Causal Inference: Identification Under Strong Ignorability

Hyunseung Kang

2025-02-04

Concepts Covered Today

▶ Identification under strong ignorability
- ▶ ATE, ATT, causal odds ratio, causal relative risk
- ▶ Single-time, static optimal policy

▶ Observational studies and strong ignorability

▶ Propensity score and its properties

▶ References:
- ▶ Chapter 3 of M. Hernán and Robins (2020)

Review: Causal Identification Under Complete Randomized Experiment I

Suppose we are interested in identifying the average treatment effect (ATE), $\mathbb{E}[Y(1) - Y(0)]$. Under an ideal, complete randomized experiment, the following assumptions are satisfied:

▶ (A1,SUTVA): $Y = AY(1) + (1 - A)Y(0)$
▶ (A2, Randomization of A): $A \perp X, Y(1), Y(0)$
▶ (A3, Positivity): $0 < P(A = 1) < 1$

Assumptions (A2) and (A3) can also be interpreted as consequences of missing at completely random (MCAR) in the missing data literature.

We have illustrated both approaches to motivate (A2) and (A3) using the following table:

Review: Causal Identification Under Complete Randomized Experiment II

	$Y(1)$	$Y(0)$	Y	A	X_{Age}
John	NA	0.94	0.94	0	23
Sally	NA	0.91	0.91	0	27
Kate	0.81	NA	0.81	1	32
Jason	0.60	NA	0.60	1	30

Review: Causal Identification Under Complete Randomized Experiment III

This lecture discusses a popular set of identifying assumptions to identify causal effects known as **strong ignorability.**

Motivating Strong Ignorability: Stratified Randomized Experiments I

In most randomized experiments, treatment is not randomized completely at random.

▶ Often, treatment is randomized within a pre-defined block of individuals.

▶ The blocks are defined by covariates X.

This type of randomized experiment is broadly known as **stratified/blocked experiments.** Some examples include

▶ From the above table, suppose we partition individuals into two blocks: those who are less than 30 years old and those who are greater than or equal to 30 years old. Treatment is randomly assigned within each block.

Motivating Strong Ignorability: Stratified Randomized Experiments II

▶ Consider a hypothetical randomized experiment to assess the causal effect of a new drug on reducing blood pressure. Suppose we recruit identical twins and we randomly assign treatment within each twin. Here, each twin defines a block.

The treatment probabilities can differ across blocks.

▶ Individuals in block one has 70% chance of getting treated
▶ Individuals in block two has 50% chance of getting treated
▶ …
▶ But, within each block, the treatment is assigned randomly.

Formalizing Strong Ignorability I

We can formalize the assumptions of a stratified randomized experiment as follows:

▶ (A1, SUTVA): $Y = AY(1) + (1 - A)Y(0)$
▶ (A2, Conditional randomization of A): $A \perp Y(1), Y(0)|X$
▶ (A3, Positivity/Overlap): $0 < \mathbb{P}(A = 1|X = x) < 1$ for all x
  where $f(x) > 0$ and $f(\cdot)$ denotes the density of X.

Assumptions (A2) states that A is now conditionally independent of $Y(1), Y(0)$ given X.

Assumption (A3) states that the treatment probability depends on X and the probability of getting treated must be between 0 and 1.

For example, if X is age:

▶ (A2) states that conditional on the age group, treatment A is randomly assigned to individuals within that age group.

Formalizing Strong Ignorability II

▶ (A3) states that conditional on the age group, each person has a non-zero probability of receiving treatment or control.
  ▶ Note that the treatment probability may be different across age groups, like a stratified randomized experiment example above.

Other remarks about the assumptions:

▶ The function $\mathbb{P}(A = 1 | X = x)$ is so famous that it's called the **propensity score** (Rosenbaum and Rubin (1983)).
  ▶ It gets its name because it quantifies the propensity for an individual with covariates X to get treated.
  ▶ We'll discuss the propensity score in greater detail later.
▶ Assumptions (A2) and (A3) are known as **strong ignorability** (Rosenbaum and Rubin (1983)).

Connection to Missing Data: MAR I

Similar to the case under a complete randomized experiment, (A2) and (A3) have connection to the **missing at random (MAR)** assumption in the missing data literature.

To illustrate, let's consider the table below:

	$Y(1)$	$Y(0)$	Y	A	$X_{\text{Under 30 years}}$
John	NA	0.94	0.94	0	1
Sally	NA	0.91	0.91	0	1
Sarah	0.70	NA	0.70	1	1
Kate	0.81	NA	0.81	1	0
Jason	0.60	NA	0.6	1	0
Jack	NA	0.88	0.88	0	0

Connection to Missing Data: MAR II

Assumption (A2) states that within the rows of the table where Xs are identical (i.e. conditional on X), the missingness indicator A is completely independent of the columns $Y(1), Y(0)$.

Assumption (A3) states that within the rows of the table where Xs are identical, some values of $Y(1)$ (or $Y(0)$) are observed and this holds for every value of X.

Identification Under Strong Ignorability: CATE I

Under strong ignorability, it's relatively straightforward to identify the ATE among a subgroup defined by X:

$$\mathbb{E}[Y(1) - Y(0)|X = x]$$

This is referred to as **the conditional average treatment effect (CATE)**.

Intuitively, to identify the CATE:

▶ We consider a smaller table where we only have individuals with $X = x$.

▶ Then, similar to a completely randomized experiment, we can identify $\mathbb{E}[Y(1)|X = x]$ by taking the average of the observed $Y(1)$.

Identification Under Strong Ignorability: CATE II

More formally, for any x, we have

$$
\begin{aligned}
&\mathbb{E}[Y|A = 1, X = x] \\
=&\mathbb{E}[AY(1) + (1 - A)Y(0)|A = 1, X = x] \qquad \text{(A1)} \\
=&\mathbb{E}[Y(1)|A = 1, X = x] \qquad\qquad\qquad\quad \text{Algebra} \\
=&\mathbb{E}[Y(1)|X = x] \qquad\qquad\qquad\qquad\quad\;\; \text{(A2)}
\end{aligned}
$$

Assumption (A3) ensures that the conditioning event $\mathbb{E}[Y|A = 1, X = x]$ is well-defined.

Identification Under Strong Ignorability: ATE I

Once we identified the CATE, we can immediately identify the average treatment effect $\mathbb{E}[Y(1) - Y(0)]$ from the law of total expectations.

For the unconditional mean $\mathbb{E}[Y(1)]$, we have

$$\mathbb{E}[Y(1)] = \mathbb{E}[\mathbb{E}[Y(1)|X]] \qquad \text{Law of total expectation}$$
$$= \mathbb{E}[\mathbb{E}[Y|A = 1, X]] \qquad \text{Argument from above}$$

Repeating the above argument identifies $\mathbb{E}[Y(0)] = \mathbb{E}[\mathbb{E}[Y|A = 0, X]]$ and thus, the ATE is identified via

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y|A = 1, X]] - \mathbb{E}[\mathbb{E}[Y|A = 0, X]]$$

Identification of the Average Treatment Effect Among the Treated (ATT) I

In addition to the ATE, there is another popular causal estimand called the **average treatment effect among the treated (ATT)**, or formally

$$\text{ATT} = \mathbb{E}[Y(1) - Y(0) \mid A = 1]$$

In words, ATT is the average treatment effect among individuals who received treatment.

Alternatively, consider the data table below.

	$Y(1)$	$Y(0)$	Y	A	X_{Age}
John	NA	0.94	0.94	0	23
Sally	NA	0.91	0.91	0	27
Kate	0.81	NA	0.81	1	32
Jason	0.60	NA	0.60	1	30

Identification of the Average Treatment Effect Among the Treated (ATT) II

▶ The ATT represents the average of the differences of $Y(1) - Y(0)$ among Kate and Jason, both of whom were treated.

▶ Note that the ATT is different than the ATE, which represents the average of the differences of $Y(1) - Y(0)$ for both treated and untreated individuals.

A feature of the ATT is that you can identify the causal effect by a weaker version of strong ignorability, i.e.

▶ (A2.ATT): $A \perp Y(0)|X$

▶ (A3.ATT): $\mathbb{P}(A = 1 \mid X = x) < 1$ for all x where $f(x) > 0$ and $0 < \mathbb{P}(A = 1)$

    ▶ Remember that f denotes the density function.

Comparing the "original" strong ignorability (A2) with the weaker versions (A2.ATT):

Identification of the Average Treatment Effect Among the Treated (ATT) III

▶ ATT identification does not need conditional independence between A and $Y(1)$.

  ▶ In the context of missing data, we only need the missingness indicator A to be independent of the column $Y(0)$, not necessarily with the column $Y(1)$.

  ▶ From my experience, the practical difference between (A2.ATT) and (A2) and the discussions about the plausibility of the identifying assumptions in an observational study is minor.

Comparing the "original" positivity (A3) with the weaker version (A3.ATT),

▶ (A3.ATT) requires that there are some controls for all values of X.

▶ Another way to say this is that there cannot be a region in the space of covariates where everyone gets treated

Identification of the Average Treatment Effect Among the Treated (ATT) IV

This weaker version of strong ignorability is often attributed to Heckman, Ichimura, and Todd (1997). But, I cannot find the exact page where (A3.ATT) is actually mentioned…

To prove identification with the weaker assumption, the term $\mathbb{E}[Y(1) \mid A = 1]$ in the ATT can be identified with just assumption (A1, SUTVA) and (A3.ATT):

$$\mathbb{E}[Y(1) \mid A = 1] = \mathbb{E}[Y \mid A = 1] \qquad \text{(A1)}$$

The second part of (A3.ATT) ensures that the conditional expectation is well-defined.

In contrast, the term $\mathbb{E}[Y(0) \mid A = 1]$ in the ATT must be identified with (A1), (A2.ATT), and (A3.ATT).

Identification of the Average Treatment Effect Among the Treated (ATT) V

$$\begin{aligned}
\mathbb{E}[Y(0) \mid A = 1] &= \mathbb{E}[\mathbb{E}[Y(0) \mid A = 1, X]|A = 1] \quad \text{Law of total expectati} \\
&= \mathbb{E}[\mathbb{E}[Y(0) \mid A = 0, X]|A = 1] \quad \text{(A2.ATT)} \\
&= \mathbb{E}[\mathbb{E}[Y \mid A = 0, X]|A = 1] \quad \text{(A1)}
\end{aligned}$$

▶ The second part of (A3.ATT) ensures that the conditional expectation on the left-hand-side of the first equality is well-defined.

▶ All of (A3.ATT) ensures that the conditional expectation on the right-hand-side of the first equality is well-defined.

   ▶ (A3.ATT) implies that there is a region of X where $\mathbb{P}(A = 1 \mid X = x) > 0$ is positive because

$$0 < \mathbb{P}(A = 1) = \int_{x, f(x) > 0} \mathbb{P}(A = 1 \mid X = x) f(x) dx$$

Identification of the Average Treatment Effect Among the Treated (ATT) VI

Technically, the inner expectation $\mathbb{E}[\mathbb{E}[Y(0) \mid A = 1, X]|A = 1]$ is taken with respect to regions of X where $f(x|A = 1) > 0$, i.e.,

$$
\mathbb{E}[\mathbb{E}[Y(0) \mid A = 1, X]|A = 1]
$$
$$
= \int_y \int_{x, f(x|A=1)>0} yf(y \mid A = 1, X = x)f(x \mid A = 1)dxdy
$$

Hence, under (A1), (A2.ATT), and (A3.ATT), we can identify the ATT via

$$
\mathbb{E}[Y(1) - Y(0) \mid A = 1]
$$
$$
= \mathbb{E}[Y \mid A = 1] - \mathbb{E}[\mathbb{E}[Y \mid A = 0, X] \mid A = 1]
$$

We list some popular causal estimands when the outcome is binary and prove that they can be identified under (A1) and strong ignorability

Under a binary outcome, some popular causal estimands are the causal relative risk (CRR) and the causal odds ratio (COR):

$$\mathrm{CRR} = \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} = \frac{\mathbb{P}(Y(1) = 1)}{\mathbb{P}(Y(0) = 1)}$$

$$\mathrm{COR} = \frac{\frac{\mathbb{P}(Y(1)=1)}{1-\mathbb{P}(Y(1)=1)}}{\frac{\mathbb{P}(Y(0)=1)}{1-\mathbb{P}(Y(0)=1)}}$$

Identification of Other Measures of Causal Effects: Causal Relative Risk (CRR) and Causal Odds Ratio (COR) II

▶ Despite their popularity, there are some issues with defining causal odds ratios (or more generally odds ratios) due to non-collapsibility issues; see Greenland, Pearl, and Robins (1999) and M. A. Hernán, Clayton, and Keiding (2011).

▶ I would recommend using CRRs instead of CORs unless the scientific question is expressed in terms of odds ratios.

▶ Note that the original ATE $\mathbb{E}[Y_i(1) - Y_i(0)]$, or a linear contrast of the outcomes, also works for binary outcomes.

Identification of the CRR and the COR proceeds by identifying $\mathbb{E}[Y(a)]$ for $a \in \{0, 1\}$, which make up the CRR and the COR. In particular, we have

$$
\begin{aligned}
\mathbb{E}[Y(a)] &= \mathbb{E}[\mathbb{E}[Y(a) \mid X]] && \text{Law of total expectation} \\
&= \mathbb{E}[\mathbb{E}[Y(a) \mid A = a, X]] && \text{(A2) and (A3)} \\
&= \mathbb{E}[\mathbb{E}[Y \mid A = a, X]] && \text{(A1)}
\end{aligned}
$$

Identification of Other Measures of Causal Effects: Causal Relative Risk (CRR) and Causal Odds Ratio (COR) III

We need (A3) to ensure that the conditioning event $\{A = a, X\}$ is well-defined. Then, under (A1), (A2), and (A3), the CRR and the COR are identified as

$$\mathrm{CRR} = \frac{\mathbb{E}[\mathbb{E}[Y \mid A = 1, X]]}{\mathbb{E}[\mathbb{E}[Y \mid A = 0, X]]}$$

$$\mathrm{COR} = \frac{\frac{\mathbb{E}[\mathbb{E}[Y \mid A=1,X]]}{1-\mathbb{E}[\mathbb{E}[Y \mid A=1,X]]}}{\frac{\mathbb{E}[\mathbb{E}[Y \mid A=0,X]]}{1-\mathbb{E}[\mathbb{E}[Y \mid A=0,X]]}}$$

Policy Learning: Single, Static, Optimal Treatment Rule (OTR) I

In personalized medicine, the goal is to operationalize how doctors assign treatment to patients. This is done by defining an optimal treatment assignment rule, which takes in patients' covariates X and outputs either treatment (i.e., 1) or control (i.e., 0) that maximizes the patient's outcome (in expectation).

▶ Consider a policy function $\pi : \mathcal{X} \to \{0, 1\}$ which assigns either treatment (i.e 1) or control (i.e 0) based on the individual's characteristic $X \in \mathcal{X}$.

▶ The goal is to find the best π, denoted as π_{OTR}, that maximizes the expected counterfactual outcome under that policy:

Policy Learning: Single, Static, Optimal Treatment Rule (OTR) II

$$\pi_{\text{OTR}} = \underset{\pi:\mathcal{X}\to\{0,1\}}{\text{argmax}} \ \mathbb{E}[Y(\pi(X))]$$

The term $Y(\pi(X))$ is the counterfactual outcome if treatment is assigned based on the policy π and can be written as

$$Y(\pi(X)) = Y(1)I(\pi(X) = 1) + Y(0)I(\pi(X) = 0)$$

The term $\mathbb{E}[Y(\pi(X))]$, which takes an average of the counterfactual outcome under policy π is frequently referred to as the **value** of the policy function π. For example,

▶ The value of a policy that always assigns treatment, i.e. $\pi(X) = 1$, is $\mathbb{E}[Y(\pi(X))] = \mathbb{E}[Y(1)]$

▶ The value of a policy that always assigns control, i.e. $\pi(X) = 0$, is $\mathbb{E}[Y(\pi(X))] = \mathbb{E}[Y(0)]$

Identifying the Value Function

Given any policy π, we can identify its value under assumptions
(A1), (A2), and (A3).

$\mathbb{E}[Y(\pi(X))]$

$=\mathbb{E}[Y(1)I(\pi(X) = 1) + Y(0)I(\pi(X) = 0)]$ Def

$=\mathbb{E}[\mathbb{E}[Y(1)I(\pi(X) = 1) + Y(0)I(\pi(X) = 0) \mid X]]$ Law

$=\mathbb{E}[I(\pi(X) = 1)\mathbb{E}[Y(1) \mid X] + I(\pi(X) = 0)\mathbb{E}[Y(0) \mid X]]$ Pro

$=\mathbb{E}[I(\pi(X) = 1)\mathbb{E}[Y \mid A = 1, X] + I(\pi(X) = 0)\mathbb{E}[Y \mid A = 0, X]]$ (A1

Once we identified the value for any policy with (A1)-(A3), identifying the optimal rule π_{OTR} does not involve more identifying assumptions.

Let $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$. Then, we can rewrite π_{OTR} as

$$
\begin{aligned}
&\pi_{\text{OTR}} \\
=&\operatorname*{argmax}_{\pi} \mathbb{E}[Y(\pi(X))] \\
=&\operatorname*{argmax}_{\pi} \mathbb{E}[I(\pi(X) = 1)\mu_1(X) + I(\pi(X) = 0)\mu_0(X)] \quad \text{From above} \\
=&\operatorname*{argmax}_{\pi} \mathbb{E}[\pi(X)\mu_1(X) + (1 - \pi(X))\mu_0(X)] \quad \text{Because } \pi(X) \\
=&\operatorname*{argmax}_{\pi} \mathbb{E}[\pi(X)(\mu_1(X) - \mu_0(X))] \quad \text{Dropped } \mathbb{E}[\mu_0(\\
=&I(\mu_1(X) - \mu_0(X) \geq 0) \quad \text{See explanation}
\end{aligned}
$$

Identifying the Optimal Treatment Rule (OTR) (i.e., the Optimal Policy) II

In words, the optimal treatment rule for a person with characteristic X is to check whether the expected outcome among those with characteristic X is larger under treatment (i.e. $\mu_1(X)$) or under control (i.e. $\mu_0(X)$).

- ▶ If $\mu_1(X) > \mu_0(X)$, the optimal rule states that the person should be treated.
- ▶ If $\mu_1(X) < \mu_0(X)$, the optimal rule is to assign the control to the person.

For more information on this topic, see NC State's course on Dynamic Treatment Regimes.

Observational Studies and Strong Ignorability I

A large number works in causal inference frame study of causal effects from observational studies as a version of a stratified randomized experiment where assumptions (A1)-(A3) are satisfied.

▶ These works assume that given the measured pre-treatment covariates X, the treatment A can be considered "as-if" random, akin to a stratified randomized experiment.

▶ For more discussions about examining observational studies from the lens of a randomized experiment, see Cochran and Chambers (1965),Rubin (2007), and Small (2024).

▶ A major takeaway from these readings is that investigators should *blind* themselves to the outcome, akin to a randomized experiment where the investigator does not see the outcome during the "design stage" of the experiment.

Observational Studies and Strong Ignorability II

Also, one of the key differences between a stratified randomized experiment (or in general, any randomized experiment) and an observational study is about the *knowledge of the propensity score* $e(x) = \mathbb{P}(A = 1 \mid X = x)$ (i.e., the treatment assignment probability)

- ▶ In a randomized experiment, $e(X)$ is known by the investigator
- ▶ In contrast, in an observational study, $e(X)$ is usually not known since individual's selection into treatment cannot be controlled by the investigator.
- ▶ We'll discuss the implications of this in later lectures on estimation.

Some other interpretations of this approach to studying observational studies are based on the presence (or absence) of **confounding** or **selection based on observables**.

Observational Studies and Strong Ignorability III

▶ We measured all the **confounders** of the treatment-outcome relationship (i.e. X) and these variables satisfy (A2) and (A3) above.

▶ There are **no unmeasured confounders**, denoted as U, that affect the treatment-outcome relationship. A bit more formally, we *do not* have the case where

$$A \perp Y(1), Y(0) | X, U$$

▶ The self-selection into treatment (or control) does not depend on anything except the observables X.

In my opinion, these are strong assumptions for an observational study and in later lectures, we'll discuss identification when strong ignorability does not hold.

Central Role of the Propensity Score $\mathbb{P}(A = 1|X)$ I

Rosenbaum and Rubin (1983) showed that the propensity score $e(X) = \mathbb{P}(A = 1|X) \in (0, 1)$ plays a critical role in identification and estimation of causal effects. Here, we highlight some important properties of the propensity score.

Consider any function $b(X)$ of the covariates. This function b is called a balancing score if conditional on $b(X)$, the treatment is independent of X, i.e.

$$A \perp X | b(X)$$

A couple of remarks:

▶ A trivial function b that satisfies this condition is the identity function $b(X) = X$.
▶ Theorem 1 of Rosenbaum and Rubin (1983) showed that the propensity score $e(X)$ is a balancing score.

Rosenbaum and Rubin (1983) proved two very interesting results about the propensity score.

Theorem 2 of Rosenbaum and Rubin (1983): $b(X)$ is a balancing score if and only if $b(X)$ is finer than the propensity score $e(X)$, i.e. if there exists a function g where $e(X) = g(b(X))$.

Some remarks

▶ The propensity score contains the "smallest'' amount of information to achieve $A \perp X|b(X)$ or **the propensity score is the coarsest balancing score**.

▶ Another way to say this is that the propensity score is the "best" dimension reducing score of X.

▶ To intuitively check Theorem 2, consider setting $b(X) = X$.
  ▶ $b(X) = X$ is not only a balancing score, but also provides much more information (i.e. finer information) than the propensity score $e(X)$, which is a number between 0 and 1.

▶ Also, the function g is simply the propensity score, i.e., we have $e(X) = e(b(X))$ where $g() = e()$.

Theorem 3 of Rosenbaum and Rubin (1983): Let $e(X) = \mathbb{P}(A = 1|X)$. If conditions (A1), (A2), and (A3) hold, we have

$$A \perp Y(1), Y(0)|e(X) \text{ and } 0 < \mathbb{P}(A = 1|e(X)) < 1$$

Technically, Rosenbaum proved this for all balancing scores, not just the propensity score. But, the implications for propensity score is more useful.

▶ Recall that (A1)-(A3) held for the entire X. The above theorem shows that these assumptions also hold for a scalar summary of X in the form of the propensity score $e(X)$.

Central Role of the Propensity Score $\mathbb{P}(A = 1|X)$ IV

▶ If we look at the proof of identification for the ATE, we can identify the ATE via

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, e(X)]] - \mathbb{E}[\mathbb{E}[Y \mid A = 1, e(X)]]$$

  ▶ The proof of this follows directly from the proof of the identification of the ATE where we replace X with $e(X)$.
  ▶ A version of this equality is in Theorem 4 of Rosenbaum and Rubin (1983).

References I

Cochran, William G, and S Paul Chambers. 1965. "The Planning of Observational Studies of Human Populations." *Journal of the Royal Statistical Society. Series A (General)* 128 (2): 234–66.

Greenland, Sander, Judea Pearl, and James M Robins. 1999. "Confounding and Collapsibility in Causal Inference." *Statistical Science* 14 (1): 29–46.

Heckman, James J, Hidehiko Ichimura, and Petra E Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *The Review of Economic Studies* 64 (4): 605–54.

Hernán, Miguel A, David Clayton, and Niels Keiding. 2011. "The Simpson's Paradox Unraveled." *International Journal of Epidemiology* 40 (3): 780–85.

References II

Hernán, Miguel, and James Robins. 2020. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 2007. "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36.

Small, Dylan S. 2024. "Protocols for Observational Studies: Methods and Open Problems." *Statistical Science* 39 (4): 519–54.