)
- Y - survived

Two back-door paths.
Any unobserved confounders are ruled out.

Example: Matching - Titanic

- 4 categories: {young male, young female, old male, old female}

$$
\begin{aligned}
E[Y|D=1] - E[Y|D=0] &= 0.354 \\
E[Y(1) - Y(0)] &= \sum_x \delta_x Pr(X=x) = 0.196 \\
E[Y(1) - Y(0)|D=1] &= \sum_x \delta_x Pr(X=x|D=1) = 0.238 \\
E[Y(1) - Y(0)|D=0] &= \sum_x \delta_x Pr(X=x|D=0) = 0.189
\end{aligned}
$$

- By stratification we lose information.
- As a reward, we get something that is easy to interpret and implement.
- If we do not stratify, we may have few observations in a certain group. There is no 12yo boy in the first class.

$$ATT = \sum_{k=1}^{K} (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \frac{N_T^k}{N_T}$$

$$ATC = \sum_{k=1}^{K} (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \frac{N_C^k}{N_C}$$

$$ATE = \sum_{k=1}^{K} (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \frac{N^k}{N}$$

- K different categories
- $\bar{Y}^{1,k}$ - mean outcome of treated in group k
- $\bar{Y}^{0,k}$ - mean outcome of control in group k
- N_T^k, N_C^k, N^k - number of treated, controls, overall within category k
- N_T, N_C, N - number of treated, controls, overall

Example: Matching - Angrist (1998)

- Voluntary military service. How did it affect wages?
- Military was the largest employer.
- Military size declined sharply in 1987.
- Compares applicants. 50% of them enlisted.
- Applicants are not chosen at random

Example: Matching - Angrist (1998)

- 698'000 observations
- Information in X: year of application, test score group, schooling level, year of birth.
- Heterogenous across race: Separate estimates for Whites and Non-whites
- 8'760 cells, but only 5'654 had at least 25 observations

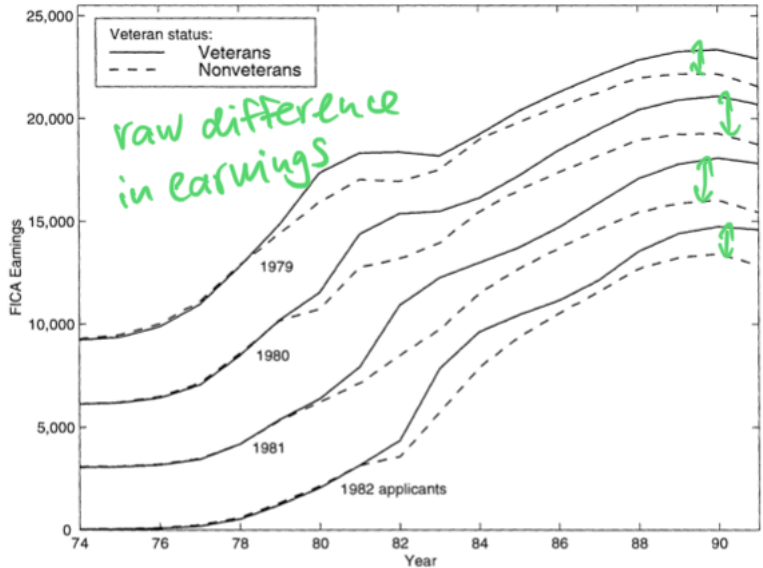
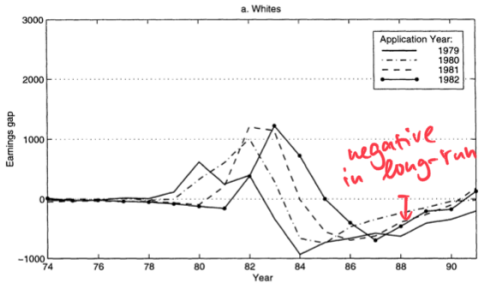


Fig.2 in Angrist (1998)



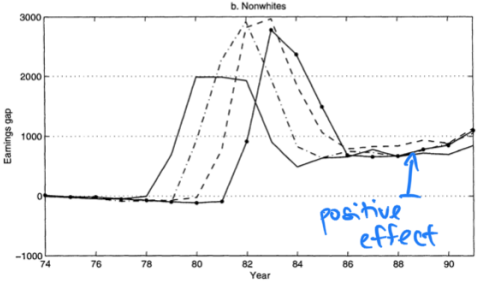


Fig.3 in Angrist (1998)

Year	Whites			
	Mean (1)	Difference in Means[c] (2)	Controlled Contrast (3)	Regression Estimates (4)
A. *Earnings*[a]		raw	Matching	Reg.
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)

Part of Table 2 in Angrist (1998)

Matching vs. Regression

These results differ. Why?

Explore the simplest possible case. Binary X.

Binary X

Saturated model (heterogenous effects)

$$Y = \beta_0 + \beta_1 X + \delta_0 D(1-X) + \delta_1 DX$$

$$\begin{aligned} \delta_1 &= E[Y|X=1, D=1] - E[Y|X=1, D=0] \\ \delta_0 &= E[Y|X=0, D=1] - E[Y|X=0, D=0] \end{aligned}$$

Non saturated model (homogenous effects)

$$Y = \alpha + \rho D + \gamma X + \varepsilon_i$$

CATT is assumed to be the same for both $X = 1$ and $X = 0$

Saturated model (heterogenous effects)

$$Y = \beta_0 + \beta_1 X + \delta_0 D(1-X) + \delta_1 DX$$

$$\begin{aligned}
\delta_1 &= E[Y|X=1,D=1] - E[Y|X=1,D=0] \\
\delta_0 &= E[Y|X=0,D=1] - E[Y|X=0,D=0]
\end{aligned}$$

$$\begin{aligned}
& E[Y(1)-Y(0)|D=1] = \sum_x \delta_x Pr(X=x|D=1) \\
&= \delta_0 Pr(X=0|D=1) + \delta_1 Pr(X=1|D=1) \\
&= \delta_0 \frac{Pr(D=1|X=0)\cdot P(X=0)}{P(D=1)} + \delta_1 \frac{Pr(D=1|X=1)\cdot P(X=1)}{P(D=1)} \\
&= \delta_0 w_0^M + \delta_1 w_1^M
\end{aligned}$$

Non-saturated model (homogenous effects)

Non saturated model (homogenous effects)

$$Y = \alpha + \rho D + \gamma X + \varepsilon_i$$

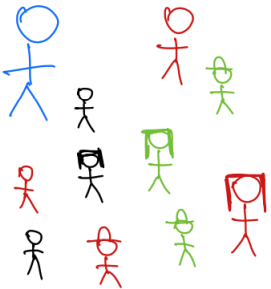
$$\begin{aligned}
\hat{\rho} &= \ldots [3.3.1 \textit{ in Angrist and Pischke} (2009)] \ldots \\
&= \frac{\sum_x \delta_x [Pr(D=1|X=x)(1-Pr(D=1|X=x))] Pr(X=x)}{\sum_x [Pr(D=1|X=x)(1-Pr(D=1|X=x))] Pr(X=x)} \\
&= \delta_0 w_0^R + \delta_1 w_1^R
\end{aligned}$$

Comparison - Matching vs Regression

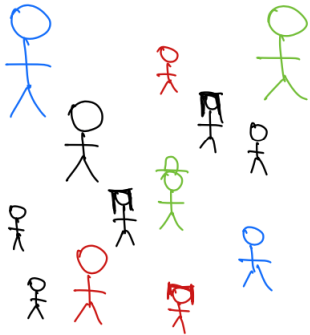
$$w_x^M = \frac{\overbrace{Pr(D=1|X=x))}^{\sim \textit{share of treated among } X=x}}{P(D=1)} \cdot Pr(X=x)$$

$$w_x^R = \frac{\overbrace{Pr(D=1|X=x)(1-Pr(D=1|X=x))}^{\sim \textit{variance of } D \textit{ given } X=x}}{\sum_x [Pr(D=1|X=x)(1-Pr(D=1|X=x))]Pr(X=x)} \cdot Pr(X=x)$$

Matching

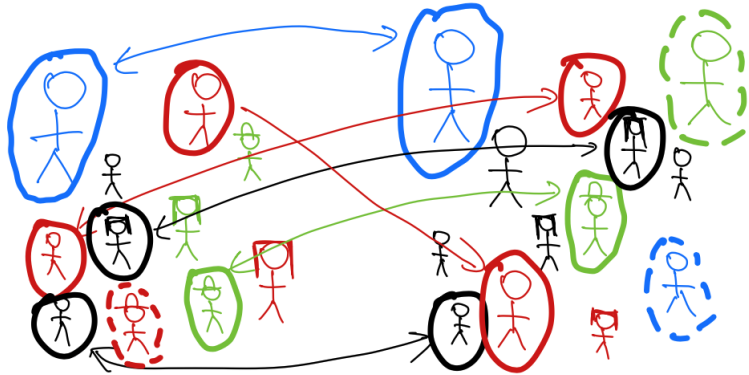


TREATED
D=1



CONTROLS
D=0

Matching



Different types of Matching

- In many interesting cases exact matches are not possible
- We need to introduce some measure on how similar different units are
- There are many ways how this can be done

Overlap

$$0 \underbrace{<}_{\text{for } E[Y(1)|D=0]} P(D=1|X) \underbrace{<}_{\text{for } E[Y(0)|D=1]} 1$$

- It is important to have comparable units.
- If we don't we may drop these observations or we may rely on extrapolation.
- Dropping observations means we estimate effects only on a subpopulation, so the object of interest changes.
- You don't want to extrapolate much, but, at the same time, you want to have your effect representative enough.

One to one matching

$$\widehat{ATT} = \frac{1}{N_T} \sum_{i:D_i=1} (Y_i - Y_{j(i)})$$

$j(i)$ is "similar" to i in terms of X in the control group

We compare Y_i to the similar unit

One to many matching

$$\widehat{ATT} = \frac{1}{N_T} \sum_{i:D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)} \right)$$

$j_m(i)$ is one of the M "similar" units from the control group to i in terms of X

We compare Y_i to the average of the similar units

Nearest neighbour covariate matching

How to measure how similar the units are?

$$||X_i - X_j|| = \sqrt{(X_i - X_j)^T (X_i - X_j)} = \sqrt{\sum_{n=1}^{p} (X_{ni} - X_{nj})^2}$$

Or weight by the variance

$$||X_i - X_j|| = \sqrt{(X_i - X_j)^T \hat{V}^{-1} (X_i - X_j)} = \sqrt{\sum_{n=1}^{p} \frac{(X_{ni} - X_{nj})^2}{\hat{\sigma}_n^2}}$$

Or weight by the covariance

$$||X_i - X_j|| = \sqrt{(X_i - X_j)^T \hat{\Sigma}^{-1} (X_i - X_j)}$$

Bias

- The larger the dimension of X, the more difficult is to find matches
- Data greedy
- X_i converges to $X_{j(i)}$ only slowly

Bias corrected matching estimator

$$\widehat{ATT}_{BC} = \frac{1}{N_T} \sum_{i:D_i=1} \Big((Y_i - Y_{j(i)}) - \underbrace{(\hat{E}[Y|X = X_i, D = 0] - \hat{E}[Y|X = X_{j(i)}, D = 0])}_{\text{bias correction term}} \Big)$$

Variance?

Without replacement

Use control units only once.

$$\hat{\sigma}^2_{ATT} = \frac{1}{N_T} \sum_{i:D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)} - \widehat{ATT} \right)^2$$

With replacement

Use control units possibly more than once.

$$\hat{\sigma}^2_{ATT} = \frac{1}{N_T} \sum_{i:D_i=1} \left(Y_i - \frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)} - \widehat{ATT} \right)^2 + \frac{1}{N_T} \sum_{i:D_i=1} \left(\frac{K_i(1-K_i)}{M^2} \frac{(Y_i - Y_j)^2}{2} \right)$$

(in this particular case bootstrap fails - Abadie and Imbens (2008)

Matching vs Regression - Practical considerations

- There are many different ways how one can perform matching.
- There are many different ways how one can perform regression.
- Researchers degree of freedom is a problem.
- Matching is appealing because it is easy to communicate to outsiders.
- Regression is appealing as there seems to be (or are?) fewer degrees of freedom

2.2 *Regression Estimators*

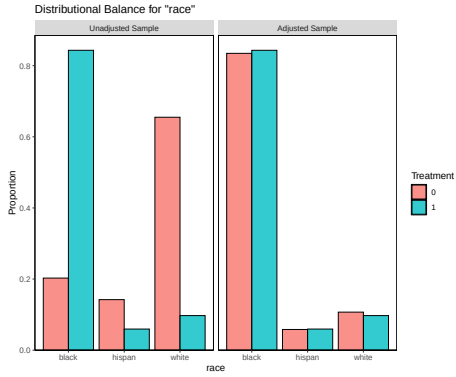
Differences between regression and matching strategies for the estimation of treatment effects are partly cosmetic. While matching methods are often more transparent to nonspecialists, regression estimation is more straightforward to implement when covariates are continuously distributed because matching on continuous covariates requires stratification or pairing (Cochran (1968)). Note, however, that both methods require a similar sort of approximation since regression on continuous covariates in any finite sample requires functional form restrictions. The fact that both stratification and functional form approximations can be made increasingly accurate as the sample size grows suggests that the manner in which continuous covariates are accommodated is not the most important difference between the two methods.

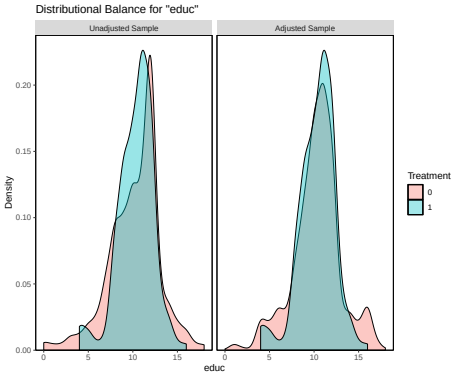
Angrist (1998), page 255: Angrist, Joshua. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." Econometrica 66.2 (1998): 249-288.

Example: LaLonde (1986)

- Very influential study.
- Does job training increase future wages?
- Having randomized treatment (**NSW** - National Support Work), LaLonde can compare matching estimators (from two different observational datasets: CPS - Current Population Survey and PSID - Panel Survey of Income Dynamics) to the one from the randomized, which served as a benchmark
- Results pessimistic: Estimates from obs. datasets are all over the place!
- E.g. **$800** vs -$8000 vs - $4400
- Well, the samples are very different

- It is important to check how comparable treated and controls are in our matched sample
- This is called a balance
- The success of matching can be shown using a balance graph.
- Excellent implementation is in `MatchIt` package in R



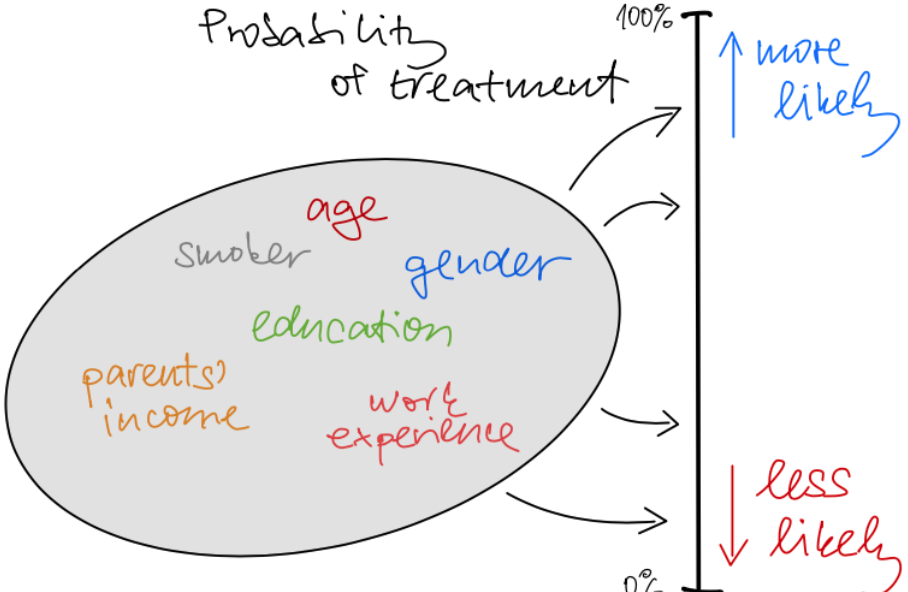


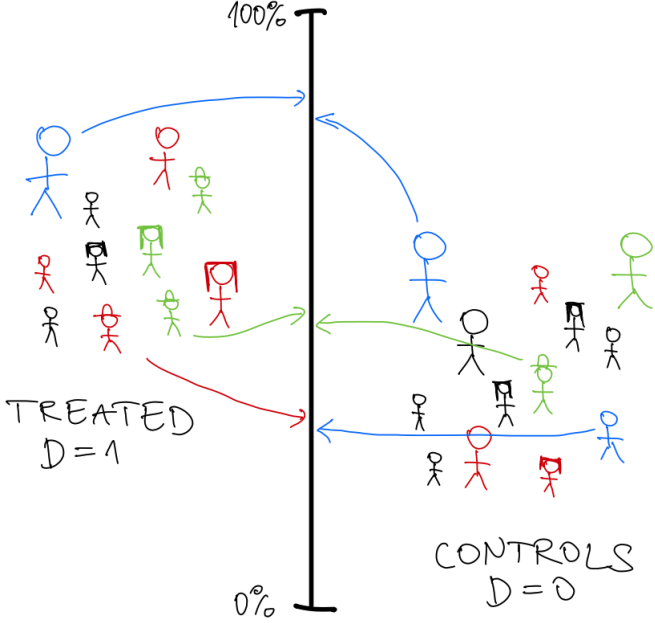
Propensity score

$$p(x) = P(D = 1|X = x)$$

We may skip the high-dimensionality of X in a very neat way.

Projecting them on the quantity that matters - probability of treatment


Probability of treatment
100%
0%
more likely
less likely
age
smoker
gender
education
parents' income
work experience


100%
0%
TREATED
D=1
CONTROLS
D=0

Propensity score matching

This idea comes from Donald Rubin (e.g. Rubin,1977) and Paul Rosenbaum (Rosenbaum and Rubin, 1983, over 30k citations).

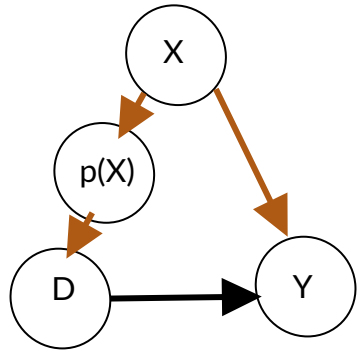
$$Y(0), Y(1) \perp\!\!\!\perp D \mid X$$

$$\Longrightarrow$$

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

$$\begin{aligned}
Pr(D=1|Y(1),Y(0),p(X)) &= E(D|Y(1),Y(0),p(X)) = E(E(D|Y(1),Y(0),X)|Y(1),Y(0),p(X)) \\
&= E(E(D|X)|Y(1),Y(0),p(X)) = E(p(X)|Y(1),Y(0),p(X)) = p(X) \\
&= p(X) = E(p(X)|p(X)) = E(E(D|X,p(X))|p(X)) = E(D|p(X)) \\
&= Pr(D=1|p(X)). \quad \square
\end{aligned}$$

Propensity score matching



Conditioning on $p(X)$ closes the backdoor path.

Also, notice that $D \perp\!\!\!\perp X | p(X)$ as Y is the collider on the path.

Propensity score matching

$$\delta_{p(X)} = E(Y|D=1, p(X)) - E(Y|D=0, p(X))$$

$$E[Y(1) - Y(0)|D=1] = E[\delta_{p(X)}|D=1]$$

Propensity score matching

1. Use logit/probit to estimate propensity scores.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = X^T\beta$$

2. Sort observations according to $\hat{p}(X)$
3. Stratify sample to blocks so that mean scores are not statistically different among treated and controls
4. Check for balance. If no balance within a block $\rightarrow$ split the block. If, for some variable, no balance in all the blocks $\rightarrow$ check the model specification in Step 1.

Implemented in Stata by Becker, Sascha O., and Andrea Ichino. "Estimation of average treatment effects based on propensity scores." The stata journal 2.4 (2002): 358-377.

There are other ways how PS matching can be implemented

- Nearest neighbour matching
- Radius matching
- Kernel matching - weight controls by a Kernel function - those controls close to propensity score of the treated get larger weight

Example: Dehejia and Wahba (2002)

- Use data from LaLonde (1986)
- Compares randomized NSW data to two observational datasets: CPS and PSID

PS Matching in detail

- With or without replacement? Smaller PS distance vs. Fewer comparison units.
- How many comparison units? Smaller PS distance vs. Increased precision.
- Which matching method to use? Caliper matching can use more (fewer) matches if (not) available.

If overlap is good, different matching will lead to similar results.

TABLE 1.—SAMPLE MEANS AND STANDARD ERRORS OF COVARIATES FOR MALE NSW PARTICIPANTS

National Supported Work Sample (Treatment and Control)		
	Dehejia-Wahba Sample	
Variable	Treatment (D=1)	Control (D=0)
Age	25.81 (0.52)	25.05 (0.45)
Years of schooling	10.35 (0.15)	10.09 (0.1)
Proportion of school dropouts	0.71 (0.03)	0.83 (0.02)
Proportion of blacks	0.84 (0.03)	0.83 (0.02)
Proportion of Hispanic	0.06 (0.017)	0.10 (0.019)
Proportion married	0.19 (0.03)	0.15 (0.02)
Number of children	0.41 (0.07)	0.37 (0.06)
No-show variable	0 (0)	n/a
Month of assignment (Jan. 1978 = 0)	18.49 (0.36)	17.86 (0.35)
Real earnings 12 months before training	1,689 (235)	1,425 (182)
Real earnings 24 months before training	2,096 (359)	2,107 (353)
Hours worked 1 year before training	294 (36)	243 (27)
Hours worked 2 years before training	306 (46)	267 (37)
Sample size	185	260

- National Supported Work Program
- Provided work experience to people with social problems
- Here is a randomized sample from LaLonde (1986)

FIGURE 1.—HISTOGRAM OF ESTIMATED PROPENSITY SCORE, NSW AND CPS

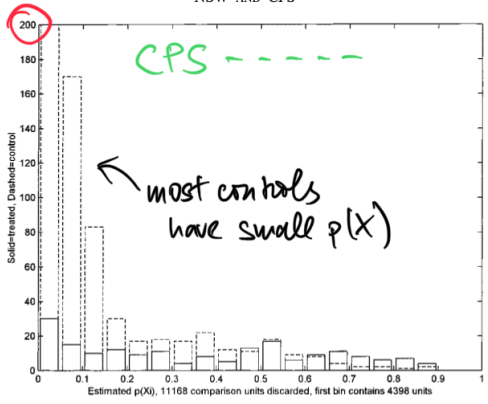


FIGURE 2.—HISTOGRAM OF ESTIMATED PROPENSITY SCORE, NSW AND PSID

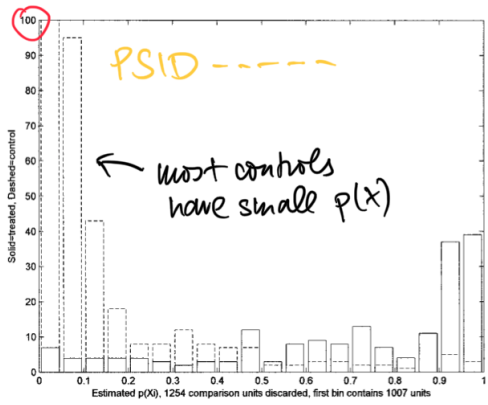


Fig 1 and 2 from Dehejia and Wahba (2002)

FIGURE 5.—PROPENSITY SCORE FOR TREATED AND MATCHED COMPARISON UNITS, HIGHEST TO LOWEST

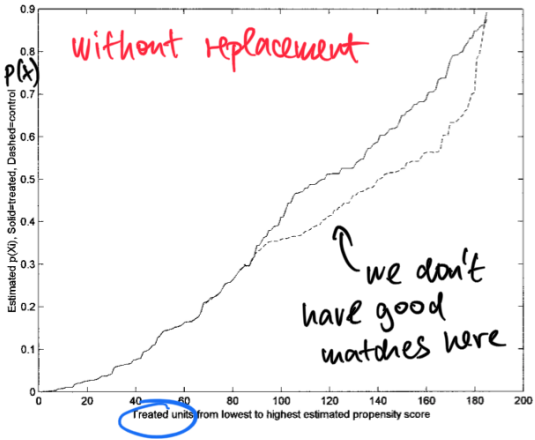


FIGURE 6.—PROPENSITY SCORE FOR TREATED AND MATCHED COMPARISON UNITS, NEAREST MATCH

CPS

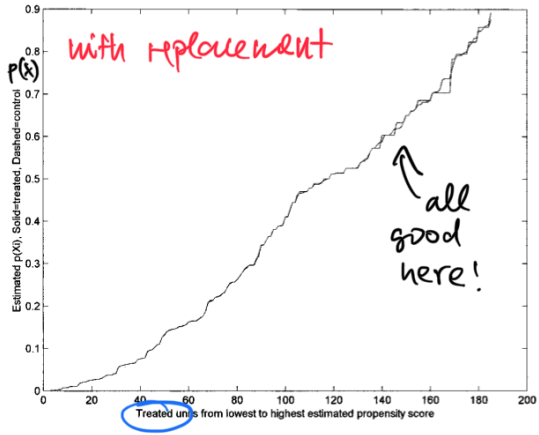


Fig 5 and 6 from Dehejia and Wahba (2002)

TABLE 2.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND CPS SAMPLES

CPS data

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74	RE75	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full CPS	15992	0.01	33.23	12.03	0.07	0.07	0.30	0.71	14017	13651	0.88	0.89	−8498	1066
		(0.02)[D]	(0.53)	(0.15)	(0.03)	(0.02)	(0.03)	(0.03)	(367)	(248)	(0.03)	(0.04)	(583)[E]	(554)
Without replacement:														
Random	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
Low to high	185	0.32	25.23	10.28	0.84	0.06	0.66	0.22	2286	1687	0.37	0.51	1605	1681
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(730)	(704)
High to low	185	0.32	25.26	10.30	0.84	0.06	0.65	0.22	2305	1687	0.37	0.51	1559	1651
		(0.03)	(0.79)	(0.23)	(0.04)	(0.03)	(0.05)	(0.04)	(495)	(341)	(0.05)	(0.05)	(733)	(709)
With replacement:														
Nearest neighbor	119	0.37	25.36	10.31	0.84	0.06	0.69	0.17	2407	1516	0.35	0.49	1360	1375
		(0.03)	(1.04)	(0.31)	(0.06)	(0.04)	(0.07)	(0.06)	(727)	(506)	(0.07)	(0.07)	(913)	(907)
Caliper, $\delta = 0.00001$	325	0.37	25.26	10.31	0.84	0.07	0.69	0.17	2424	1509	0.36	0.50	1119	1142
		(0.03)	(1.03)	(0.30)	(0.06)	(0.04)	(0.07)	(0.06)	(845)	(647)	(0.06)	(0.06)	(875)	(874)
Caliper, $\delta = 0.00005$	1043	0.37	25.29	10.28	0.84	0.07	0.69	0.17	2305	1523	0.35	0.49	1158	1139
		(0.02)	(1.03)	(0.32)	(0.05)	(0.04)	(0.06)	(0.06)	(877)	(675)	(0.06)	(0.60)	(852)	(851)
Caliper, $\delta = 0.0001$	1731	0.37	25.19	10.36	0.84	0.07	0.69	0.17	2213	1545	0.34	0.50	1122	1119
		(0.02)	(1.03)	(0.31)	(0.05)	(0.04)	(0.06)	(0.06)	(890)	(701)	(0.06)	(0.06)	(850)	(843)

Table 2 from Dehejia and Wahba (2002)

TABLE 3.—SAMPLE CHARACTERISTICS AND ESTIMATED IMPACTS FROM THE NSW AND PSID SAMPLES

Control Sample	No. of Observations	Mean Propensity Score[A]	Age	School	Black	Hispanic	No Degree	Married	RE74 US$	RE75 US$	U74	U75	Treatment Effect (Diff. in Means)	Regression Treatment Effect
NSW	185	0.37	25.82	10.35	0.84	0.06	0.71	0.19	2095	1532	0.29	0.40	1794[B]	1672[C]
													(633)	(638)
Full PSID	2490	0.02	34.85	12.12	0.25	0.03	0.31	0.87	19429	19063	0.10	0.09	−15205	4
		(0.02)[D]	(0.57)	(0.16)	(0.03)	(0.02)	(0.03)	(0.03)	(449)	(361)	(0.04)	(0.03)	(657)[E]	(1014)
Without replacement:														
Random	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1035)	(983)
Low to high	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1135)	(983)
High to low	185	0.25	29.17	10.30	0.68	0.07	0.60	0.52	4659	3263	0.40	0.40	−916	77
		(0.03)	(0.90)	(0.25)	(0.04)	(0.03)	(0.05)	(0.05)	(554)	(361)	(0.05)	(0.05)	(1135)	(983)
With replacement:														
Nearest Neighbor	56	0.70	24.81	10.72	0.78	0.09	0.53	0.14	2206	1801	0.54	0.69	1890	2315
		(0.07)	(1.78)	(0.54)	(0.11)	(0.05)	(0.12)	(0.11)	(1248)	(963)	(0.11)	(0.11)	(1202)	(1131)
Caliper, $\delta = 0.00001$	85	0.70	24.85	10.72	0.78	0.09	0.53	0.13	2216	1819	0.54	0.69	1893	2327
		(0.08)	(1.80)	(0.56)	(0.12)	(0.05)	(0.12)	(0.12)	(1859)	(1896)	(0.10)	(0.11)	(1198)	(1129)
Caliper, $\delta = 0.00005$	193	0.70	24.83	10.72	0.78	0.09	0.53	0.14	2247	1778	0.54	0.69	1928	2349
		(0.06)	(2.17)	(0.60)	(0.11)	(0.04)	(0.11)	(0.10)	(1983)	(1869)	(0.09)	(0.09)	(1196)	(1121)
Caliper, $\delta = 0.0001$	337	0.70	24.92	10.73	0.78	0.09	0.53	0.14	2228	1763	0.54	0.70	1973	2411
		(0.05)	(2.30)	(0.67)	(0.11)	(0.04)	(0.11)	(0.09)	(1965)	(1777)	(0.07)	(0.08)	(1191)	(1122)
Caliper, $\delta = 0.001$	2021	0.70	24.98	10.74	0.79	0.09	0.53	0.13	2398	1882	0.53	0.69	1824	2333
		(0.03)	(2.37)	(0.70)	(0.09)	(0.04)	(0.10)	(0.07)	(2950)	(2943)	(0.06)	(0.06)	(1187)	(1101)

Table 3 from Dehejia and Wahba (2002)

Lessons to take

- When few control units are available, use sampling with replacement (you can use the same control twice)
- When enough control units are available, sampling without replacement would be fine
- Careful diagnostics aid the right choices.
- So perhaps it is not as bad as LaLonde (1986) suggested?

→ Reply: Smith, Jeffrey A., and Petra E. Todd. "Does matching overcome LaLonde's critique of nonexperimental estimators?." Journal of econometrics 125.1-2 (2005): 305-353.

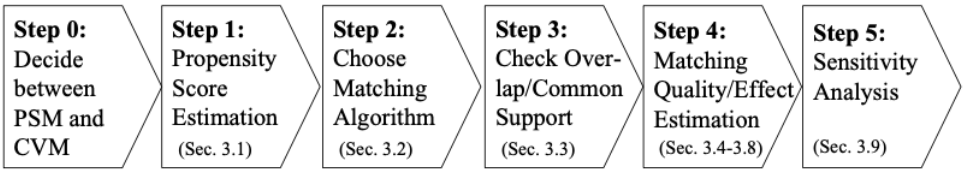
- Results are sensitive to covariates in PS estimation and to choice of the sample.
- PSM "...does not represent a general solution to the evaluation problem"

→ Rejoinder: Dehejia, Rajeev. "Practical propensity score matching: a reply to Smith and Todd." Journal of econometrics 125.1-2 (2005): 355-364.

- Yes, one should check the sensitivity of estimates to the PS model specification.
- High quality comparison group should not be too sensitive.
- With this on your mind, PSM works fine. Even in the different subsamples of LaLonde (1996)

Implementation issues

There are other ways how PS matching can be implemented



CVM: Covariate Matching, PSM: Propensity Score Matching

Fig 1 in Caliendo, Marco, and Sabine Kopeinig. "Some practical guidance for the implementation of propensity score matching." Journal of economic surveys 22.1 (2008): 31-72.

Inverse Propensity Score Weighting

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

$$\Longrightarrow$$

$$ATE = E[Y(1)] - E[Y(0)] = E\left[\frac{Y \cdot D}{p(X)}\right] - E\left[\frac{Y \cdot (1-D)}{1-p(X)}\right]$$

$$ATT = E[Y(1)|D=1] - E[Y(0)|D=1] = E[Y \cdot D] - E\left[Y \cdot (1-D)\frac{p(X)}{1-p(X)}\right]$$

Inverse Propensity Score Weighting

$$
\begin{aligned}
E\left[\frac{Y \cdot D}{p(X)}\right] &= E\left[E\left[\frac{Y \cdot D}{p(X)}|X\right]\right] = E\left[E\left[\frac{Y(1)}{p(X)}|D=1,X\right] Pr(D=1|X)\right] \\
&= E\left[E\left[\frac{Y(1)}{p(X)}|D=1,X\right] p(X)\right] = E\left[E\left[Y(1)|D=1,X\right]\right] = E[Y(1)]
\end{aligned}
$$

and other quantities similarly.

Inverse Propensity Score Weighting

First: estimate $\hat{p}$.

Then:

$$\widehat{ATE} = \frac{1}{N}\sum_i \frac{Y_i D_i}{\hat{p}(X_i)} - \sum_i \frac{Y_i(1-D_i)}{1-\hat{p}(X_i)}$$

$$\widehat{ATT} = \frac{1}{N}\sum_i Y_i D_i - \sum_i Y_i(1-D_i)\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}$$

Inverse Propensity Score Weighting

Normalized versions (more stable):

$$\widehat{ATE} = \left[\frac{1}{N}\sum_i \frac{Y_i D_i}{\hat{p}(X_i)}\right] \Big/ \left[\frac{1}{N}\sum_i \frac{D_i}{\hat{p}(X_i)}\right] - \left[\sum_i \frac{Y_i(1-D_i)}{1-\hat{p}(X_i)}\right] \Big/ \left[\sum_i \frac{(1-D_i)}{1-\hat{p}(X_i)}\right]$$

$$\widehat{ATT} = \left[\frac{1}{N}\sum_i Y_i D_i\right] \Big/ \left[\frac{1}{N}\sum_i D_i\right] - \left[\sum_i Y_i(1-D_i)\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}\right] \Big/ \left[\sum_i (1-D_i)\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}\right]$$

Weigthing: Hirano and Imbens (2001)

Performance for different constructions of standard errors: Bodory, Camponovo, Huber, and Lechner, (2020)

R package `treatweight` by Bodory and Huber (2021)

- Sensitive to specification of $p(\cdot)$
- May require trimming
- Does not rely on stratification nor matching (less degrees of freedom?)
- Standard errors need to take into account that the propensity scores are only estimated (Hirano, Imbens and Ridder, 2003)

Wrap-up

There are different ways how we can estimate the quantity of interest (e.g. ATE, ATT) if our observables are informative in explaining the selection bias.

Regression, Matching, IPW.

They all have pros and cons.

It is the selection on observables assumption that drive the identification. Without this, any estimator is dubious at best.

Thank you for your attention!

References

- Tips and tricks on implementation of randomization: Duflo, Esther, Rachel Glennerster, and Michael Kremer. "Using randomization in development economics research: A toolkit." Handbook of development economics 4 (2007): 3895-3962.
- This book is a classic. Somewhat opinionated. By the pioneers of the field: Angrist, Joshua D., and Jörn-Steffen Pischke. Mostly harmless econometrics. Princeton university press, 2008.
- Very readable and engaging book, highly recommended: Cunningham, Scott. Causal Inference. Yale University Press, 2021.
- Adjusting for X in RCT? Negi, Akanksha, and Jeffrey M. Wooldridge. "Revisiting regression adjustment in experiments with heterogeneous treatment effects." Econometric Reviews 40.5 (2021): 504-534. Or this twitter summary: https://twitter.com/jmwooldridge/status/1457001530985492495?s=21
- Angrist, Joshua. "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants." Econometrica 66.2 (1998): 249-288.
- Some practical recommendations on Matching: Imbens, Guido W. "Matching methods in practice: Three examples." Journal of Human Resources 50.2 (2015): 373-419.
- Why bootstrap fails in matching: Abadie, Alberto, and Guido W. Imbens. "On the failure of the bootstrap for matching estimators." Econometrica 76.6 (2008): 1537-1557.
- Book length treatment of causal inference. Long and very rich and detailed exposition. Imbens, Guido W., and Donald B. Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015.
- Rubin, Donald B. "Assignment to treatment group on the basis of a covariate." Journal of educational Statistics 2.1 (1977): 1-26.
- PSM paper: Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70.1 (1983): 41-55.
- Large sample theory for PS matching: Abadie, Alberto, and Guido W. Imbens. "Matching on the estimated propensity score." Econometrica 84.2 (2016): 781-807.
- Very popular article on implementation issues in PSM: Caliendo, Marco, and Sabine Kopeinig. "Some practical guidance for the implementation of propensity score matching." Journal of economic surveys 22.1 (2008): 31-72.

References

- Pessimistic view on policy evaluations based on observational data: LaLonde, Robert J. "Evaluating the econometric evaluations of training programs with experimental data." The American economic review (1986): 604-620.
- Addressing the LaLonde critique: Dehejia, Rajeev H., and Sadek Wahba. "Propensity score-matching methods for nonexperimental causal studies." Review of Economics and statistics 84.1 (2002): 151-161.
- Reply: Smith, Jeffrey A., and Petra E. Todd. "Does matching overcome LaLonde's critique of nonexperimental estimators?." Journal of econometrics 125.1-2 (2005): 305-353.
- Rejoinder: Dehejia, Rajeev. "Practical propensity score matching: a reply to Smith and Todd." Journal of econometrics 125.1-2 (2005): 355-364.
- IPW estimator: Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. "Efficient estimation of average treatment effects using the estimated propensity score." Econometrica 71.4 (2003): 1161-1189.
- Hirano, Keisuke, and Guido W. Imbens. "Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization." Health Services and Outcomes research methodology 2.3 (2001): 259-278.
- Bodory, H., Camponovo, L., Huber, M., and Lechner, M. "The finite sample performance of inference methods for propensity score matching and weighting estimators." Journal of Business & Economic Statistics 38.1 (2020): 183-200.