Missing data lecture 16: Causal inference foundations

Set up

We're often interested in making causal statements about the processes we study. Suppose we're interested in whether a year-long jobs training program for disadvantaged workers leads to lower rates of unemployment in the year following the training program for these workers. Suppose we had data on post-training program employment records for these workers as well as those of workers with comparable backgrounds to the trainees in the same labor markets with putative access to the training program. This is a simplification of a question was examined in Ashenfelter (1978). We'd probably compare the observed employment status in the year following the training program Y_i for participants to nonparticipants. Let's identify the units $i = 1, \ldots, n$ in the training program as those with $W_i = t$, and those not in the program as those with $W_i = c$. Let the value n_t be the number of people in the training program, and n_c as those who aren't. The simple comparison of mean employment is:

$$\hat{\tau}^{\text{dif}} = \frac{1}{n_t} \sum_{i=1}^{n} Y_i \mathbb{1}\left(W_i = t\right) - \frac{1}{n_c} \sum_{i=1}^{n} Y_i \mathbb{1}\left(W_i = c\right)$$

We'd probably be tempted to ascribe causality to the comparison: `Participation in a jobs program decreased unemployment by 0.05 percentage points", but we've learned that one should only say something like:`Participation in a jobs program predicted a decrease in unemployment of 0.05 percentage points". Why can't we use the causal language we'd prefer to use? When can we do so?

Core principles

Potential outcomes and causal effects

The key idea, shared in Imbens and Rubin (2015) is that the effect of a cause is really a statement about the comparison of two outcomes for a single individual that correspond to different actions taken. In the job training example, there are two actions: participating in

a job training program, and not participating in a job training program, which we denote $W_i \in \{t, c\}$ with c indicating lack of participation. The outcome would be the employment status the year after each action; $Y_i^{W=t}$ would be the employment status after participating in the job training program, while $Y_i^{W=c}$ would be employment status if one did not participate in the jobs program. These variables are called *potential outcomes* because they exist without regards to W_i, the treatment actually chosen by the participants. The causal effect of the action is defined as the difference between the potential outcomes, namely employment status if one had participated in the program and employment status if one had not participated in the program:

$$\tau_i = Y_i^{W=t} - Y_i^{W=c}$$

We can write the different scenarios for each outcome to get the values for τ_i:

τ_i	$Y_i^{W=t}$	$Y_i^{W=c}$	Description
0	0	0	Always unemployed, no causal effect
1	1	0	Training led to employment
-1	0	1	Training led to unemployment
0	1	1	Always employed, no causal effect

Note that these causal effects are not dependent on which treatment was actually chosen because the effect compares two variables that don't depend on the treatment chosen. The idea of potential outcomes is called the Rubin Causal Model (RCM). *Counterfactual outcome* is another term for these variables. In the RCM, these are considered fixed for each individual.

Stable Unit Treatment Value Assumption

Crucially, this individual-level causal effect is also not estimable, because we only get to see one outcome, namely the outcome chosen by the individual. The data we would see for a single set of individuals is

W_i	$Y_i^{W=t}$	$Y_i^{W=c}$	$Y_{i(0)}$
0	?	0	0
1	0	?	0
1	1	?	1
0	?	1	1

This table encodes two assumptions that are present in many causal inference problems: consistency, and lack of interference. These are combined into the Stable Unit Treatment Value Assumption (SUTVA):

Assumption 1. SUTVA

1. There are not different forms or versions of treatments that lead to different potential outcomes.
2. The potential outcomes for each unit do not depend on treatments assigned to other units.

The first, consistency, means that we assume the following:

$$P(Y_{i(0)} = Y_i^{W=t} \mid W_i = t) = 1,\ P(Y_{i(0)} = Y_i^{W=c} \mid W_i = c) = 1$$

This has several implications, all of which are important to chew through:

1. We assume that everyone receives the same type of treatment in the same way. This would be violated for instance if some job trainees received online instruction while others received in-person classes.
2. We assume that the act of treatment does not change the observation. This is a well-known problem in psychology studies, namely that when people are aware that they are being watched they will change their behavior, also known as the Hawthorne effect.

The second assumption is the no-interference assumption, which states that the potential outcomes for a single unit do not depend on treatment assignment of other units. This is the sort of assumption that would fail in a vaccine trial that measured, say, symptomatic disease caused by an infectious pathogen in households. In this setting, a vaccine may make participants less infectious, thus a person's disease status would depend on the vaccination status of those around them.

Both of these assumptions make it possible to define an individual's observed outcome, $Y_{i(0)}$ as a function of their treatment assignment W_i and their potential outcomes alone:

$$Y_{i(0)} = \mathbb{1}\left(W_i = t\right) Y_i^{W=t} + \mathbb{1}\left(W_i = c\right) Y_i^{W=c}$$

This implies that each unit is missing an observation, which we'll denote as $Y_{i(1)}$, in keeping with our missing data notation:

$$Y_{i(1)} = \mathbb{1}\left(W_i = c\right) Y_i^{W=t} + \mathbb{1}\left(W_i = t\right) Y_i^{W=c}$$

Causal inference as a missing data problem

We can invert these relationships to make the missingness more explicit:

$$Y_i^{W=t} \begin{cases} Y_{i(0)} & \text{if } W_i = t \\ Y_{i(1)} & \text{if } W_i = c \end{cases}, \quad Y_i^{W=c} \begin{cases} Y_{i(1)} & \text{if } W_i = t \\ Y_{i(0)} & \text{if } W_i = c \end{cases}$$

Thus, if we wanted to predict individual causal effects, we would need to impute the missing observation depending on which treatment group they have been assigned to.

Missingness mechanisms

We know from our missing data material that if we want to make an inference about model parameters under missing data, we need to be in the MCAR or MAR setting. We can think of W_i as the causal version of M_i. In the missing data realm, we would have M_{it} and M_{ic}, with 1s representing missingness of Y_{it} or Y_{ic}, respectively. These variables give too many degrees of freedom for causal inference, however, because they are constrained so that $M_{it} + M_{ic} = 1$ for every i. Thus, we can write our missing data mechanism as a mechanism for $\mathbf{W}$, the set of assignments for all n units under study. Let $Y_{(0)}$ be the observed data for all participants, and $Y_{(1)}$ be the missing values for all units. MCAR would be the following, with a slight change of notation:

$$P(\mathbf{W} = \tilde{w} \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}) = P(\mathbf{W} = \tilde{w} \mid Y_{(0)} = y^{\star}_{(0)}, Y_{(1)} = y^{\star}_{(1)}), \forall y_{(0)}, y_{(1)}, y^{\star}_{(0)}, y^{\star}_{(1)}$$

While MACAR would be:

$$P(\mathbf{W} = w \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}) = P(\mathbf{W} = w \mid Y_{(0)} = y^{\star}_{(0)}, Y_{(1)} = y^{\star}_{(1)}), \forall w, y_{(0)}, y_{(1)}, y^{\star}_{(0)}, y^{\star}_{(1)}$$

One way to ensure this is to randomly assign units to treatment. If units are not assigned randomly to treatment, it may be that we can consider covariates for every individual, arranges into a $n \times p$ matrix $\mathbf{X}$,

$$P(\mathbf{W} = \tilde{w} \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, \mathbf{X}) = P(\mathbf{W} = \tilde{w} \mid Y_{(0)} = y^{\star}_{(0)}, Y_{(1)} = y^{\star}_{(1)}, \mathbf{X}), \forall w, y_{(0)}, y_{(1)}, y^{\star}_{(0)}, y^{\star}_{(1)}$$

However, if we don't control the treatment assignment, like in the jobs training program, we likely have MNAR missingness:

$$P(\mathbf{W} = \tilde{w} \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \mathbf{X}) \neq P(\mathbf{W} = \tilde{w} \mid Y_{(0)} = \tilde{y}, Y_{(1)} = y^{\star}_{(1)}, \mathbf{X}) \text{ for some } y_{(1)} \neq y^{\star}_{(1)}$$

For instance, given that the job training program is a one-year program, people who decide to select into the program might have fewer employment prospects than the people who don't select into the program.

Causal Estimands

Typically, we want to define an estimable quantity that is related to the population of interest. In causal inference, the population of interest tends to be the sample at hand, which is assumed to arise from an infinite superpopulation. In causal inference, we know that causal effects are comparisons of unit-level potential outcomes, so averages of these values would make for natural causal estimands.

One would be be the finite-sample mean of the individual causal effects:

$$\tau_{\text{fs}} = \frac{1}{n}\sum_{i=1}^{n}(Y_i^{W=t} - Y_i^{W=c})$$

We can write the contrast above as the means over the two groups:

$$\begin{aligned}
\hat{\tau}^{\text{dif}} &= \frac{1}{n_t}\sum_{i=1}^{n}\mathbb{1}\left(W_i = t\right)Y_{i(0)} - \frac{1}{n_c}\sum_{i=1}^{n}\mathbb{1}\left(W_i = c\right)Y_{i(0)} \\
&= \frac{1}{n_t}\sum_{i=1}^{n}\mathbb{1}\left(W_i = t\right)Y_i^{W=t} - \frac{1}{n_c}\sum_{i=1}^{n}\mathbb{1}\left(W_i = c\right)Y_i^{W=c} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\left(W_i = t\right)Y_i^{W=t}}{n_t/n} - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}\left(W_i = c\right)Y_i^{W=c}}{n_c/n}
\end{aligned}$$

If we were under the MACAR scenario, we could take the expectation of both sides with respect to $W_i \mid Y_{i(0)}, Y_{i(1)}$ to get

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}^{\text{dif}} \mid Y_{i(0)}, Y_{i(1)}\right] &= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left[\mathbb{1}\left(W_i = t\right) \mid Y_{i(0)}, Y_{i(1)}\right]Y_i^{W=t}}{n_t/n} - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left[\mathbb{1}\left(W_i = c\right) \mid Y_{(0)}, Y_{(1)}\right]Y_i^{W=c}}{n_c/n} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left[\mathbb{1}\left(W_i = t\right)\right]Y_i^{W=t}}{n_t/n} - \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{E}\left[\mathbb{1}\left(W_i = c\right)\right]Y_i^{W=c}}{n_c/n} \\
&= \frac{1}{n}\sum_{i=1}^{n}Y_i^{W=t} - \frac{1}{n}\sum_{i=1}^{n}Y_i^{W=c} \\
&= \tau_{\text{fs}}
\end{aligned}$$

The second line came from our MACAR assumption, which we could satisfy in a randomized experiment.

Assignment mechanisms

As we showed above, we can write $Y_{i(0)}$ and $Y_{i(1)}$ in terms of the assignment W_i and the potential outcomes $Y_i^{W=t}$ and $Y_i^{W=c}Y_i^{W=t}$, so we can write

$$P(\mathbf{W} = w \mid Y_{(0)} = \tilde{y}_{(0)}, Y_{(1)} = y_{(1)}, \mathbf{X} = x)$$

as

$$P(\mathbf{W} = w \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x)$$

where we let $Y^{W=t}$ represent the n-vector of all unit potential outcomes under treatment and $Y^{W=c}$ is the n-vector of all unit potential outcomes under control, and the vectors y_t and y_c are dummy vectors in the space of binary n-vectors, or $\{0,1\}^n$.

This probability distribution over vectors $w \in \{t, c\}^n$ is called the *assignment mechanism*. It sums to one over all possible assignments:

$$\sum_{w \in \{t,c\}^n} P(\mathbf{W} = w \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x) = 1$$

We can write the probability that $W_i = t$

$$P(W_i = t \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x) = \sum_{w \in \{t,c\}^n | w_i = t} P(\mathbf{W} = w \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x) =$$

We can define something called the *propensity score*, which is the probability of treatment for someone with a covariate $X_i = x$

$$e(x) = \frac{1}{N(x)} \sum_{i | X_i = x} P(W_i = t \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x)$$

where $N(x)$ is the number of units with $X_i = x$. If $N(x) = 0$ we set $e(x) = 0$.

A typical assignment mechanism restriction is that of individualistic assignment, where

$$P(W_i = t \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x) = P(W_i = t \mid Y_i^{W=t} = y_{it}, Y_i^{W=c} = y_{ic}, \mathbf{X}_i = x_i)$$

Probabilistic assignment is an individual assignment mechanism that is in the interval $(0, 1)$:

$$0 < P(W_i = t \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x) < 1$$

Finally, we can define the most important assignment mechanism restriction, which is called unconfoundedness:

Definition 1. Unconfounded Assignment

$$P(\mathbf{W} = w \mid Y^{W=t} = y_t, Y^{W=c} = y_c, \mathbf{X} = x) = P(\mathbf{W} = w \mid Y^{W=t} = y_t^\star, Y^{W=c} = y_c^\star, \mathbf{X} = x)$$

for all $w, x, y_t, y_c, y_t^\star, y_c^\star$.

Note that this definition is just MACAR, but for treatment assignment.

We can relax this assumption in the following way, as shown in Rubin (1978):

Definition 2. Ignorable Treatment Assignment

$$P(\mathbf{W} = w \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}, \mathbf{X} = x) = P(\mathbf{W} = w \mid Y_{(0)} = y_{(0)}, Y_{(1)} = y_{(1)}^\star, \mathbf{X} = x)$$

for all $y_{(1)}, y_{(1)}^\star$.

Note that this is a definition akin to MAR for missingness, where we moved back to missing data notation to highlight the potential dependence on observed outcomes but not unobserved outcomes.

Ashenfelter, Orley. 1978. "Estimating the Effect of Training Programs on Earnings." *The Review of Economics and Statistics*, 47–57.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge university press.

Rubin, Donald B. 1978. "Bayesian Inference for Causal Effects: The Role of Randomization." *The Annals of Statistics* 6 (1). https://doi.org/10.1214/aos/1176344064.