MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 5: Selection on Observables 3

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2025

Topics of this lecture

1. Getting More Credible
2. Falsification Tests
3. Partial Identification
4. Sensitivity Analyses

Selection on Observables

Observational settings where the assignment mechanism for D is either unknown or not under our control.

We are willing to make the (1) conditional independence assumption:

$$(Y_1, Y_0) \perp\!\!\!\perp D \mid X$$

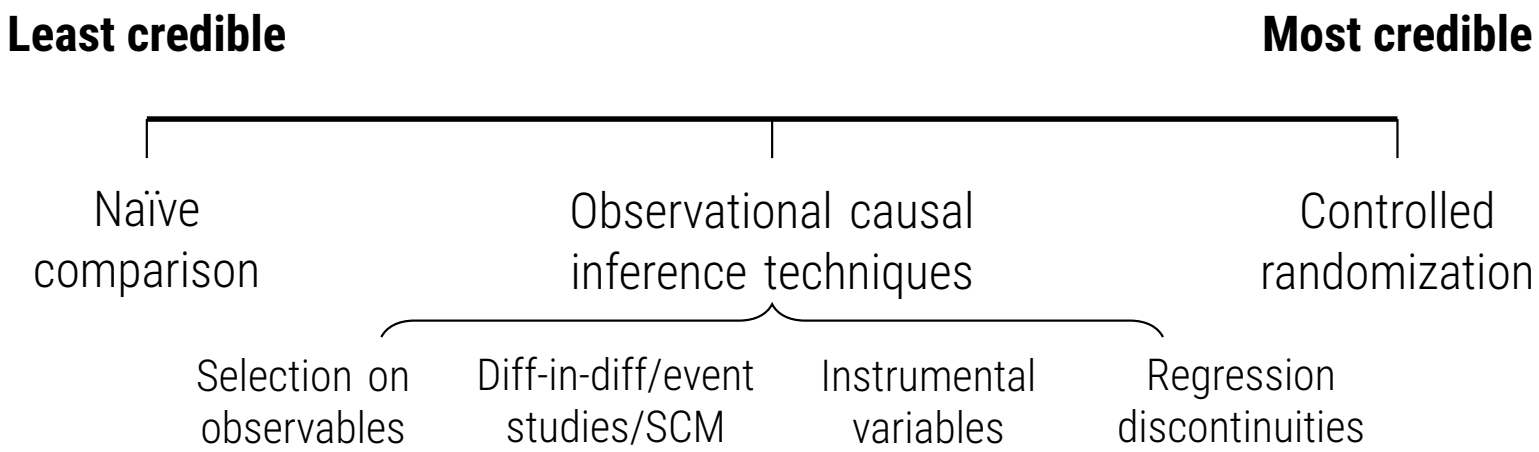
And the (2) common support assumption:

$$0 < P(D_i = 1 \mid X_i = x) < 1 \quad \text{for any} \quad x \in \mathcal{X}$$

This allows us to nonparametrically identify the ATE for D on Y Estimation via

subclassification, matching, weighting, or regression adjustment.

The Assumption-Credibility Trade-off



Recall that across designs, the stronger our assumptions are the less credible our inferences. Can we apply the general logic to a specific design?

We will explore approaches to improving our confidence in any given design:

1. Try to falsify our assumptions empirically
2. Weaken our assumptions and see what happens
3. Assess how wrong we would have to be to change our conclusions aka - sensitivity analysis

YES

Note: These approaches can (should?) also be used to assess other designs.

1 Getting More Credible

2 Falsification Tests

3 Partial Identification

4 Sensitivity Analyses

What is Falsification?

Falsification is a scientific principle (a 'criterion of demarcation') that comes from (at least) Popper (1934/1959):

> *The point is that, whenever we propose a solution to a problem, we ought to try as hard as we can to overthrow our solution, rather than defend it.*

Non incentive viable solution because we want to defend what we made

Contrast this with another term: 'validation' (a.k.a verification).

This is not somethign we can do because of the FPOCI

Validation implies a test that provides evidence in favour of our assumptions.

By contrast, falsification implies a test that, if failed weighs against our assumptions. These sometimes look equivalent but they are not!

"Covariates are balanced $\rightsquigarrow$ our assumptions are met."

vs.

"Covariates are balanced $\rightsquigarrow$ no evidence that our assumptions are not met."

This is better because its falsification and leaves the door open for potential thigns that we not control for but what we control for are as expected

Placebo Tests

One important type of falsification test is the placebo test.

Consider the following setup:

1. Treatment D, outcome Y
2. Assumption: $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i$ (or, alternatively, conditional independence)
3. Estimand is the ATE of D on Y

Suppose we remain concerned about the possible presence of a confounder U which throws into doubt our assumption. We design tests that, if the confounder is present, will falsify our assumption.

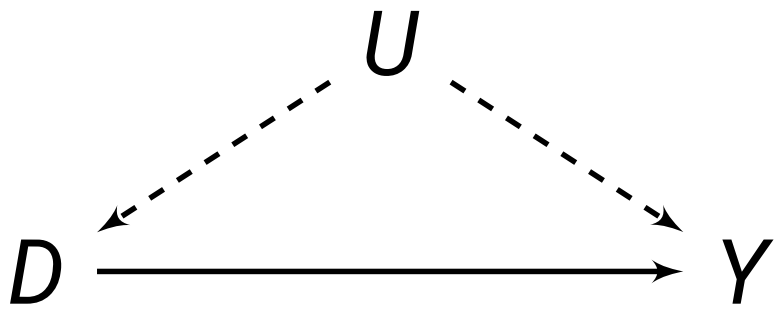
Eggers et al (2023) propose a typology of placebo tests that can be used to falsify assumptions about our design:

1. Placebo outcome test: $D \rightarrow Y'$ where $Y' \neq Y$
2. Placebo treatment test: $D' \rightarrow Y$ where $D' \neq D$
3. Placebo population test: $D \rightarrow Y$ for population where $\tau_{ATE} = 0$

Placebo Tests as Graphical Models: The Problem

Problem: We wish to estimate the effect of D on Y, but we remain concerned about the presence of a confounder U.

We can present this graphically:



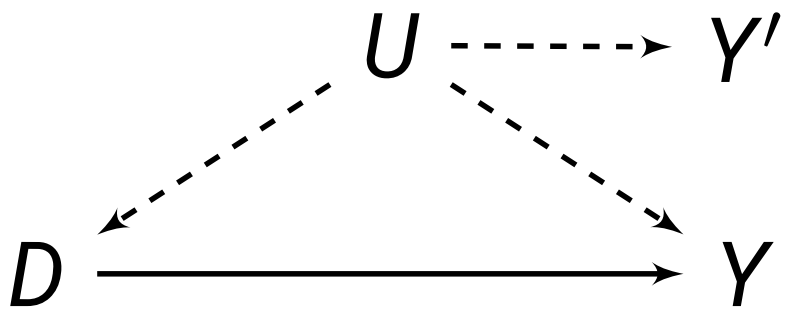
Reminder: U simultaneously sets D and Y, opening a back-door path that renders the effect of D on Y is not identified. How might we seek to falsify the claim that U is not a concern?

Placebo Tests as Graphical Models: Placebo Outcome Test

Test for an 'effect' of D on Y', under two assumptions:

1. The true effect of D on Y' is ≈ 0.
2. The confounder of interest affects Y'.

Graphically:



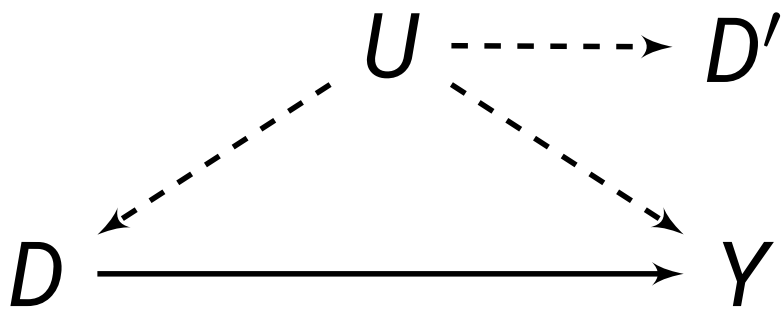
Insight: If U is present, then we should find a relationship between D and Y' (if our model is correct). Evidence of no relationship fails to falsify our design.

Placebo Tests as Graphical Models: Placebo Treatment Test

Test for an 'effect' of D' on Y, under two key assumptions:

1. The true effect of D' on Y is ≈ 0.
2. The confounder of interest affects D'.

Graphically:



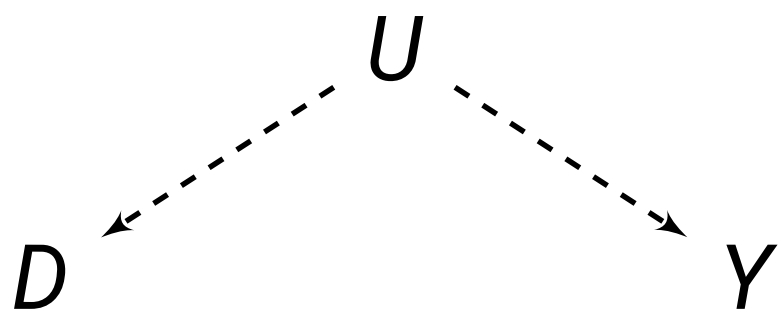
Insight: If U is present, then we should find a relationship between D' and Y (if our model is correct). Evidence of no relationship fails to falsify our design.

Placebo Tests as Graphical Models: Placebo Population Test

Test for an 'effect' of D on Y in a new population, under two key assumptions:

1. The true effect of D on Y is ≈ 0.
2. The confounder of interest still affects D and Y.

Graphically:



Insight: If U is present, then we should still find a relationship between D and Y (if our model is correct). Evidence of no relationship fails to falsify our design.

Are Falsification Tests Helpful?

The value of such tests (or 'robustness tests' in the specific case) is debated (Gelman, 2018, "Robustness checks are a joke"):

> Part of the problem is that robustness checks are typically done for purpose of confirming one's existing beliefs, and that's typically a bad game to be playing. More generally, the statistical properties of these methods are not well understood. Researchers typically have a deterministic attitude, identifying statistical significance with truth (as for example here).

But they can be redeemed by good scientific practice:

- Design tests that are meaningful and fair.
- Falsify first, not last.
- Do not null-hack or store failed tests in the file-drawer! (I will know!)
- Try pre-registering observational designs and falsification tests if feasible.
- Pay attention to point estimation and statistical significance (more when we do diff-in-diffs)

1 Getting More Credible

2 Falsification Tests

3 Partial Identification

4 Sensitivity Analyses

"Woah woah woah! There's still plenty of meat on that bone. Now you take this home, throw it into Markdown, add some conditional ignorability, then loosen up those assumptions a little. Baby, you've got a paper going."

Carl Weathers (1948 - 2024, Rest In Peace)

Partial Identification Setup

Consider again a simple research setting:

- Binary treatment: $D_i \in \{0, 1\}$
- Potential outcomes: Y_{di}
- Estimand is the ATE:

$$\tau_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}]$$

Previously we asked: What assumptions must I make to point identify the ATE from observed data?

Now we ask: How much can we learn about the ATE from observed data, without any assumptions? This goal is what is called partial identification.


decomposing
things
statisticians
microbes
imgflip.com

Decomposing the ATE

Let's begin by decomposing the ATE into its constituent parts:

$$
\begin{aligned}
\tau_{ATE} &= \mathbb{E}[Y_{1i} - Y_{0i}] \\
&= \mathbb{E}[Y_{1i} \mid D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]P(D_i = 0) \\
&\quad -\mathbb{E}[Y_{0i} \mid D_i = 1]P(D_i = 1) - \mathbb{E}[Y_{0i} \mid D_i = 0]P(D_i = 0) \\
&= \Big(\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]\Big)P(D_i = 1) \\
&\quad + \Big(\mathbb{E}[Y_{1i} \mid D_i = 0] - \mathbb{E}[Y_i \mid D_i = 0]\Big)P(D_i = 0)
\end{aligned}
$$

Quantities in green are observed in our data, quantities in red are unobserved.

Previously we have made assumptions to fill in the unobserved quantities.

But we don't have to make those assumptions! We can make whatever (defensible!) assumptions we want...

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$P(D_i = 0)$	0	$\mathbb{E}[Y_{0i}\|D_i = 0]$	?
$P(D_i = 1)$	1	?	$\mathbb{E}[Y_{1i}\|D_i = 1]$

An extreme option is to assume the 'best' and 'worst' possible outcomes.

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$P(D_i = 0)$	0	$\mathbb{E}[Y_{0i}\|D_i = 0]$	$\underline{Y}$
$P(D_i = 1)$	1	$\overline{Y}$	$\mathbb{E}[Y_{1i}\|D_i = 1]$

An extreme option is to assume the 'best' and 'worst' possible outcomes.

Consider the following assumptions:

1. If left untreated, treated units would have had 'best' (highest) possible outcome ($\overline{Y}$)
2. If treated, control units would have had 'worst' (lowest) possible outcome ($\underline{Y}$)

Substituting into our decomposition, this gives the sharp lower bound on τ:

$$\underline{\tau} = \left(\mathbb{E}[Y_i|D_i = 1] - \overline{Y}\right) P(D_i = 1) + \left(\underline{Y} - \mathbb{E}[Y_i|D_i = 0]\right) P(D_i = 0)$$

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$P(D_i = 0)$	0	$\mathbb{E}[Y_{0i}\|D_i = 0]$	$\overline{Y}$
$P(D_i = 1)$	1	$\underline{Y}$	$\mathbb{E}[Y_{1i}\|D_i = 1]$

An extreme option is to assume the 'best' and 'worst' possible outcomes.

Conversely:

1. If left untreated, treated units would have had 'worst' (lowest) possible outcome ($\underline{Y}$)
2. If treated, control units would have had 'best' (highest) possible outcome ($\overline{Y}$)

This gives the sharp upper bound on τ:

$$\overline{\tau} = (\mathbb{E}[Y_i|D_i = 1] - \underline{Y})\, P(D_i = 1) + \left(\overline{Y} - \mathbb{E}[Y_i|D_i = 0]\right) P(D_i = 0)$$

Note: These 'extreme case' sharp upper and lower bounds are assumption free given the observed data. The most precise bounds we can derive only by looking at the data are $[\underline{\tau}, \overline{\tau}]$.

Adding Assumptions

Our 'extreme case' bounds are often not very useful. Let's add an assumption and see how the bounds change.

The monotone treatment selection (MTS) assumption (Manski & Pepper, 2000):

$$\mathbb{E}[Y_{0i} \mid D_i = 0] \leq \mathbb{E}[Y_{0i} \mid D_i = 1]$$

$$\mathbb{E}[Y_{1i} \mid D_i = 0] \leq \mathbb{E}[Y_{1i} \mid D_i = 1]$$

Read: The expected values of the potential outcomes for units who are in treatment are always higher than for those in the control.

This implies a tighter sharp upper bound on τ:

$$\begin{aligned} \tau &\leq (\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0])\, P(D_i = 1) \\ &\quad + (\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_i|D_i = 0])\, P(D_i = 0) \\ \therefore \overline{\tau} &= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \end{aligned}$$

Example: Sentencing and Recidivism

A classic example considers how the type of sentence for juvenile offenders affect recidivism (Manski, 2007):

- $D_i = 1$ if sentence involves confinement in residential facilities; 0 if not
- $Y_{di} = 1$ if commits a crime again given sentence type d; 0 if not

Observed strata:

	Y	
D	0	1
0	.36	.53
1	.03	.08

Point estimate (random assignment):
$.08/(.03 + .08) - .53/(.36 + .53) = .13$

Nonparametric sharp bounds:

- Assumption free: $[-.53 - .03,\ .36 + .08] = [-.56,\ .44]$
- MTS: $[-.56,\ .13]$

Doing More With Bounds

MTS is one possible assumption that can be used to tighten our bounds. There are many more you could make. This can, however, be analytically challenging.

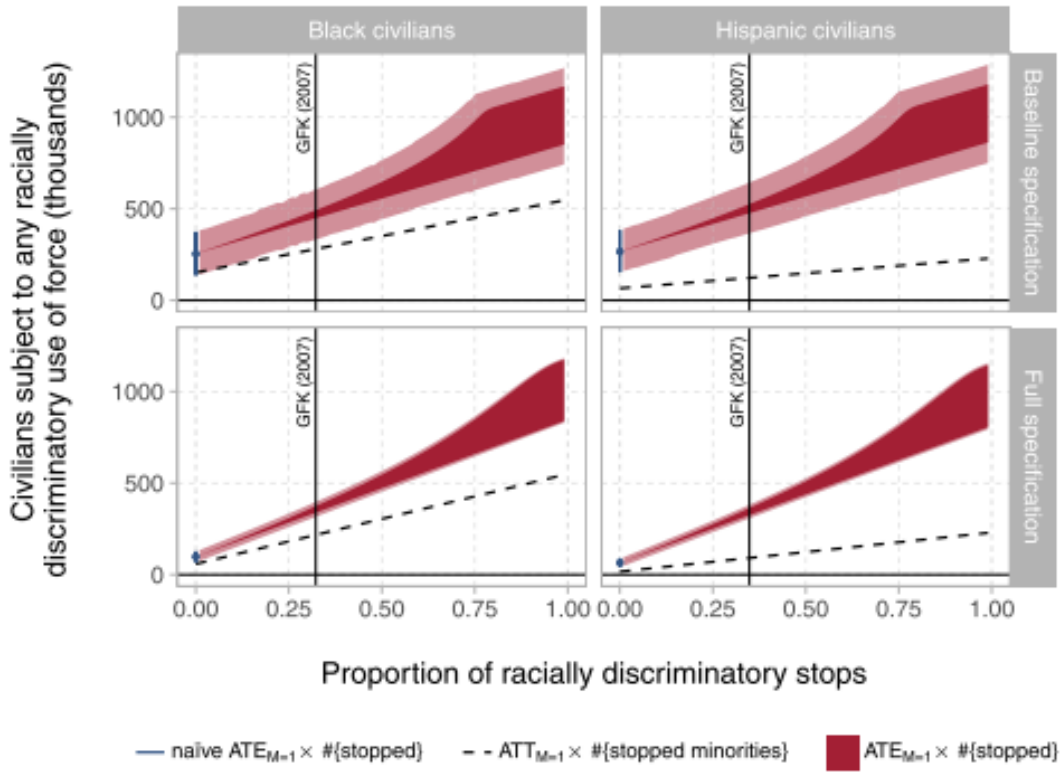
Duarte et al (2023) develop an algorithmic approach to deriving sharp bounds for discrete settings (packaged as `autobounds` in Python.)

'Thinking with bounds' can prove very useful:

- Fryer (2019) studies racial bias in US police violence with four datasets on police-denizen interactions.
- Key analyses: regressions of lethal and non-lethal police violence on the race of the denizen, controlling for many observable covariates. Key finding: some racial bias in non-lethal violence, but none in lethal violence.
- Knox et al (2020) study a key challenge in the paper: who is stopped is not random. If police exercise bias in who they stop, the above analysis is flawed. (Why?)
- They derive bounds on discrimination in Fryer's (2019) data based on varying assumptions about police bias in who they stop.

Fragility of Discrimination Findings to Selection

FIGURE 4. Bounds for Racially Discriminatory Use of Force, any Severity



Notes: These plots present the ATE$_{M=1}$ (ATT$_{M=1}$) for excess racial force, scaled by the number of stops (number of minority stops) to obtain the total number of civilians affected. The left panels consider the difference in the use of force if black civilians were substituted into each encounter of any race (each black encounter), versus white civilians; the right panels show the same quantities for Hispanic civilians. Blue points (error bars) denote the naïve estimator (95% confidence intervals), which, conditional on the typical selection-on-observables assumption, is unbiased for the ATE$_{M=1}$ if there are no discriminatory stops of minority civilians (zero on the *x*-axis). The dark (light) regions represent the range of possible values (95% CI) for (1) the ATE$_{M=1}$ and (2) the proportion of discriminatory stops in reported data jointly, per Proposition 1. The vertical line corresponds to an estimate of the proportion of discriminatory stops from Gelman, Fagan, and Kiss (2007), suggesting a plausible value for this unobservable parameter. The top (bottom) panels present bounds based on a model with no controls (the main specification, adjusting for a wide range of covariates).

(source: Knox et al (2020))

1 Getting More Credible

2 Falsification Tests

3 Partial Identification

4 Sensitivity Analyses

Confounding as Omitted Variable Bias

Consider a setting with a binary D, covariates $\mathbf{X}$, and candidate unobserved confounder U. We believe the following about the data generating process:

$$Y_i = \tau D_i + \mathbf{X}\beta + \delta U_i + \varepsilon_i$$

As U is unobserved, we estimate instead:

$$Y_i = \hat{\tau} D_i + \mathbf{X}\hat{\beta} + \hat{\eta}_i$$

What does $\hat{\tau}$ give us? (Cinelli & Hazlett, 2020)

$$\begin{aligned}
\hat{\tau} =& \frac{cov(D^{\perp\mathbf{X}}, Y^{\perp\mathbf{X}})}{var(D^{\perp\mathbf{X}})} \\
=& \frac{cov(D^{\perp\mathbf{X}}, \tau D^{\perp\mathbf{X}} + \gamma U^{\perp\mathbf{X}})}{var(D^{\perp\mathbf{X}})} \\
=& \tau + \delta \frac{cov(D^{\perp\mathbf{X}}, U^{\perp\mathbf{X}})}{var(D^{\perp\mathbf{X}})} \\
=& \tau + \delta\gamma
\end{aligned}$$

where $V^{\perp X}$ is residual V after removing components linearly explained by $\mathbf{X}$

Decomposing Confounding

Consider again our derived equality:

$$\hat{\tau} = \tau + \delta\gamma$$

Omitted variable bias is $\delta\gamma$. This is the interaction of two parameters:

1. δ: The marginal change in $\boldsymbol{Y}$ for different levels of $\boldsymbol{U}$ (may or may not be causal)
2. γ: The imbalance in $\boldsymbol{U}$ between treated and control (may or may not be causal)

Insights:

- Bias is multiplicative in δ and γ.
- As our classical confounding DAG tells us, if either $\delta = 0$ or $\gamma = 0$, then there is no confounding issue.
- While we cannot observe τ, δ, or γ, we can observe $\hat{\tau}$. We can thus explore what combinations of the three unobserved parameters could account for our estimate. This is the goal of sensitivity analysis.

Traditional Approach: Imbens (2003)

Imbens (2003) uses a slightly different setup, but gets us to a similar place, where δ and γ can be understood as:

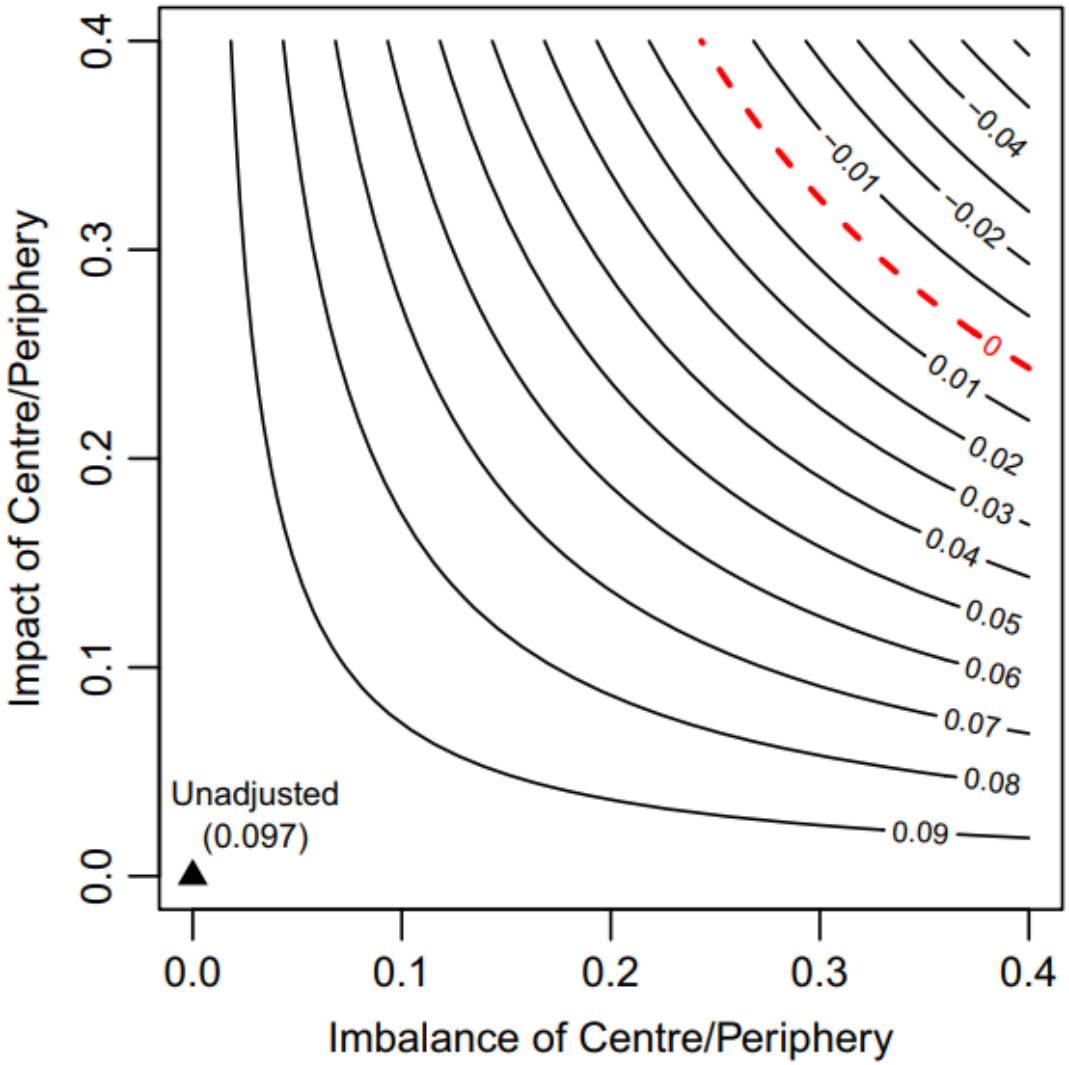
- $R^2_{Y,par}(\delta)$: residual variation in Y explained by unobserved confounder U
- $R^2_{D,par}(\gamma)$: residual variation in D explained by unobserved confounder U

Practically, as both values are bounded by $[0, 1]$, we can try any number of combinations of hypothetical values, and see how our estimate of τ changes.

These values are hypothetical, but a common strategy is to benchmark them against the explanatory power of observed covariates. This can be very powerful in settings where canonically important covariates are observed.

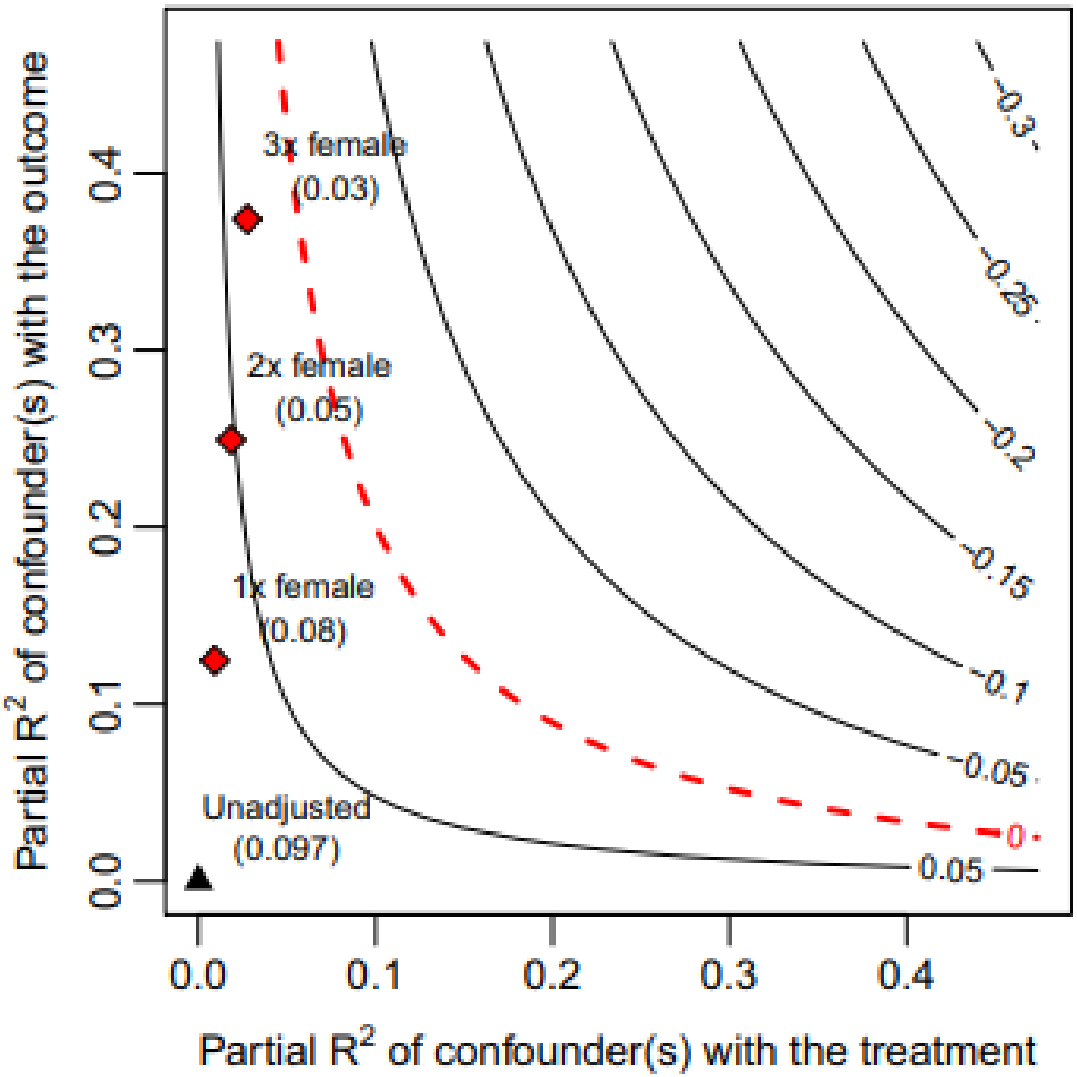
Visualisation typically through 'contour plots' – see `sensemakr` in `R`

Contour Plot for Single Hypothetical Confounder



(source: Cinelli & Hazlett, 2020)

Cinelli & Hazlett's (2020) Partial R^2 Parameterization



(source: Cinelli & Hazlett, 2020)