ECON42720 Causal Inference and Policy Evaluation

4a Matching and Re-weighting

Ben Elsner (UCD)

About this Lecture

This lecture is all **about adjusting for confounders**

- Why we want to adjust for confounders
- How we can adjust for confounders using re-weighting
- Limits of re-weighting: the curse of dimensionality
- How we find suitable control units using matching
- Differences between regression and matching

Matching is a **powerful tool**, but it's also an art in itself

- There are many techniques out there
- Learning to use them takes practice

Resources

As an **introduction**, I recommend Chapter 5 in Scott Cunningham's Mixtape

Slightly **more detailed coverage** can be found in

- Huntington-Klein's The Effect, Chapter 14
- Huber's Causal Analysis, Chapter 4

Many examples in this chapter, in particular the R codes, have been taken from The Effect or inspired by it.

Credits

Stephen Pettigrew produced some very instructive graphs on matching. You can find his slides on matching here. He has lots of interesting materials on causal inference on his website.

Gary King has done fundamental work on matching and has a website with lots of resources. I have used some of his materials, especially the illustrations of matching, in this lecture. One paper I learned a lot from is Ho *et al.* (2007).

Starting Point: Conditional Independence

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

For **causal identification**, we require the assumption that the **treatment** D is as good as **randomly assigned conditional on the covariates** X

Formally, this means that the potential outcomes are **conditionally independent** of the treatment assignment given the covariates

$$E\left[Y^1 \mid D = 1, X\right] = E\left[Y^1 \mid D = 0, X\right]$$
$$E\left[Y^0 \mid D = 1, X\right] = E\left[Y^0 \mid D = 0, X\right]$$

Conditional Independence and Selection on Observables

If CIA holds, we speak of **selection on observables**

- **Independence does not hold** in general
- But it holds in the **subpopulations** defined by the covariates X

The **groups defined by X** (think age, gender, neighbourhood, etc) determine the **treatment assignment**

- But **within each group**, who gets treated is **as good as random**

This is a **strong assumption!**

Example: Smoking and Lung Cancer

Does smoking cause lung cancer?

- Today we would say "yes, of course"
- But answering this question was far from clear in the 1950s
- There is a **strong correlation** between smoking and lung cancer, but is it causal?

(Potential) problem: confounders

- There could be genetic determinants of smoking and lung cancer
- There could be environmental factors that cause both smoking and lung cancer

We don't have **experimental evidence**

Example: Death Rates per 1,000

The following example from Cochran (1968) will illustrate what **selection on observables** and do for us

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

In all countries, the **highest death rates are for cigar and pipe smokers**

- Does this mean that smoking pipes and cigars is more dangerous than smoking cigarettes?
- And given the minor differences between cigarette smokers and non-smokers, are cigarettes harmless?

Smoking and Lung Cancer: Independence?

The **independence assumption** would imply that all three groups have the **same potential outcomes on average**

$$E[Y^1 \mid \text{Non-Smoker}] = E[Y^1 \mid \text{Cigarette}] = E[Y^1 \mid \text{Pipe}] = E[Y^1 \mid \text{Cigar}]$$
$$E[Y^0 \mid \text{Non-Smoker}] = E[Y^0 \mid \text{Cigarette}] = E[Y^0 \mid \text{Pipe}] = E[Y^0 \mid \text{Cigar}]$$

Suppose that the **independence assumption** holds

- This would/should also mean that observable characteristics X are similar between the groups
- I.e. the **covariates should be balanced** between groups

Are cigarette smokers similar to pipe and cigar smokers?

Let's ask Dall-E: show me a picture of a cigarette smoker and a cigar smoker

Age as a Confounder?

Smoking group	Canada	UK	US
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Clearly, **age affects what people smoke and also their death rates**

- Independence is violated: the **distribution of age** is different between the groups
- There may be other confounders, but let's focus on age for now

We have **covariate imbalance!**

Potential remedy: condition on age (**subclassification**)

Subclassification: Divide Age into Strata

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

Subclassification: Divide Age into Strata

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

The **death rate of cigarette smokers in the population** is:

$$20 \times \frac{65}{100} + 40 \times \frac{25}{100} + 60 \times \frac{10}{100} = 29$$

But: the **age distribution is (heavily) imbalanced** between the groups

Re-weighting: Age-Adjusted Death Rates

Let's **re-weight** the death rates of cigarette smokers by the **age distribution of pipe/cigar smokers**

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

The **age-adjusted death rate of cigarette smokers** is:

$$20 \times \frac{10}{100} + 40 \times \frac{25}{100} + 60 \times \frac{65}{100} = 51$$

If **cigarette smokers** had the **same age distribution as pipe/cigar smokers**, their death rate would be 51

Age-Adjusted Death Rates

Cochran **computes age-adjusted death rates** (based on the population age distribution)

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	29.5	14.8	21.2
Cigars/pipes	19.8	11.0	13.7

Here we **achieved balance on one covariate: age**

- The **age-adjusted death rates** are now more similar between the groups
- But there may be an **imbalance on other covariates** (SES, income, health, etc)

We need to **use a DAG** to identify **all confounders** and adjust for them

Identifying Assumptions

In presence of confounders X, we can **identify a causal effect under two assumptions**

1. **Conditional Independence**: $Y^0, Y^1 \perp D \mid X$
2. **Common Support**: $0 < P(D = 1 \mid X) < 1$ with probability one

Common support: for each stratum, we need some units that are treated and others that are control units

- We need **common support** to calculate the **weights for the adjustment**

Summary: Subclassification and Re-weighting

Treated and control units often differ in the **distribution of X (confounders)**

We can make **both groups** (somewhat) **comparable** by

1. dividing the sample into **strata based on** X (**subclassification**)
2. re-weighting the strata to **achieve balance on** X (**re-weighting**)

After re-weighting, both groups have the **same distribution of** X by construction

Causal Identification with Selection on Observables

Under **conditional independence and common support**, the following holds:

$$
\begin{aligned}
E\big[Y^1 - Y^0 \mid X\big] &= E\big[Y^1 - Y^0 \mid X, D = 1\big] \\
&= E\big[Y^1 \mid X, D = 1\big] - E\big[Y^0 \mid X, D = 0\big] \\
&= E\big[Y \mid X, D = 1\big] - E\big[Y \mid X, D = 0\big]
\end{aligned}
$$

The **estimator for the ATE** is as follows:

$$
\widehat{\delta_{ATE}} = \int \Big(E\big[Y \mid X, D = 1\big] - E\big[Y \mid X, D = 0\big]\Big) d\,\Pr(X)
$$

The Limits of Subclassification: The Curse of Dimensionality

In the example of **smoking and death rates**, we **adjusted for just one confounder**

- The hope was that, by slicing up age into three groups, achieve balance in treated and control groups
- We did achieve balance on age, but what about other confounders?
- Also, are three age groups enough or do we need more?

In practice, we have the **problem of a finite sample size**

- There are **limits to how many strata we can create**
- We cannot have an infinite number of groups defined by one variable (such as age)
- We cannot have an infinite number of variables to adjust for

This problem is known as the **curse of dimensionality**

The Limits of Subclassification: The Curse of Dimensionality

Let's say we have $k = 1, \dots, K$ groups (for example defined by gender and age). We can calculate the ATT as

$$\widehat{\delta}_{ATT} = \sum_{k=1}^{K} \left(\overline{Y}^{1,k} - \overline{Y}^{0,k} \right) \times \left(\frac{N_T^k}{N_T} \right)$$

where $\overline{Y}^{1,k}$ and $\overline{Y}^{0,k}$ are the average outcomes in group k for treated and control units, and N_T^k is the number of treated units in group k.

In **large groups** (small K) we will easily find a **control unit for every treated unit**

As K increases and **groups get smaller**, we will have **more and more groups** that only contain **control or treated units but not both**

Possible Solution: Matching

Source: Dall-E

Possible Solution: Matching

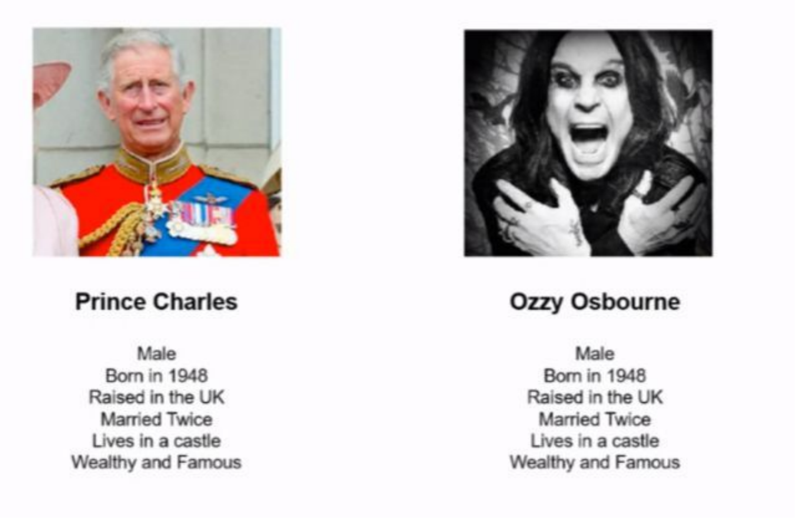
Idea of matching:

- for each **treated unit**, find a **control unit** that is **similar on all confounders**
- **compare the outcomes** of **treated and control units**
- The **comparison** gives us an **estimate of the ATT**

Control units: **statistical twins** of treated units

It if also possible to have **multiple control units for each treated unit**

Statistical Twins?



Source: somewhere on X, before 2023

Why Don't We just Run a Regression

If treated and untreated units have different X and X are confounders, we can include them in a regression

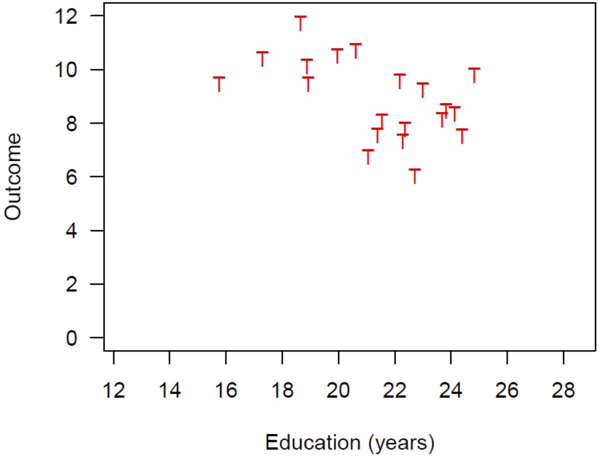
$$Y_i = \alpha + \beta D_i + \beta \boldsymbol{X_i} + u_i$$

Don't we then **compare like with like**?

- Answer: it depends on the **functional form** of the relationship between X and Y
- Regression can get it wrong if the relationship is non-linear and/or
- If there is **not much common support** in the distribution of X between treated and control units

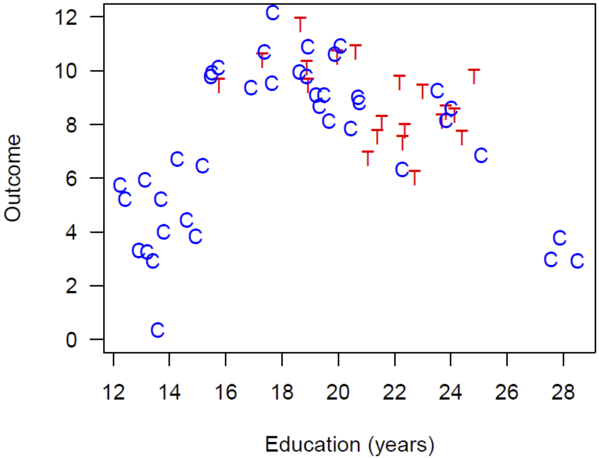
Regression vs. Matching

Suppose we want to look at the effect of a treatment D on an outcome Y. Education is a confounder.



Regression vs. Matching

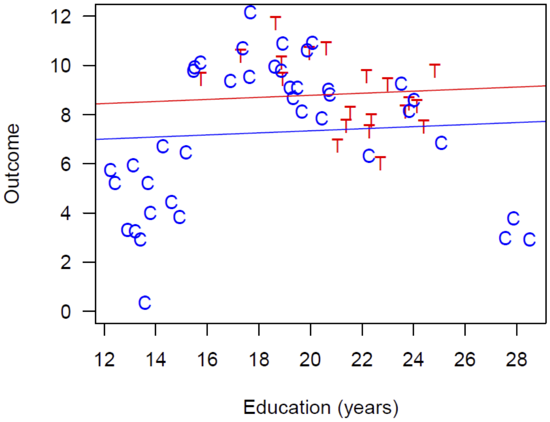
Enter the control units; for high and low levels of education, we have no common support



Regression vs. Matching

Separate regression lines for treated and control groups:

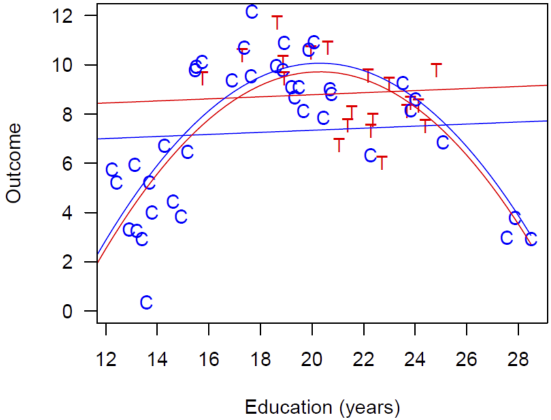
- the difference is $\widehat{\beta} > 0$



Regression vs. Matching

If we use a **quadratic term for education**, we get a different result

- The estimate $\widehat{\beta}$ is small and negative



Regression vs. Matching

The previous slides highlight a **problem with regression**

- with a **lack of common support**, control and treated units are not comparable
- this can even be the problem if both groups have the same average level of education

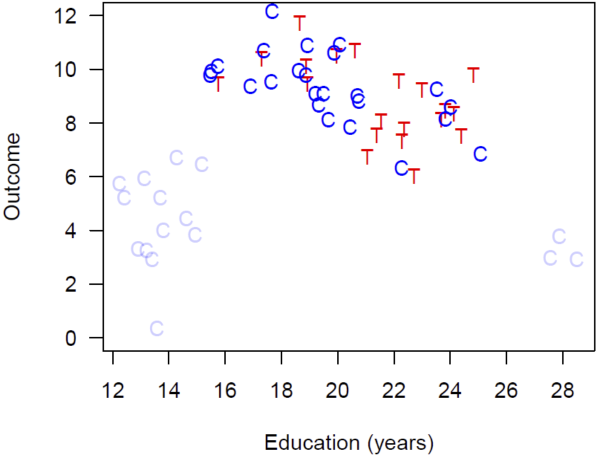
Control units with high and low levels of education influence the regression line

- but these units cannot be compared to any treated units
- so our regression compares fundamentally different units (apples and oranges)

We have a **covariate imbalance**; regression does not (always) solve the problem

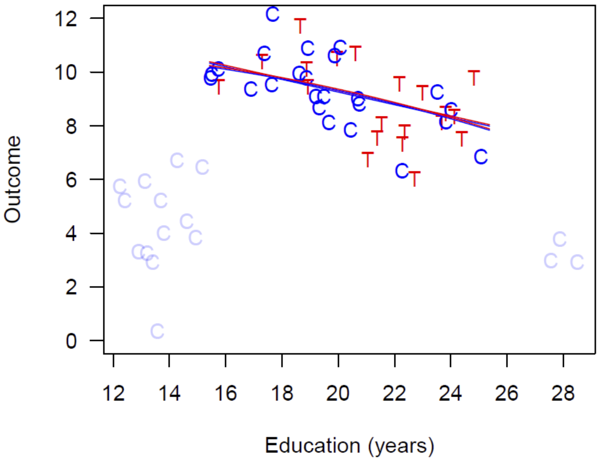
Regression vs. Matching

Matching **selects units with common support** in the distribution of X



Regression vs. Matching

Among these units, there is **no difference between treatment and outcome**



Matching Stage 1: Preparation

1: Choose the **variables you want to match** on

- Match on **confounders**, but not on colliders or mediators

2: Choose a **matching method** (more on this later)

- The method determines how you select control observations

3: **Match treated and control observations**

- Select control observations that are similar in X to treated ones
- prune observations without good matches

Matching: Stage 2: Refinement and Estimation

4: Check if your **dataset is balanced on covariates**

- Treated and control observations should have similar values of X
- If you don't have balance, go back to stage 1

5: Run a **simple regression of the outcome on the treatment**

- Or do a simple difference in outcomes and run a t-test

6: **Run sensitivity checks** to see if the results depend on the matching procedure

- Change matching methods
- Change parameters of the matching method

Matching and the ATT: One Control Unit per Treated Unit

With one control unit for each treated unit, we **can calculate the ATT** as

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

- Y_i is the outcome for treated unit i
- $Y_{j(i)}$ is the outcome for the control unit $j(i)$

Matching and the ATT: Multiple Control Units per Treated Unit

Or if we find M matches for each treated unit, we can calculate the ATT as

$$\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(1)} \right] \right)$$

- $Y_{j_m(1)}$ is the outcome for the mth control unit matched to treated unit i

Matching and the ATE

We can also use **matching to estimate the ATE**. For this, we need to

- Find a similar control unit for each treated unit
- Find a similar treated unit for each control unit

The **estimator for the ATE** is as follows:

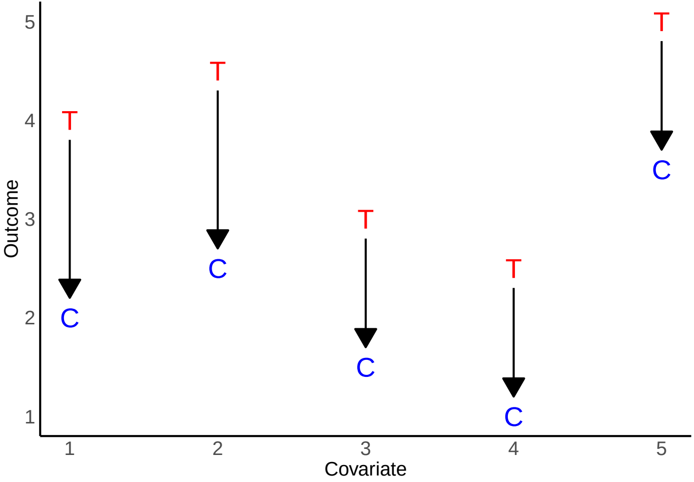
$$\widehat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left[Y_i - \left(\frac{1}{M}\sum_{m=1}^{M} Y_{j_m(i)}\right)\right]$$

Exact Matching

Match each treated unit to a control unit that has **exactly the same covariate values**

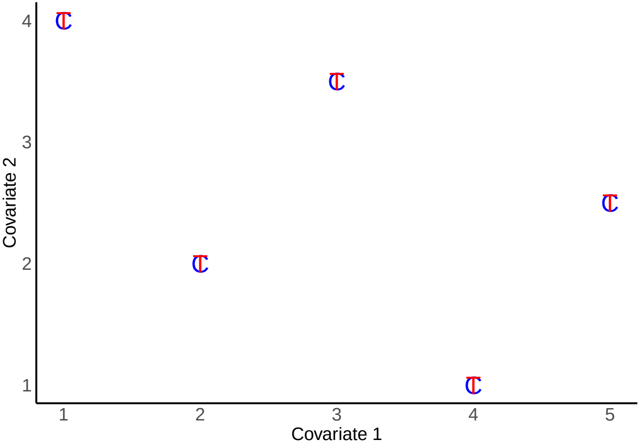
This is called **exact matching** and can be thought of as the **gold standard for matching**

Exact Matching with One Covariate



For each treated unit, we find a **control unit with the same covariate value**

Exact Matching with Two Covariates

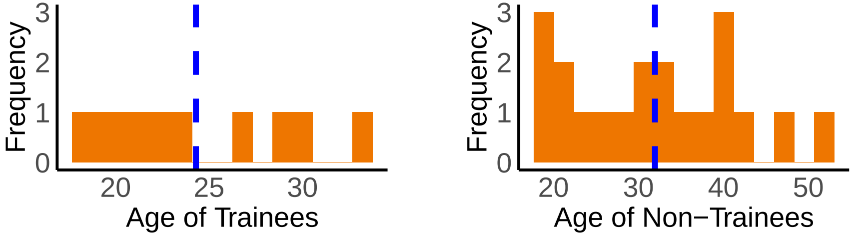


For each treated unit, we find a control unit with the **same values of covariates 1 and 2**

Example: Job Training Programme

Trainees			Non-Trainees		
Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500
2	29	12250	2	27	10075
3	24	11000	3	21	8725
4	27	11750	4	39	12775
5	33	13250	5	38	12550
6	22	10500	6	29	10525
7	19	9750	7	39	12775
8	20	10000	8	33	11425
9	21	10250	9	24	9400
10	30	12500	10	30	10750
			11	33	11425
			12	36	12100
			13	22	8950
			14	18	8050
			15	43	13675
			16	39	12775
			17	19	8275
			18	30	9000
			19	51	15475
			20	48	14800
Mean	24.3	$11,075	Mean	31.95	$11,101.25

Age Distribution of Trainees vs. Non-Trainees

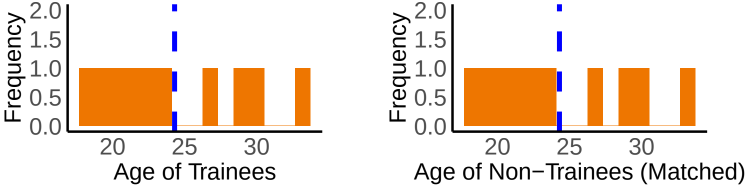


Clearly, the age distribution of trainees and non-trainees is different (mean 24.3 vs. 31.95)

Creating an (exactly) Matched Sample

Trainees			Non-Trainees			Matched Sample		
Unit	Age	Earnings	Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500	14	18	8050
2	29	12250	2	27	10075	6	29	10525
3	24	11000	3	21	8725	9	24	9400
4	27	11750	4	39	12775	8	27	10075
5	33	13250	5	38	12550	11	33	11425
6	22	10500	6	29	10525	13	22	8950
7	19	9750	7	39	12775	17	19	8275
8	20	10000	8	33	11425	1	20	8500
9	21	10250	9	24	9400	3	21	8725
10	30	12500	10	30	10750	10,18	30	9875
			11	33	11425			
			12	36	12100			
			13	22	8950			
			14	18	8050			
			15	43	13675			
			16	39	12775			
			17	19	8275			
			18	30	9000			
			19	51	15475			
			20	48	14800			
Mean	24.3	$11,075	Mean	31.95	$11,101.25	Mean	24.3	$9,380

Treated Sample vs. Matched Control Sample



With **exact matching**, the age distribution of **treated and matched control units are the same**

If age is the only confounder, we can **estimate the ATT** as

$$\text{ATT} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - Y_{i'}) = 11,075 - 9,380 = 1,695$$

So the **estimated causal effect** of the **training programme** is 1,695 dollars

References

Cochran, W. G. 1968. The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics*, **24**(2).

Ho, Daniel E., Imai, Kosuke, King, Gary, & Stuart, Elizabeth A. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, **15**(3), 199–236.

benjamin.elsner@ucd.ie

www.benjaminelsner.com

Sign up for office hours

YouTube Channel

@ben_elsner

LinkedIn

Contact

Prof. Benjamin Elsner
University College Dublin
School of Economics
Newman Building, Office G206
benjamin.elsner@ucd.ie

Office hours by appointment. Please email me.