Causal Inference and Impact Evaluation

Rubin Causal Model and Selection

Roadmap

Two roads diverged in a yellow wood,

And sorry I could not travel both

And be one traveler...

[...]

Two roads diverged in a wood, and I —

I took the one less traveled,

And that has made all the difference.

("The Road Not Taken", Robert Frost)

Two roads diverged in a yellow wood,

And sorry I could not travel both

And be one traveler...

[...]

Two roads diverged in a wood, and I –

I took the one less traveled,

And that has made all the difference.

("The Road Not Taken", Robert Frost)

Two roads diverged in a yellow wood,

And sorry I could not travel both

And be one traveler...

[...]

Two roads diverged in a wood, and I —

I took the one less traveled,

And that has made all the difference.

("The Road Not Taken", Robert Frost)

— **How would you know?**

- Frost's life in the path not taken was observable *a priori*...

- Frost's life in the path not taken was observable *a priori*...
- *A posteriori*, once the road he took was taken, he could ("should", from an economist perspective) only wonder.

- Frost's life in the path not taken was observable *a priori*...
- *A posteriori*, once the road he took was taken, he could ("should", from an economist perspective) only wonder.
- In this sense, it constitutes a **counterfactual** — an alternative state of the world.

Fundamental Problem of Causal Inference

For each pair of observational unit (Frost) $\times$ action (taking the road he took), one can only observe outcomes in one state of the world.

- Assume that Frost's most important reason Y_i^{obs} for making this statement was winning Congressional Gold Medal in 1960, which he did.

- Assume that Frost's most important reason Y_i^{obs} for making this statement was winning Congressional Gold Medal in 1960, which he did.

- And let us denote the event of taking the road he took $W_i^{obs} = 1$.

- Assume that Frost's most important reason Y_i^{obs} for making this statement was winning Congressional Gold Medal in 1960, which he did.

- And let us denote the event of taking the road he took $W_i^{\text{obs}} = 1$.

- $Y_i(1) = 1$, but what about $Y_i(0)$? These were potential endpoints, again, from an *a priori* perspective.

- Assume that Frost's most important reason Y_i^{obs} for making this statement was winning Congressional Gold Medal in 1960, which he did.
- And let us denote the event of taking the road he took $W_i^{\text{obs}} = 1$.
- $Y_i(1) = 1$, but what about $Y_i(0)$? These were potential endpoints, again, from an *a priori* perspective.
- The following identity sums up the problem:

$$Y_i^{\text{obs}}(W_i^{\text{obs}}) = Y_i\,(1)\,W_i^{\text{obs}} + Y_i\,(0)\left(1 - W_i^{\text{obs}}\right) = \begin{cases} Y_i(0), & \text{if } W_i^{\text{obs}} = 0 \\ Y_i(1), & \text{if } W_i^{\text{obs}} = 1 \end{cases}$$

- Assume that Frost's most important reason Y_i^{obs} for making this statement was winning Congressional Gold Medal in 1960, which he did.
- And let us denote the event of taking the road he took $W_i^{\text{obs}} = 1$.
- $Y_i(1) = 1$, but what about $Y_i(0)$? These were potential endpoints, again, from an *a priori* perspective.
- The following identity sums up the problem:

$$Y_i^{\text{obs}}(W_i^{\text{obs}}) = Y_i(1)\, W_i^{\text{obs}} + Y_i(0)\left(1 - W_i^{\text{obs}}\right) = \begin{cases} Y_i(0), & \text{if } W_i^{\text{obs}} = 0 \\ Y_i(1), & \text{if } W_i^{\text{obs}} = 1 \end{cases}$$

- The **individual causal effect** of the road taken is $Y_i(1) - Y_i(0)$ on winning the medal.

Fundamental Problem of Causal Inference, Formally Holland (1986)

$Y_i(1) - Y_i(0)$ cannot be observed or gauged from available data for **any** unit i (Robert Frost, included).

- Important detour — **two perspectives** on observational units and what we may observe:

Population

Set of units for which we observe **all** the values we are interested.

Example

The set of municipalities in Brazil.

- Important detour — **two perspectives** on observational units and what we may observe:

Population

Set of units for which we observe **all** the values we are interested.

Example

The set of municipalities in Brazil.

Sample

Set of units for which we observe **some** values we are interested on.

Super-population

Large (but finite) population from which the sample is drawn.

Example

A set of people observed after sampling from these municipalities.

- From now on, if a definition/claim appears two times with different colors:

"finite population" (fp) perspective

"super-population" (sp) perspective

- If a definition/claim appears only one time in one color:[1]

"fp" + "sp" perspective

[1] I hope this it is obvious from context that both perspectives would yield the same concept.

Potential outcomes

Counter-factual behavior of **fixed quantities** in alternative states of the world.

Potential outcomes

Counter-factual behavior of **random variables** (with respect to a sampling distribution) in alternative states of the world.

- $\{1, \ldots, N\}$, **population** of size N;
- $(Y_i^{\text{obs}}, \mathbf{X}_i) \in \mathbb{R}^{k+1}$, **vector of observables** for the i-th unit, **outcomes** Y_i^{obs} and, possibly, **attributes** $\mathbf{X}_i$ – stacked versions $(\mathbf{Y}^{\text{obs}}, \mathbf{X})$;
- $W_i^{\text{obs}} \in \mathbb{T}_i$ is the **treatment** assigned for the i-th unit – stacked version $\mathbf{W}^{\text{obs}} \in \prod_{i=1}^{N} \mathbb{T}_i = \mathbb{T}_1 \times \ldots \times \mathbb{T}_N$;
- $\left\{Y_i(\mathbf{W}^{\text{obs}})\right\}_{\mathbf{W}^{\text{obs}} \in \prod_{i=1}^{N} \mathbb{T}_i}$ is the vector of of **potential outcomes** – stacked versions

- $\{1, \ldots, N_{\text{sp}}\}$, **super-population** of size N_{sp}...
- ...from which a **sample** $\{1, \ldots, N\}$ of size N is drawn;
- $W_i^{\text{obs}} \in \mathbb{T}_i$ is the **treatment** assigned for the i-th unit — stacked version $\mathbf{W}^{\text{obs}} \in \prod_{i=1}^{N} \mathbb{T}_i = \mathbb{T}_1 \times \ldots \times \mathbb{T}_N$;
- $\left\{Y_i(\mathbf{W}^{\text{obs}})\right\}_{\mathbf{W}^{\text{obs}} \in \prod_{i=1}^{N} \mathbb{T}_i}$ is the vector of of **potential outcomes** – stacked versions

- Although the **definition** of causal effects does not require more than one unit...
- ... **learning** about causal effects may be achieved with multiple units.

- Notice that, here, the counter-factuals or “alternative states of the world” are defined in the most general way possible.

- Notice that, here, the counter-factuals or “alternative states of the world” are defined in the most general way possible.

- If we observed data on Frost’s network of writers-friends and what this road is, we could **define** the (finite sample perspective) treatment effect of a specific combination of roads taken.

- Notice that, here, the counter-factuals or “alternative states of the world” are defined in the most general way possible.
- If we observed data on Frost’s network of writers-friends and what this road is, we could **define** the (finite sample perspective) treatment effect of a specific combination of roads taken.
- So, we (typically) impose some **stability** on how one’s road can affect other people...

- Notice that, here, the counter-factuals or "alternative states of the world" are defined in the most general way possible.

- If we observed data on Frost's network of writers-friends and what this road is, we could **define** the (finite sample perspective) treatment effect of a specific combination of roads taken.

- So, we (typically) impose some **stability** on how one's road can affect other people...

- and we assume that these roads are somewhat similar, at least in terms of the **causal mechanisms** they may activate.

Stable Unit Treatment Value Assumption (SUTVA)

The potential outcomes for any unit...

1. ... **do not vary** with the treatments assigned to other units ("No interference");
2. ... for each unit, there are **no different forms** or versions for each treatment level, which lead to potential different outcomes ("No hidden versions").

- Both components of SUTVA are examples of **exclusion restrictions**...

- They rely on external information to rule out the existence of causal effects of a particular treatment relative to an alternative.

1. "No interference":

   - Rules out **general equilibrium** effect, in contrast to **partial equilibrium** effect under a *ceteris paribus* assumption.

   - Allows us to project the whole vector of treatments on the i−th coordinate: $\left\{Y_i(\mathbf{W}^{\text{obs}})\right\} = \left\{Y_i(W_i^{\text{obs}})\right\}$.

2. "No hidden versions":
   - Ensures well-defined potential outcomes.

Causal Estimand

Formally, if $\mathbb{T} = \{0, 1\}$ a **causal estimand** τ is a row-exchangeable function:

$$\tau = \tau\left(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W}^{\text{obs}}\right)$$

These are the ultimate objects of interest, since they sum up features of the distribution of individual causal effects.

Example

$Y_i(1) - Y_i(0)$ and $\frac{Y_i(1)}{Y_i(0)}$ are **unit-level causal effects**.

Example

$Y_i(1) - Y_i(0)$ and $\frac{Y_i(1)}{Y_i(0)}$ are **unit-level causal effects.**

Average Treatment Effect Parameter (ATE)

$$\tau_{\mathtt{ATE}}^{\mathrm{fs}} \overset{\mathrm{def}}{=} N^{-1} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$$

Average Treatment Effect Parameter (ATE)

$$\tau_{\mathtt{ATE}}^{\mathrm{sp}} \overset{\mathrm{def}}{=} \mathbb{E}_{\mathrm{sp}}[Y_i(1) - Y_i(0)]$$

Statistic

A **statistic** T is a known, real-valued function $T(\mathbf{Y}, \mathbf{W}^{\text{obs}}, \mathbf{X})$.

Identification step

Stating assumptions that allows one to relate the causal estimands to moments we would be able observe in a sample or population.

Identification step

Stating assumptions that allows one to relate the causal estimands to moments we would be able observe in a sample or population.

Estimation step

Defining functions of the data that, in some sense, approximate the causal estimands identified.

Identification step

Stating assumptions that allows one to relate the causal estimands to moments we observe in a sample or population.

Estimation step

Defining statistics that approximate the causal estimands identified.

Identification step

Stating assumptions that allows one to relate the causal estimands to moments we observe in a sample or population.

Estimation step

Defining functions of the data that, in some sense, approximate the causal estimands identified.

Inference step

Incorporating uncertainty into our conclusions about estimates of causal estimands.

Confidence Interval

Given an estimand τ, a **confidence interval** with coefficient $1-\alpha$ is a pair of real-valued functions $C_L(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X})$ and $C_U(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X})$, defining an interval

$$[C_L(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}), C_U(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X})]$$

such that:

$$\mathrm{P}_W\left[C_L(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X}) \leq \tau \leq C_u(\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}}, \mathbf{X})\right] \geq 1-\alpha$$

Confidence Interval

Internal validity

Quality of estimating causal estimands credibly for our particular sample or population.

External validity

Quality of estimating causal quantities that would carry over to other samples or populations.

The Law of Decreasing Credibility (Manski, 2003)

The credibility of inference decreases with the strength of the assumptions maintained.

Assignment Mechanism

Given a population of N units, the **assignment mechanism** is a row-exchangeable function $\mathrm{P}(\mathbf{W}^{\mathrm{obs}}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ taking values between 0 and 1 and satisfying:

$$\sum_{\mathbf{W}^{\mathrm{obs}} \in \{0,1\}^N} \mathrm{P}_W \left[\mathbf{W}^{\mathrm{obs}}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right] = 1,$$

for all $\mathbf{X}$, $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$.

We denote by $\mathbb{W}^+$ the set of assignments with strictly positive probability.

Assignment Mechanism

Given a population of N units, the **assignment mechanism** is a row-exchangeable function $P(\mathbf{W}^{obs}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ taking values between 0 and 1 and satisfying:

$$\sum_{\mathbf{W}^{obs} \in \{0,1\}^N} P_W\left[\mathbf{W}^{obs}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right] = 1,$$

for all $\mathbf{X}$, $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$.

We denote by $\mathbb{W}^{+}$ the set of assignments with strictly positive probability.

Assignment Mechanism

Given a population of N units, the **assignment mechanism** is a row-exchangeable function $P(\mathbf{W}^{obs}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ taking values between 0 and 1 and satisfying:

$$\sum_{\mathbf{W}^{obs} \in \{0,1\}^N} P_W \left[\mathbf{W}^{obs}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right] = 1,$$

for all $\mathbf{X}$, $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$.

We denote by $\mathbb{W}^+$ the set of assignments with strictly positive probability.

Example

$\mathbb{W}^+ = \left\{ \binom{0}{0}, \binom{1}{0}, \binom{0}{1}, \binom{1}{1} \right\}$ with equal probability

Assignment Mechanism

Given a sample of N units from a super-population, the **assignment mechanism** is the first term of the probability density function f induced by random sampling on the quadruple $(Y_i(0), Y_i(1), W_i^{obs}, \mathbf{X}_i)$:

$$\begin{aligned} f_{W,Y(0),Y(1),\mathbf{X}}(W_i^{obs}, Y_i(0), Y_i(1), \mathbf{X}_i; \theta, \lambda, \phi) = \\ f_{W|Y(0),Y(1),X}(W_i^{obs}|Y_i(0), Y_i(1), \mathbf{X}_i; \theta) \times \\ f_{Y(0),Y(1)|X}(Y_i(0), Y_i(1)|\mathbf{X}_i; \lambda) \times \\ f_X(\mathbf{X}_i|\phi) \end{aligned}$$

where the first one is the **assignment mechanism**, and (θ, λ, ϕ) is a global parameter.

Assignment Mechanism

Given a sample of N units from a super-population, the **assignment mechanism** is the first term of the probability density function f induced by random sampling on the quadruple $(Y_i(0), Y_i(1), W_i^{obs}, \mathbf{X}_i)$:

$$
\begin{aligned}
f_{W,Y(0),Y(1),\mathbf{X}}(W_i^{obs}, Y_i(0), Y_i(1), \mathbf{X}_i; \theta, \lambda, \phi) = \\
f_{W|Y(0),Y(1),X}(W_i^{obs}|Y_i(0), Y_i(1), \mathbf{X}_i; \theta) \times \\
f_{Y(0),Y(1)|X}(Y_i(0), Y_i(1)|\mathbf{X}_i; \lambda) \times \\
f_X(\mathbf{X}_i|\phi)
\end{aligned}
$$

where the first one is the **assignment mechanism**, and (θ, λ, ϕ) is a global parameter.

Assignment Mechanism

Given a sample of N units from a super-population, the **assignment mechanism** is the first term of the probability density function f induced by random sampling on the quadruple $(Y_i(0), Y_i(1), W_i^{\text{obs}}, \mathbf{X}_i)$:

$$\begin{aligned} f_{W,Y(0),Y(1),\mathbf{X}}(W_i^{\text{obs}}, Y_i(0), Y_i(1), \mathbf{X}_i; \theta, \lambda, \phi) = \\ f_{W|Y(0),Y(1),X}(W_i^{\text{obs}}|Y_i(0), Y_i(1), \mathbf{X}_i; \theta) \times \\ f_{Y(0),Y(1)|X}(Y_i(0), Y_i(1)|\mathbf{X}_i; \lambda) \times \\ f_X(\mathbf{X}_i|\phi) \end{aligned}$$

where the first one is the **assignment mechanism**, and (θ, λ, ϕ) is a global parameter.

Example

?

Unit-Level Assignment Probability

The **unit-level assignment probability** for an arbitrary unit $i \in \{1, ..., N\}$ is defined as:

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = \sum_{\mathbf{W}^{\text{obs}}: W_i^{\text{obs}}=1} \mathrm{P}(\mathbf{W}^{\text{obs}} | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)).$$

Unit Assignment Probability

Function $f_{W|Y(0),Y(1),X}(1|Y_i(0), Y_i(1), \mathbf{X}_i, \theta)$.

Propensity Score

For units with $\mathbf{X}_i = x$, the finite population **propensity score** is defined as:

$$p(\mathbf{x}) = \frac{1}{N(\mathbf{x})} \sum_{\mathbf{X}_i : \mathbf{X}_i = \mathbf{x}} p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)).$$

where $N(x)$ is the number of units in the population for which $\mathbf{X}_i = x$.

For values of x with $N(x) = 0$, the propensity score is defined to be zero.

Propensity Score

For all x in the support of $\mathbf{X}_i = x$, it is defined as:

$$p(\mathbf{x}) =$$
$$\mathbb{E}_{sp}\left[f_{W|Y(0),Y(1),X}(1|Y_i(0), Y_i(1), \mathbf{X}_i, \theta) \times f_{Y(0),Y(1)|X}(Y_i(0), Y_i(1)|\mathbf{X}_i, \lambda)|\mathbf{X}_i = \mathbf{x}\right]$$

Roadmap

Individualistic Assignment Mechanisms

An assignment mechanism $\mathrm{P}(\mathbf{W}^{\text{obs}}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ is **individualistic** if, for some function $q(.)$ taking values between 0 and 1:

$$p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = q(\mathbf{X}_i, Y_i(0), Y_i(1)), \ \forall\, i$$

and:

$$\mathrm{P}(\mathbf{W}^{\text{obs}}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = c \prod_{i=1}^{N} \left[q(\mathbf{X}_i, Y_i(0), Y_i(1))\right]^{W_i^{\text{obs}}} \left[1 - q(\mathbf{X}_i, Y_i(0), Y_i(1))\right]^{1-W_i^{\text{obs}}}$$

for $\mathbf{W}^{\text{obs}}, \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1) \in \mathbb{A}$, for some set $\mathbb{A}$, and zero elsewhere (c is the constant that ensures that the probabilities sum to unity).

Intuitively, this happens when the unit assignment probabilities $p_i(.)$ for each i do not depend on the outcomes and assignments for other units.

Probabilistic Assignment Mechanisms

An assignment mechanism $\mathrm{P}(\mathbf{W}^{\mathrm{obs}}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ is **probabilistic** if:

$$0 < p_i(\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 0, \ \forall\, i$$

that is, when the unit assignment probabilities $p_i(.)$ are strictly between 0 and 1 and every unit has the possibility of being assigned to the active treatment and the possibility of being assigned to the control condition.

Unconfounded Assignment Mechanisms

An assignment mechanism $P(\mathbf{W}^{obs}|\mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1))$ is **unconfounded** if:

$$P(\mathbf{W}^{obs}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = P(\mathbf{W}^{obs}|\mathbf{Y}'(0), \mathbf{Y}'(1), \mathbf{X})$$

for all $\mathbf{W}^{obs}$, $\mathbf{X}$, $\mathbf{Y}(0)$, $\mathbf{Y}(1)$ and $\mathbf{Y}'(0)$, $\mathbf{Y}'(1)$.

Intuitively, unconfoundedness rules out the dependence between the assignment mechanism and potential outcomes.

In this case, we write the assignment mechanism as $P(\mathbf{W}^{obs}|\mathbf{X})$.

Proposition

If an assignment mechanism is individualistic and unconfounded, then the assignment mechanism is the product of the propensity scores:

$$p(x) = q(x) \tag{1}$$

for all x in the support of $\mathbf{X}_i$

Randomized Experiment

An assignment mechanism corresponds to a **randomized experiment** if it is probabilistic and has a known functional form that is controlled by the researcher.

Classical Randomized Experiment

An assignment mechanism corresponds to a **classical randomized experiment** if it is individualistic, probabilistic, unconfounded and has a known functional form that is controlled by the researcher.

Regular Assignment Mechanisms

We say that an assignment mechanism is **regular** if it is individualistic, probabilistic and unconfounded.

Observational Study with Regular Assignment Mechanism

An assignment mechanism corresponds to a **observational study with regular assignment mechanism** if it is individualistic, probabilistic, unconfounded and has unknown functional form.

Bernoulli Trial

A **Bernoulli trial** is a classical randomized experiment with an assignment mechanism such that the assignments for all units are independent.

Completely Randomized Experiment

A **completely randomized experiment** is a classical randomized experiment in which a fixed number N_t of subjects is assigned to receive the active treatment, i.e.:

$$\mathbb{W}^{+} = \left\{ \mathbf{W}^{\text{obs}} \middle| \sum_{i=1}^{N} W_i^{\text{obs}} = N_t \right\}$$

for some number $N_t \in \{1, \ldots, N-1\}$. In this case:

$$\mathrm{P}(\mathbf{W}^{\text{obs}} | \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = \binom{N}{N_t}^{-1} \text{if } \sum_{i=1}^{n} W_i^{\text{obs}} = N_t$$

and 0 otherwise.

Stratified and Paired Randomized Experiment

A **stratified randomized experiment** with J groups (called blocks or strata) defined by pre-treatment variables is a classical randomized experiment in which a fixed number of subjects $N_t(j)$ for each group is assigned to receive the active treatment i.e.:

$$\mathbb{W}^+ = \left\{ \mathbf{W}^{obs} \middle| \sum_{i:B_i=j}^{n} W_i^{obs} = N_t(j), \text{for } j = 1, 2, \ldots, J \right\}$$

and

$$P(W^{obs}|\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = \begin{cases} \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1} & \text{if } \mathbf{W}^{obs} \in \mathbb{W}^+ \\ 0, & \text{otherwise} \end{cases}$$

A **paired randomized experiment** is a stratified randomized experiment with $N(j) = 2$ and $N_t(j) = 1$.

- Notice that the cardinality of $\mathbb{W}^+$ gradually decreases as we move from Bernoulli trials to paired randomized experiments.
- This helps to eliminate "unhelpful" assignment vectors that are *a priori* unlikely to lead to precise causal inferences.
- Think of a Bernoulli trial, where there exists a chance that almost every unit be assigned to treatment.
- If blocks are chosen based on characteristics that are related to the distribution of potential outcomes, the same logic applies.

Roadmap

1. What is (are) the **causal relationship(s)** of interest?
    - What is the relationship in the most concrete way you can describe it?

1. What is (are) the **causal relationship(s)** of interest?
   - What is the relationship in the most concrete way you can describe it?

- Nice, but it is easy to get lost with so many details...

2. What would be the **perfect experiment** to capture this causal relationship?
    - If one follows the maximum “no causation without manipulation”, the question of Dupas (2014) is certaily different from the question of

3. What is the **identifying variation**?

4. What is the **empirical strategy** (data, estimation and inference)?

Roadmap

Proposition (Identification of $\tau_{\texttt{ATE}}$ under Random Assignment Mechanisms — Or "Randomization Solves the Selection Problem")

Assume that the assignment mechanism satisfies the following assumption of randomization:

$$\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp W_i^{obs} \tag{CRE.1}$$

Then $\tau_{\texttt{ATE}}$ is identified, i.e., it can be written as a function of moments of observable data:

$$\tau_{\texttt{ATE}} = \mathrm{E}[Y_i | W_i^{obs} = 1] - E[Y_i | W_i^{obs} = 0]$$

where the expectation is taken over the sampling distribution and the distribution induced by the assignment mechanism.

Proof.

Use the identity $Y_i^{\text{obs}} = Y_i(1)\,W_i^{\text{obs}} + Y_i(0)\left(1 - W_i^{\text{obs}}\right)$ to write:

$$\begin{aligned} \mathrm{E}(Y_i|W_i^{\text{obs}} = 1) - \mathrm{E}(Y_i|W_i^{\text{obs}} = 0) \equiv \\ & \underbrace{\mathrm{E}(Y_i(1) - Y_i(0)|W_i^{\text{obs}} = 1)}_{=\tau_{\mathtt{ATE}},\ \text{given (CRE.1)}} + \\ & \underbrace{[\mathrm{E}(Y_i(0)|W_i^{\text{obs}} = 1) - \mathrm{E}(Y_i(0)|W_i^{\text{obs}} = 0)]}_{\mathcal{B},\ \text{for "selection bias"}} \end{aligned}$$

and use (CRE.1) to conclude that $\mathcal{B} = 0$ □

- Let us look a little closes to $\mathcal{B}$, the selection bias that would arise in a difference-in-means absent randomization:

$$\mathcal{B} \stackrel{\text{def}}{=} [\mathrm{E}(Y_i(0)|W_i^{\text{obs}} = 1) - \mathrm{E}(Y_i(0)|W_i^{\text{obs}} = 0)]$$

- Let us look a little closes to $\mathcal{B}$, the selection bias that would arise in a difference-in-means absent randomization:

$$\mathcal{B} \stackrel{\text{def}}{=} [\mathrm{E}(Y_i(0)|W_i^{\text{obs}} = 1) - \mathrm{E}(Y_i(0)|W_i^{\text{obs}} = 0)]$$

- It captures the difference in potential outcomes in the untreated state between treatment and comparison units.

Example (Textbook Distribution to Schools Absent Randomization)

1. $E(Y_i(0)|W_i^{obs} = 1) - E(Y_i(0)|W_i^{obs} = 0) > 0$ would be consistent with:
    - parents that value education more highly are likely to encourage home investments;
    - principals have private knowledge on teachers' human capital limitations and self-select into treatment based on these gains.
2. $E(Y_i(0)|W_i^{obs} = 1) - E(Y_i(0)|W_i^{obs} = 0) < 0$ would be consistent with:
    - distribution is targeted to schools with *ex ante* bad performance;
    - schools have private knowledge on teachers' human capital limitations and self-select into treatment based on these gains

Example

Roadmap

- Assuming we have a design satisfying $\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp W_i^{\text{obs}}$, it is important to keep in mind the distinction between:

  1. **reduced-form estimates of treatment impacts** — total derivatives of production functions, model-free approach is sufficient.
  2. **structural estimates of treatment impacts** — partial derivatives of production functions, model-free approach is **not** sufficient.

- Assuming we have a design satisfying $\big(Y_i(0), Y_i(1)\big) \perp\!\!\!\perp W_i^{\text{obs}}$, it is important to keep in mind the distinction between:

  1. **reduced-form estimates of treatment impacts** — total derivatives of production functions, model-free approach is sufficient.
  2. **structural estimates of treatment impacts** — partial derivatives of production functions, model-free approach is **not** sufficient.

- In general Partial derivatives can only be obtained if researchers specify the model that links various inputs to the outcomes of interest and collect data on these intermediate inputs.

References I

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica*, 82(1):197–228.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Manski, C. F. (2003). *Partial identification of probability distributions.* Springer Science & Business Media.