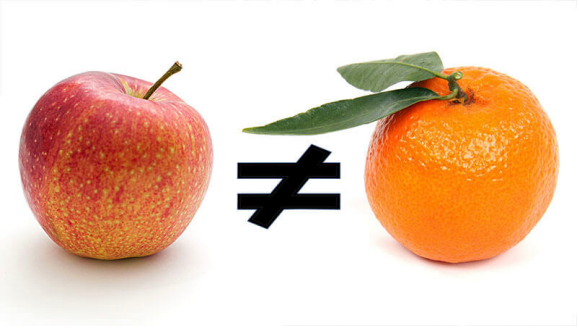
Introduction to Empirical Economics

Confounding & Matching

Léo Zabrocki
leo.zabrocki@psemail.eu
https://lzabrocki.github.io/

École Normale Supérieure



After holidays

One big course on all causal inference methods

Then focus on specific topics

I will invite young researchers to present their work

Materials of the course

https://github.com/lzabrocki/empirical_economics

Today's slides are heavily based on the fantastic ones made by Matt Blackwell & Gary King

Who can summarize the previous class?

Fundamental Problem of Causal Inference

$$Y_i = W_i \times Y_i(1) + (1 - W_i)Y_i(0)$$

The Fertilizer Experiment

Field i	W_i	$Y_i(0)$	$Y_i(1)$	τ_i	Yield (pounds)
1	0	29.2	?	?	?
2	0	11.4	?	?	?
3	1	?	26.6	?	26.6
4	1	?	23.7	?	23.7
5	0	25.3	?	?	25.3
6	1	?	28.5	?	28.5
7	1	?	14.2	?	14.2
8	1	?	17.9	?	17.9
9	0	16.5	?	?	?
10	0	21.1	?	?	21.1
11	1	?	24.3	?	24.3

Science Table

Randomized Controlled Trial

Assignment to the treatment is controlled by the researcher.

Positivity: assignment is probabilistic: $0 < P(W_i = 1) < 1$

Unconfoundedness: $W_i \perp\!\!\!\perp (Y(1), Y(0))$

The Average Treatment Effect

$$
\begin{aligned}
ATE &= E[Y_i(1) - Y_i(0)] \\
&= E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 0]
\end{aligned}
$$

When the Treatment is not Randomized

$$E[Y_i|W_i = 1] - E[Y_i|W_i = 0] = \underbrace{E|Y_i(1)|W_i = 1] - E(Y_i(0)|W_i = 0]}_{\text{Average Treatment on the Treated}}$$
$$+ \underbrace{E[Y_i(0)|W_i = 1] - E[Y_i(0)|W_i = 0]}_{\text{Selection Bias}}$$

More Stringent Assumptions

For some observed covariate X

Positivity: assignment is probabilistic: $0 < P(W_i = 1 \mid X, Y(0), Y(1)) < 1$

No unmeasured confounding: $P(W_i = 1 \mid X, Y(0), Y(1)) = P(W_i = 1 \mid X)$

What is Confouding?

Confounding is the bias caused by common causes of the treatment and outcome

In observational studies, the goal is to avoid confounding inherent in the data.

No unmeasured confounding assumes that we've measured all sources of confounding

Do you have examples of confouding?

What is Confouding?

Effect of income on voting (confounding: age)

Effect of job training program on employment (confounding: motivation)

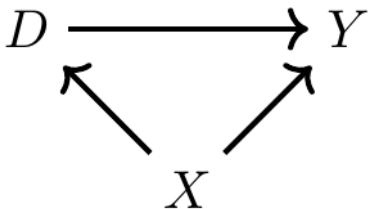
Effect of obesity on heart disease (confounding:age, diet, smoking...)

Big Problem

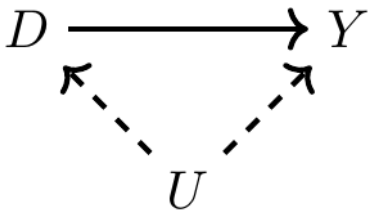
How can we determine if no unmeasured confounding holds if we didn't assign the treatment?

Which covariate should we adjust for?

Using Theory to Think About Confounding



Using Theory to Think About Confounding



No unmeasured confounders is not testable

Are *unobserved* $(Y_i(0)|W_i = 1, X_i)$ similar to *observed* $(Y_i(0)|W_i = 0, X_i)$?

A Classic Strategy

Multivariate regression model:

$$Y_i = \alpha + \beta W_i + \mathbf{X}\gamma + \epsilon_i$$

Adjust for all measured confounders

Assume that there are no omitted variables

An Example

Imagine a researcher is interested in estimating the effect of education on wages.

She gather data 526 workers and observes:

- `educ` (years of education)
- `exper` (years of labor market experience)
- `tenure` (years with the current employer)
- `log(wage)` (the log of the wage)

An Example

$$\widehat{\log(wage)} = .284 + .092\,educ + .0041\,exper + .022\,tenure$$
$$(.104)\quad (.007)\qquad (.0017)\qquad (.003)$$
$$n = 526,\ R^2 = .316,$$

Interpretation: holding `exper` and `tenure` fixed, another year of education is predicted to increase log(wage) by .092, which translates into an approximate 9.2% increase in wage.

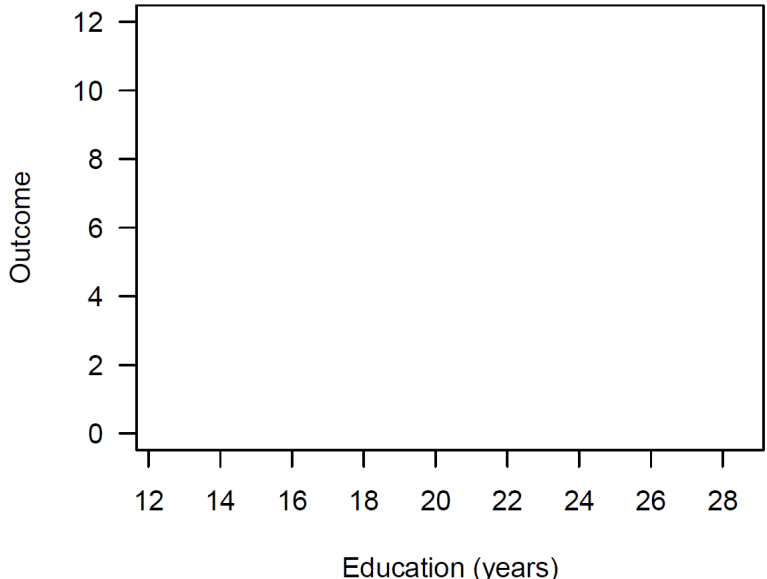
Uncertainty: the data are compatible with an increase in wage ranging from $(0.092 - 1.96 \times 0.007) \times 100 = 7.8\%$ to $(0.092 + 1.96 \times 0.007) \times 100 = 10.6\%$.

Issues with this Approach

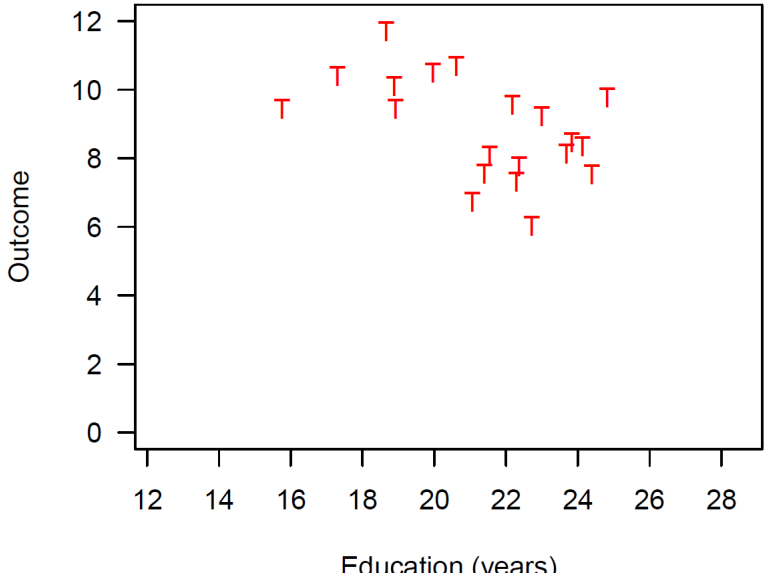
If we assume that all measured confounders are observed, the estimate could be still biased because:

- We assume a linear & additive model
- We did not check visually if we were comparing apples to apples
- The model could extrapolate to regions without empirical counterfactuals

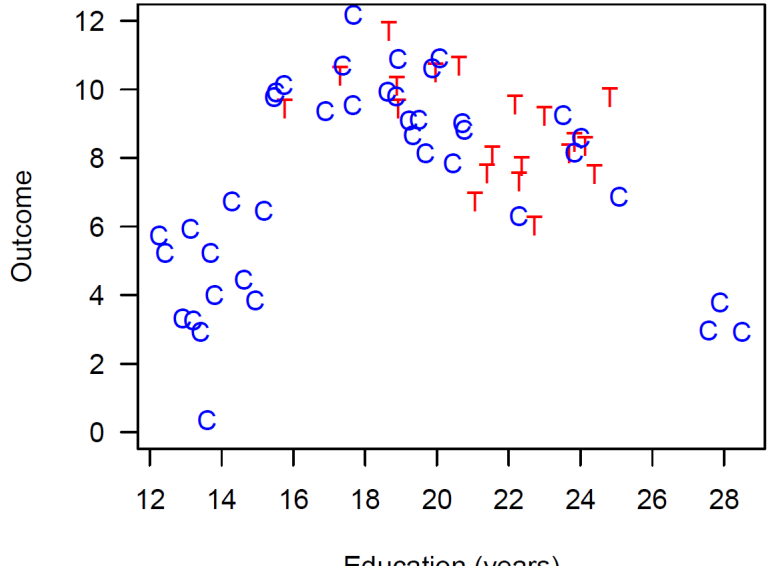
Extrapolation Bias



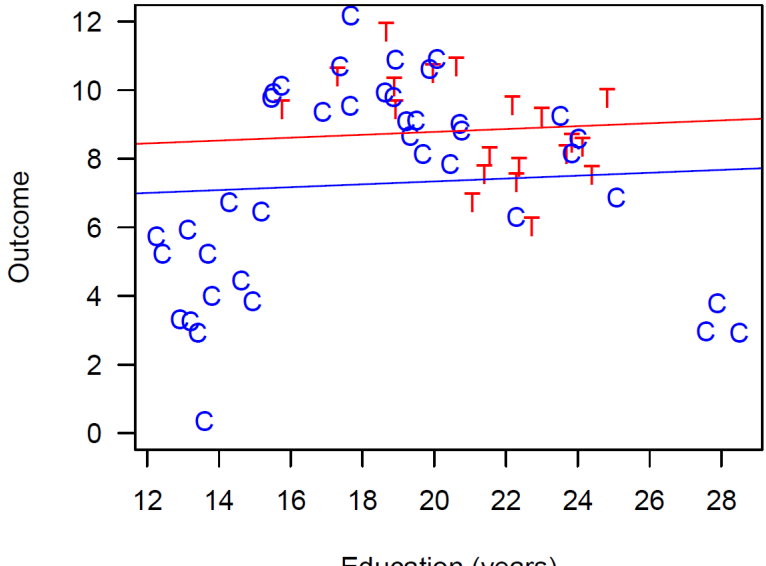
Extrapolation Bias



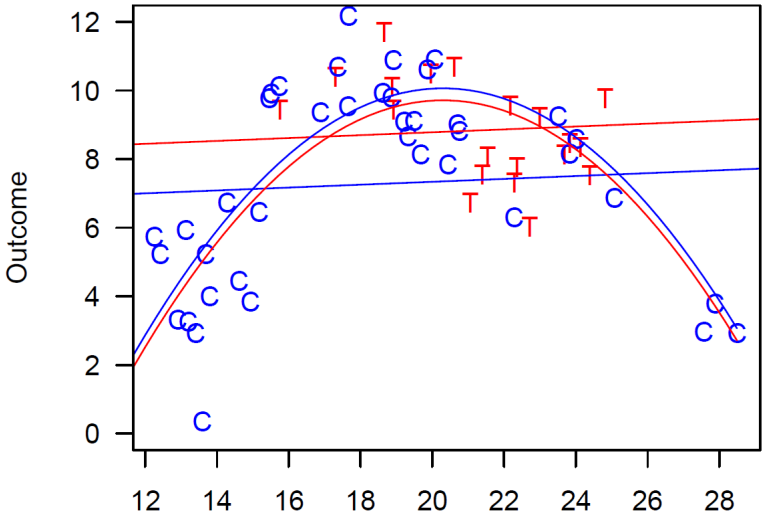
Extrapolation Bias



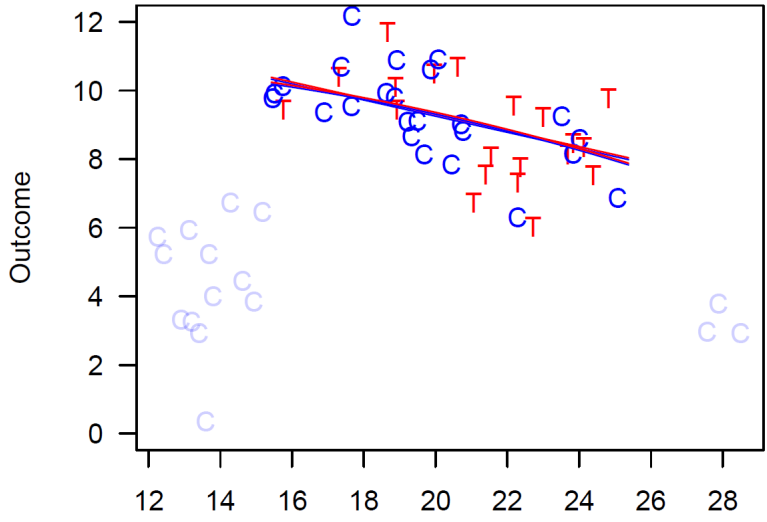
Extrapolation Bias



Extrapolation Bias



Extrapolation Bias



Solution: Matching

Recall the fundamental problem of causal inference:

$$Y_i = W_i \times Y_i(1) + (1 - W_i)Y_i(0)$$

$$Y_i = \begin{cases} Y_i(1) & \text{if } W_i = 1 \\ Y_i(0) & \text{if } W_i = 0 \end{cases}$$

It is as a *missing* data problem

We can try to *impute* the missing outcomes using matching

Matching

The causal estimand is now the average treatment on the treated (ATT) :

$$\tau_{ATT} = E[Y_i(1) - Y_i(0)|W_i = 1]$$

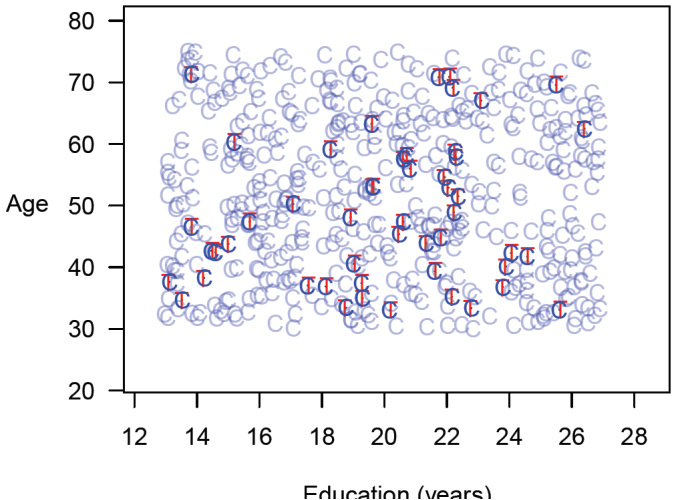
For each treated unit i: find the "closest" control unit j and impute j's outcome as the unobserved potential outcome for i

$$\widehat{\tau_{ATT}} = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i - Y_j(i))$$

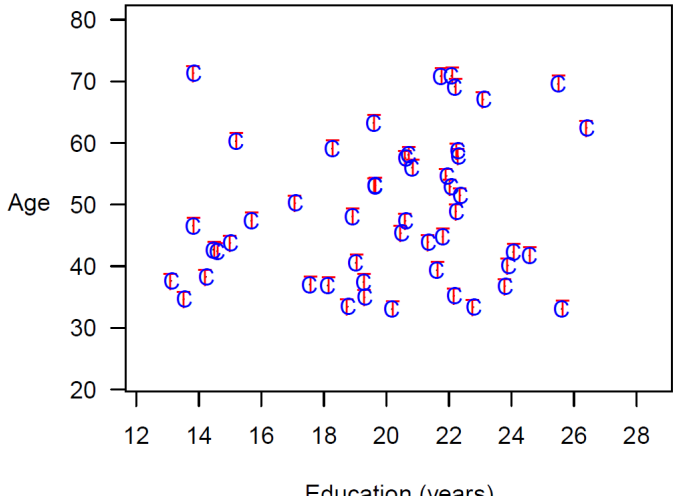
where $Y_j(i)$ is the observed outcome for (untreated) unit j, the closest match to i, i.e. $X_j(i)$ is closest to X_i among the untreated observations

Which distance metric should we use?

Exact Matching



Exact Matching



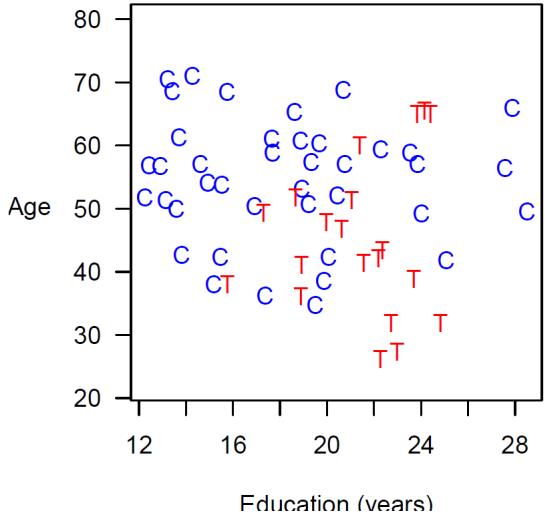
Propensity Score Matching

The curse of dimensionality makes exact matching not feasible in practise

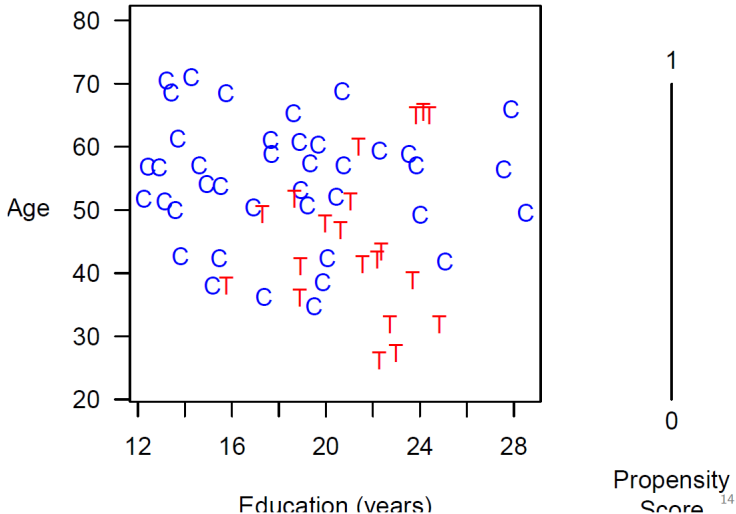
Solution: for each unit, predict its probability of being treated using observed covariate

Match on the *propensity score* which summarises covariate information in a single dimension.

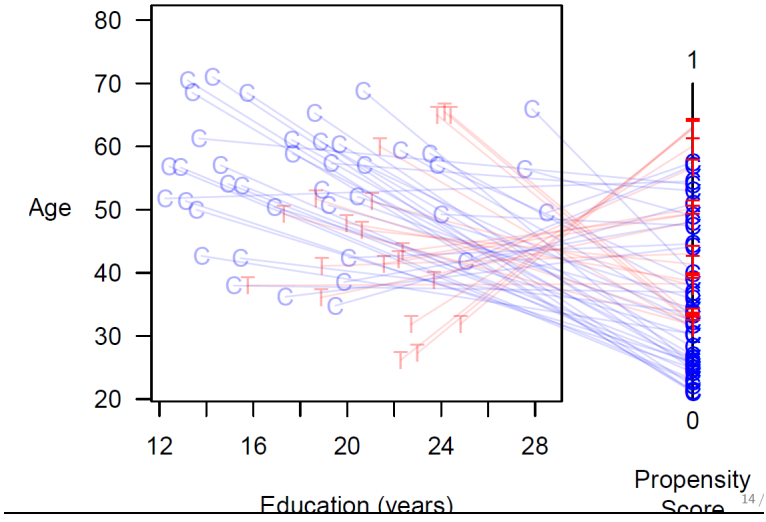
Propensity Score Matching



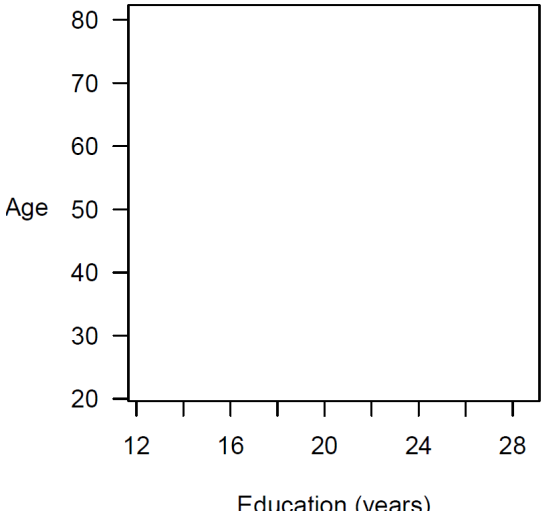
Propensity Score Matching

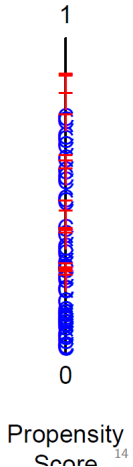


Propensity Score Matching



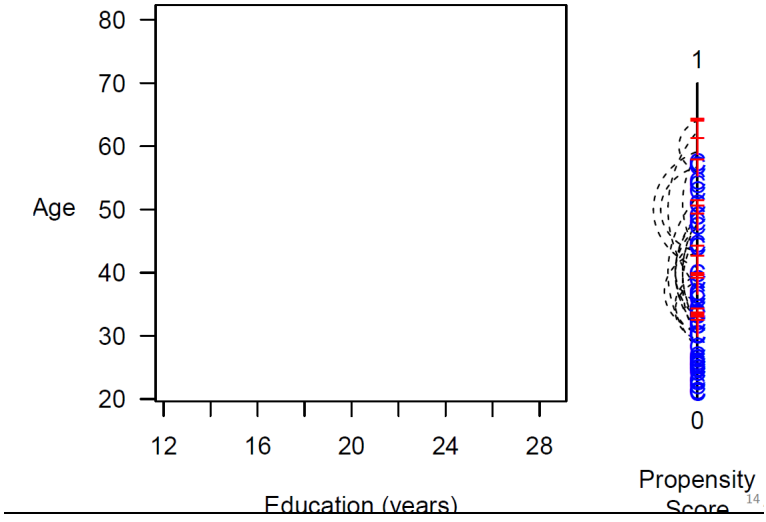
Propensity Score Matching



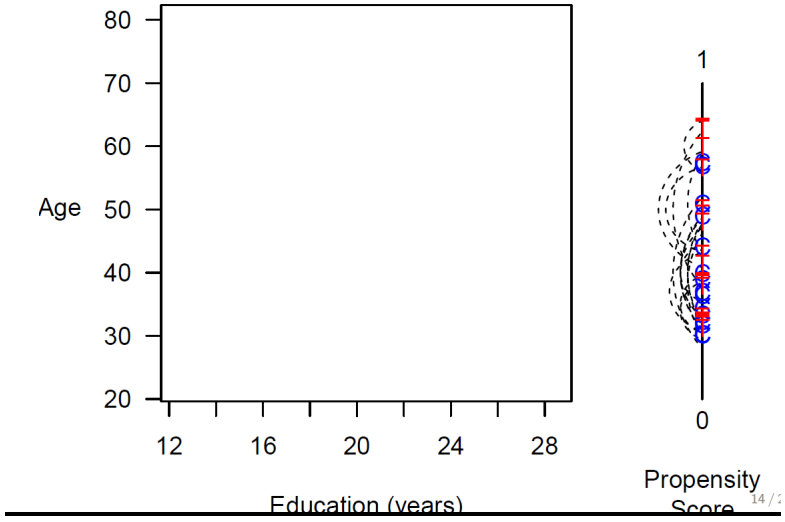


14

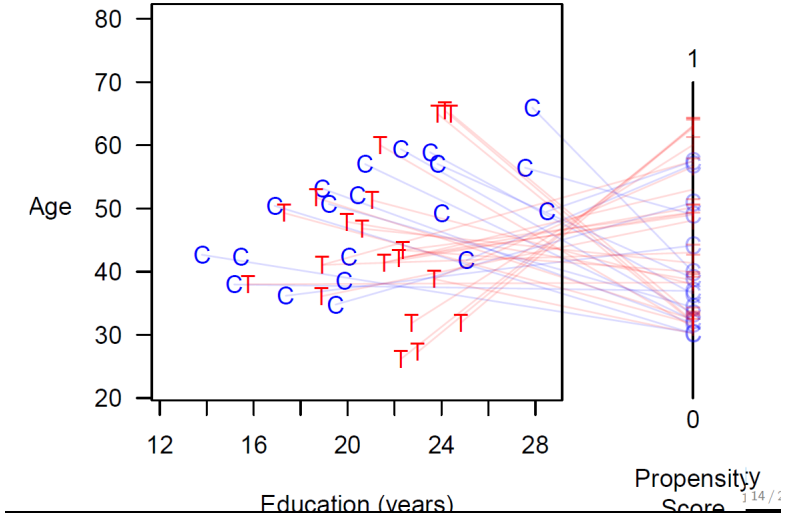
Propensity Score Matching



Propensity Score Matching



Propensity Score Matching



Example: Air Pollution in Milan

Michela Baccini *et al.* (EH, 2017):

Data: 1,461 days (2003-2006)

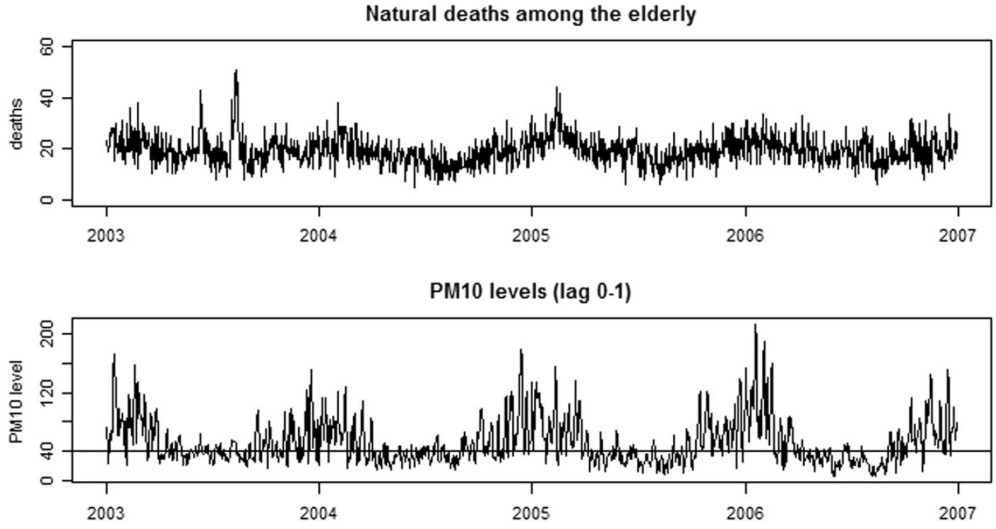
Control days: $PM_{10} < 40\ \mu g/m^3$

Treated days: $PM_{10} \geq 40\ \mu g/m^3$

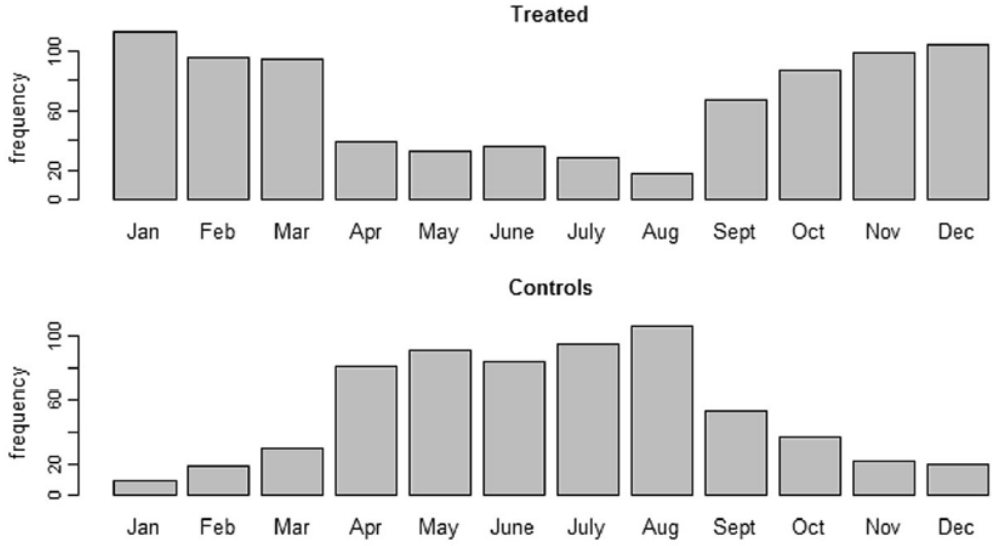
Outcome: Daily mortality

Causal Estimand: AD = $Y_i(1) - Y_i(0)$

Confounding



Confounding



Propensity Score Matching

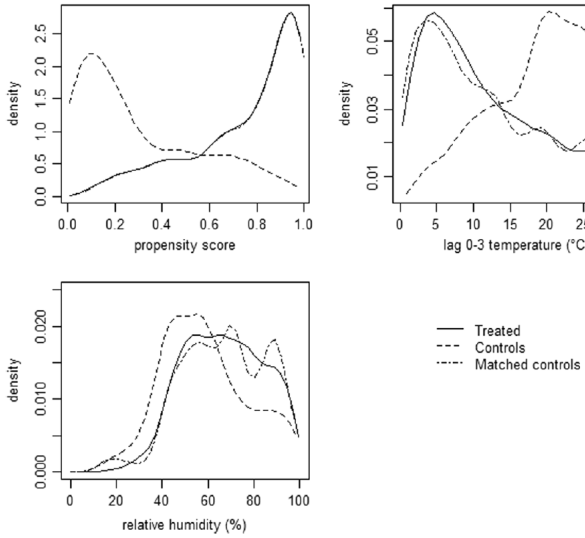
Goal: for each treated unit, impute its missing Y(0).

PS: $W_i \sim Bernoulli(e_i)$ $\text{logit}(e_i) = f(Z_i, \beta)$

Z: temperature, humidity, weekday, day of the year, etc...

Nearest neighbor matching: each treated day i is matched to the control day with estimated propensity score closest to i

Confounding



Confounding

Table 1 Covariates balance before and after matching, Milan, Italy, 2003–2006

	Mean/Proportion			Standardized difference[d]		
Background characteristic	Treated ($n = 812$)	Controls ($n = 649$)	Matched Controls ($n = 649$)	Pre-matching	Post-matching	% Bias[e]
Estimated propensity score	0.756	0.306	0.756	1.810	0	100.0
Temperature (°C)[a]	11.4	18.3	11.3	0.914	0.013	98.5
Relative humidity (%)	66.8	58.6	67.1	0.456	0.014	97.0
Saturdays and Sunday	0.243	0.341	0.195	0.217	0.106	51.0
Day of year	-	-	-	405.5[c]	15.9[c]	96.1[f]
Influenza epidemics	0.128	0.009	0.054	0.483	0.315	34.7
Heat episodes[b]	0.032	0.028	0.025	0.001	0.002	−77.8
Summer days	0.225	0.664	0.252	0.037	0.002	93.8

Confounding

Table 2 Estimated number of attributable deaths by cause and age class, Milan, Italy, 2003–2006

	Age 15–64		Age 65–74		Age 75+		All ages (15+)	
	AD	90% CI	AD	90% CI	AD	90% CI	AD	90% CI
Cardiovascular causes	−172	−368, 24	91	−244, 426	797	305, 1288	716	117, 1315
Respiratory causes	−25	−133, 83	87	11, 163	243	−22, 508	305	17, 593
Other natural causes	153	−246, 552	−157	−401, 87	62	−414, 538	58	−496, 612
All natural causes	−44	−609, 521	21	−425, 467	1102	388, 1816	1079	116, 2042

AD attributable deaths, *90% CI* 90% confidence interval

Matching

Adjusts non-parametrically for observed confounders

Reveals the common support of the data

Helps create a balanced sample (*i.e.*, comparing apples to apples)

Therefore reduces model dependence

It a principled approach to analyze the data under the assumption of no unmeasured confounders

Evaluating Reduction in Model Dependence

Carpenter, AJPS (2002)

Hypothesis: Democratic senate majorities slow FDA drug approval time

Data: $N = 408$ new drugs (262 approved, 146 pending)

Measured confounders: 18 (clinical factors, firm characteristics, media variables, etc.)

Model: lognormal survival

Question: Causal effect of Democratic Senate majority (identified by Carpenter as not robust)

Match: prune 49 units (2 treated, 17 control units)

Run: 262,143 possible specifications; calculate the estimate for each

Evaluate: Variability in estimates across specifications

Evaluating Reduction in Model Dependence