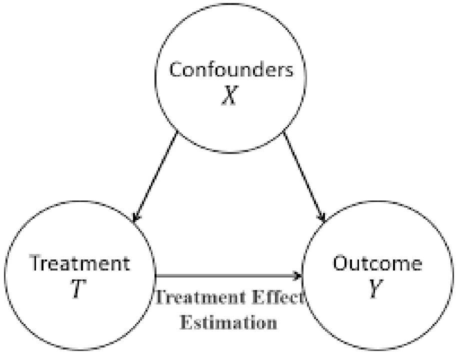


Behavioral Causal Inference

Ran Spiegler

TAU & UCL

EEA-ESEM, Barcelona

August 2023

Introduction

- Drawing causal inferences from correlational data:
  - Professional empirical researchers do this for a living.
  - Lay people perform this activity for everyday personal decisions.
    - Nutrition $\rightarrow$ Health
    - Education $\rightarrow$ Future income
    - Social distancing $\rightarrow$ Viral disease

Introduction

- The challenge: Confounding variables
  - Observed correlations do not represent causal effects.
- Professionals use various methods to cope with this problem.
- A basic method: Control variables
  - Professionals distinguish between "good" and "bad" controls (Angrist-Pischke 2009, Cinelli et al. 2022).

Introduction

- What about lay decision makers (DMs)? Two differences:
  - They are less likely to use sound/sophisticated methods (more likely to use bad controls).
  - Their aggregate behavior affects the very correlations from which they draw causal inferences.
- "Behavioral" causal inference: Addressing these two differences

Today’s Model

- A DM makes a binary decision; tries to infer its causal effect on a binary outcome from long-run correlational data.
- Exogenous variables potentially confound this relation.
- The DM’s “data type” is defined by his set of (exogenous) control variables.
- In equilibrium, long-run data is consistent with each data type best-replying to his causal belief (based on his subjective controls).
- My question: What is the maximal expected welfare loss due to “bad controls” that can be sustained in equilibrium?

Example I

- $a \in \{0,1\}$ is an action.
- $y \in \{0,1\}$ is an outcome.
- The DM's utility is $y - ca$, where $c \in (0,1)$.
- $x \in \{0,1\}$ is an exogenous variable that the DM **may** observe prior to taking his action. It is the **only true cause** of y.
- The DM's type is defined by whether he conditions on or adjusts for x.

Example I: Types' Estimated Causal Effects

Conditioning on x:

$$p(y = 1 \mid a = 1, x) - p(y = 1 \mid a = 0, x) = 0$$

Adjusting for x:

$$\sum_x p(x)[p(y = 1 \mid a = 1, x) - p(y = 1 \mid a = 0, x)] = 0$$

No controls:

$$p(y = 1 \mid a = 1) - p(y = 1 \mid a = 0) \leq 1$$

Example I

- The types that do **not** condition on x will not vary their action with x, by definition.
- The type that **does** condition on x could potentially vary his action with x…
- …but will choose $a = 0$ for every x, because he correctly estimates a null causal effect.

Example I

- Consequently, no data type varies his action with x in equilibrium.
- Therefore, if p is consistent with equilibrium, a and x are independent, and the confounding effect of x **disappears**!

  $\Rightarrow$ All types will estimate a null causal effect.
- The equilibrium condition "protects" DMs from their causal errors: It **eliminates** any welfare loss due to bad controls.
- How general is this effect?

Some Background Literature

- The model here could be reformulated by adapting existing languages:
  - Analogy-based expectations (Jehiel 2005), Bayesian networks (Spiegler 2016), Berk-Nash equilibrium (Esponda-Pouzo 2016)
  - Earlier works rule out latent variables that directly cause DM actions.
- Behavioral implications of causal misperceptions: Spiegler 2016,2020
- Worst-case belief errors due to misspecified models: Eliaz-Spiegler-Weiss 2021
- “Non-Bayesian persuasion”: Hagenbach-Koessler 2020, Eliaz-Spiegler-Thysen 2021, Schwartzstein-Sunderam 2021, Levy-Moreno-Razin 2022

A Model

- $a \in \{0,1\}$ is the DM's action.
- $y \in \{0,1\}$ is an outcome.
- $t \in \{0,1\}$ is the DM's preference type.
- The DM's vNM utility is $u(t,a,y) = y - c \cdot \mathbf{1}[a \neq t]$.
  - $c \in (0,1)$ is a constant cost.
- $x = (x_1, \dots, x_K)$ is a collection of exogenous variables realized jointly with t, before the realization of a and y.
- **Baseline model**: a has no causal effect on y.

Data Types

- There is a set of "data types" N, enumerated $i = 1, .., n$.
- Each type i is defined by a distinct pair (C_i, D_i).

$$C_i \subseteq D_i \subseteq \{1, \dots, K\}$$

- The type has data revealing the long–run joint distribution of x_{D_i}, a, y.
- C_i is the set of x variables the type conditions on.
- $D_i \backslash C_i$ is the set of x variables the type adjusts for.
- The DM never has long-run statistical data on t.

Strategies

- $\lambda \in \Delta(N)$ is an independent distribution over data types.
- A strategy for type (t, i) is a strategy $\sigma_{t,i}: X \to \Delta\{0,1\}$.
  - $\sigma_{t,i}$ is measurable w.r.t x_{C_i}.
- p is a long-run distribution over x, t, a, y:

$$p(t, x, a, y) = p(t, x)p(a \mid t, x)p(y \mid t, x)$$

$$p(a \mid t, x) = \sum_{i \in N} \lambda_i \sigma_{t,i}(a \mid t, x)$$

Subjective Causal Belief

- Data type i's estimated consequence of choosing a:

$$\tilde{p}(y \mid do(a), x_{C_i}) = \sum_{x_{D_i \setminus C_i}} p(x_{D_i \setminus C_i} \mid x_{C_i}) p(y \mid a, x_{D_i})$$

- Pearl's *do* notation emphasizes that the conditioning represents a causal quantity, rather than a purely probabilistic one
- This formula would be correct if the DM employed "good controls".
- "Bad controls": Failing to control for confounders, or wrongly controlling for certain non-confounders.

Equilibrium

- A strategy profile σ with full support is an ε-equilibrium if for every

  x, a and every (t, i), $\sigma_{t,i}(a \mid t, x) > \varepsilon$ only if a maximizes

$$\sum_{y} \tilde{p}(y \mid do(a), x_{C_i}) u(t, a, y)$$

- An equilibrium is a limit of ε-equilibria for some sequence of $\varepsilon \rightarrow 0$.

The Case of Constant t

- Suppose $t = 0$ with certainty.
- The DM's type consists entirely of his data type i.
- The rational benchmark: Always play $a = 0$.
- The DM's expected welfare loss is $c \cdot \Pr(a = 1)$.
- What is the largest $\Pr(a = 1)$ we can sustain in equilibrium?
- In the Introduction's example, this probability was **zero**.

Example II

- $K = 2$ (two x variables, both take values in $\{0,1\}$)
- $p(y = 1 \mid x_1, x_2) = x_1 x_2$
- $n = 2$ (two data types), $\lambda_1 = \lambda_2 = 0.5$
- $C_i = D_i = \{i\}$ (type i conditions on x_i)
- Story: Business analysts with different expertise

Example II

- Suppose type i plays $a_i = x_i$ with certainty.
- Let's calculate type 1's estimated causal effect for every x_1.
- Note that $p(y = 1 \mid a, x_1 = 0) = 0$ for **every** a.
  - This conditional probability is based on "aggregate data" (across types).
- Therefore, $\Delta_1(x_1 = 0) = 0 < c$.
  - When $x_1 = 0$, type 1 prefers to play $a = 0$.

Example II

$$\Delta_1(x_1 = 1) = p(y = 1 \mid a = 1, x_1 = 1) - p(y = 1 \mid a = 0, x_1 = 1)$$

$$p(y = 1 \mid a = 1, x_1 = 1) = p(x_2 = 1 \mid a = 1, x_1 = 1)$$

$$= \frac{p(x_2 = 1 \mid x_1 = 1)}{p(x_2 = 1 \mid x_1 = 1) + p(x_2 = 0 \mid x_1 = 1)\lambda_1}$$

$$p(y = 1 \mid a = 0, x_1 = 1) = p(x_2 = 1 \mid a = 0, x_1 = 1) = 0$$

Example II

$$\Longrightarrow \qquad \Delta_1(x_1 = 1) = \frac{p(x_2 = 1 |\, x_1 = 1)}{p(x_2 = 1 |\, x_1 = 1) + 0.5p(x_2 = 0 |\, x_1 = 1)}$$

- If $p(x_2 = 1 |\, x_1 = 1) \approx 1$, then $\Delta_1(x_1 = 1) > c$.

$\Rightarrow$ When $x_1 = 1$, type 1 prefers to play $a = 1$.

Example II

- Type 1's strategy is consistent with equilibrium if $p(x_2 = 1|\, x_1 = 1) \approx 1$.
- The same reasoning works for type 2.
- If $p(x_1 = x_2 = 1) \approx 1$, the expected welfare loss is close to c.
- The equilibrium condition does **not** protect the DM from his erroneous causal inference due to "bad controls".
- The behavior of one type creates a confounding pattern for the other.

A Binary Relation

- Define a binary relation P over the set of data types N:
  - iPj if $D_i \supseteq C_j$
  - I.e., type i controls for every variable that type j conditions on.
- A binary relation is quasitransitive (Sen 1969) if its asymmetric part is transitive.

First Set of Characterization Results

Proposition 1: Suppose P is complete and quasitransitive. Then, the DM's equilibrium expected welfare loss is zero.

Proposition 2: Suppose P violates completeness or quasitransitivity. Then, there exist λ and $(p(x, y))$ that sustain $\Pr(a = 1) \approx 1$ in equilibrium.

First Set of Characterization Results

Proposition 1: Suppose P is complete and quasitransitive. Then, the DM's equilibrium expected welfare loss is zero.

Idea of proof:

- P partitions types into layers. At the top layer, types control for all relevant confounders.
- Therefore, top-layer types don't generate variation in a. This effectively removes confounders for the 2nd layer…
- …and by induction, this argument "infects" all layers.

First Set of Characterization Results

Proposition 2: Suppose P violates completeness or quasitransitivity.

Then, there exist λ and $(p(x,y))$ that sustain $\Pr(a=1) \approx 1$ in

equilibrium.

Idea of proof:

- When P is incomplete, we can construct something like Example II.
- When P violates quasitransitivity, we can construct a more elaborate version of Example II that involves **three** types.

The Case of Variable t

- Suppose t is the sole cause of y; the x variables are proxies of t.
- Denote $\delta_t = p(y = 1 \mid t)$. W.l.o.g, $\delta_1 \geq \delta_0$.
- Denote $\Pr(t = 1) = \gamma \in (0,1)$.
- The DM's expected welfare loss is

$$c \cdot [\gamma \cdot \Pr(a = 0 \mid t = 1) + (1 - \gamma) \cdot \Pr(a = 1 \mid t = 0)]$$

- Restrict attention to "simple" data types: $C_i = D_i$ for every i.
  - If P is complete, it is a linear ordering.

Example III

- Suppose $y = t$ deterministically.

- No x variables; the DM uses no controls:

$$\Delta = \Pr(y = 1 \mid a = 1) - \Pr(y = 1 \mid a = 0)$$

$$=$$

$$\Pr(t = 1 \mid a = 1) - \Pr(t = 1 \mid a = 0)$$

- The DM's best-reply to his belief increases with t.

$\Rightarrow$ $\Delta \geq 0$. The DM always plays $a = 1$ when $t = 1$.

Example III

- The DM's expected welfare loss is $c \cdot (1-\gamma) \cdot \sigma_{t=0}(a=1)$.
- The DM plays $a=1$ at $t=0$ only if $c \leq \Delta$.
- Therefore, the welfare loss is bounded from above by

$$[\Pr(t=1 \mid a=1) - \underbrace{\Pr(t=1 \mid a=0)}_{0}] \cdot (1-\gamma) \cdot \sigma_{t=0}(a=1)$$

$$\frac{\gamma \sigma_{t=1}(a=1)}{\gamma \sigma_{t=1}(a=1) + (1-\gamma)\sigma_{t=0}(a=1)} \cdot (1-\gamma) \cdot \sigma_{t=0}(a=1)$$

Example III

$$\frac{\gamma \sigma_{t=1}(a=1)}{\gamma \sigma_{t=1}(a=1) + (1-\gamma)\sigma_{t=0}(a=1)} \cdot (1-\gamma) \cdot \sigma_{t=0}(a=1)$$

- $\sigma_{t=1}(a=1) = 1$; the expression increases with $\sigma_{t=0}(a=1)$.
- This gives an upper bound of $\gamma(1-\gamma)$, which can be approximated arbitrarily well by selecting $c \approx \gamma$.
- **Intuition**: Error size (due to strong a-y correlation) is negatively related to error frequency.

Second Set of Characterization Results

- Recall $C = D$ for all data types; P is complete iff it is a linear order.
- Recall t is the only cause of y.

Proposition 3: Suppose P is complete. The DM's maximal expected equilibrium welfare loss is $\gamma(1-\gamma)$.

Proposition 4: Suppose P is incomplete. The DM's maximal expected equilibrium welfare loss is $\max\{\gamma, 1-\gamma\}$.

Proposition 3: A Few Words about the Proof

- The proof is based on an inductive argument that for every x and every type $i = 1, \ldots, n$, $\Delta_i(x) \geq 0$ and $\sigma_{t=1,i}(a = 1|x) = 1$.
  - All types agree on the causal effect's **sign**. This feature is crucial for the upper bound $\gamma(1-\gamma)$.
- The argument holds for $i = 1$, fundamentally because this type controls for every x variable the other types condition on.
- But unlike the constant t case, type 1 **does** vary his behavior, and thus exerts a "**confounding externality**" on the other types.

Proposition 3: A Few Words about the Proof

- This externality across types makes the inductive argument trickier.
- In particular, the way $\Pr(a = 1|t, x_{C_i})$ and $\Pr(a = 1|t, x_{C_{i+1}})$ vary with t could in principle exhibit "Simpson's paradox" (recall $C_{i+1} \subset C_i$):
  - $\Pr(a = 1|t, x_{C_i})$ increases in t for **every** x_{C_i}, yet the **coarser** conditional probability $\Pr(a = 1|t, x_{C_{i+1}})$ decreases in t.
  - The subtle part of the proof is showing **this anomaly does not arise** when P is complete.

More Stuff

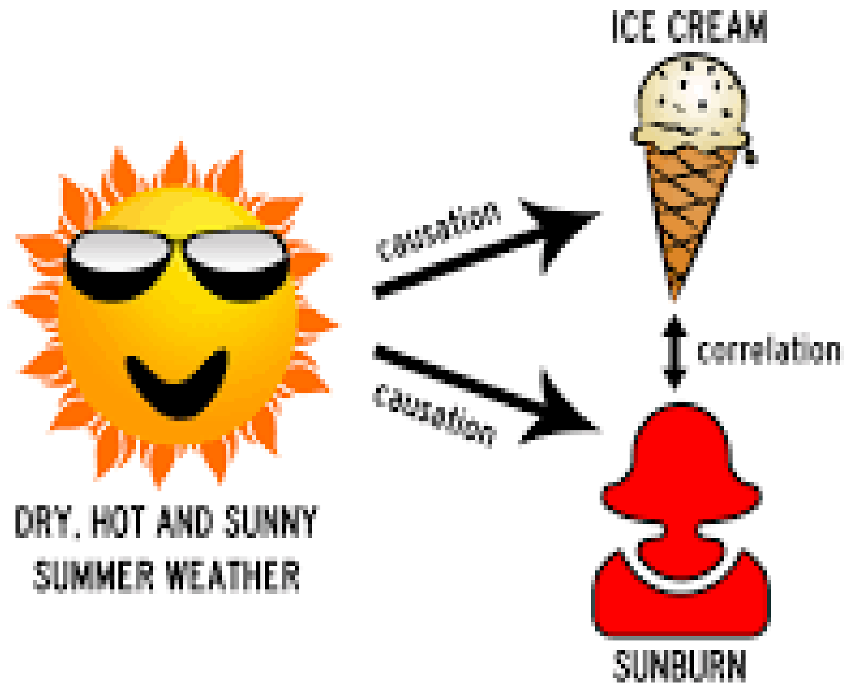
- When P is incomplete and we relax the assumption that $y \perp x \mid t$, the upper bound on the equilibrium welfare loss is 1.
- **Open problem**: Completing the characterization of upper bounds for general type spaces and data-generating processes
- Extension to non-null causal effects
  - Additively separable formulation: Results essentially intact
  - An "application": Partying in a pandemic

Summary

- DMs commit errors of causal inference from correlational data due to “bad controls”.
- The behavioral consequences of these errors shape the confounding patterns that lead to causal errors in the first place.
- Yet, equilibrium forces can drastically lower the cost of these errors.
- This equilibrium effect depends on the structure of the sets of control variables that different types of DMs employ.

Summary

- When the differentiation between data types is "**vertical**", the maximal welfare loss is substantially lower than in the non-equilibrium benchmark.
  - In some cases, the welfare loss can disappear entirely.
- When the differentiation is "**horizontal**", the "protective" equilibrium effect is much weaker (in the worst-case analysis).


ICE CREAM
causation
correlation
causation
DRY, HOT AND SUNNY
SUMMER WEATHER
SUNBURN