Causality

Chen Zhou

August 28, 2016

Abstract

This note reviewed the concept of causality stepping from its definition to its connection with regression methods. The definition of causality is based on *potential outcomes* here. In a random experiment, once the stable unit treatment value assumption is satisfied, the treatment effects can be evaluated by causal estimands.

1 Potential Outcomes

Causality[1] is tied to an *action* applied to a *unit*. A unit here can be a physical object, a firm, an individual person, or collection of objects or persons, at a particular point in time[2]. For instance, when deciding to take an aspirin to relieve your headache, you could also have chosen not to take the aspirin.

Given a unit and a set of actions, we associate each action-unit pair with a *potential outcome*. For one unit, there are must be only one action can be performed, therefore outcomes associated with other actions are never observed. That's why we name it "potential".

To convey a causal inference, we first must identify the potential outcome, otherwise that causal statement is ill-defined.

2 Definition of Causal Effects

Based on the definition of potential outcome, the definition of causal effects can be derived. Consider a single unit, I contemplating whether or not to take an aspirin for my headache. The headache may be gone or remain after I take the aspirin. We denote either outcome by $Y(\text{Aspirin})$. Similarly, if I do not take the aspirin, my headache may be gone or remain. We denote this potential outcome by $Y(\text{No Aspirin})$, which can be either "Headache" or "No Headache".

All possibilities can be organized as

1. Headache gone only with aspirin:

$$Y(\text{Aspirin}) = \text{ No Headache}, \quad Y(\text{No Aspirin}) = \text{ Headache}$$

2. No effect of aspirin, with a headache in both class:

$$Y(\text{Aspirin}) = \text{ Headache}, \quad Y(\text{No Aspirin}) = \text{ Headache}$$

[1] The section of note is adapted from Imbens and Rubin (2015, Chapter 1)

[2] The same subject but at another time is presumed to be different.

3. No effect of aspirin, with the headache gone in both cases:

$$Y(\text{Aspirin}) = \text{ No Headache}, \quad Y(\text{No Aspirin}) = \text{ No Headache}$$

4. Headache gone only without aspirin:

$$Y(\text{Aspirin}) = \text{ Headache}, \quad Y(\text{No Aspirin}) = \text{ Headache}$$

There are two important aspects of this definition of causal effect. First, the definition depends only on the potential outcomes, but it does *not* depend on which outcome is actually observed[3]. Second, the causal effect is the comparison of potential outcomes, for the same unit, at the same moment in time post-treatment. Simple before-after comparison is not causal effect.

The fundamental problem of causal inference is therefore the problem that at most one of the potential outcomes can be realized and thus observed. For the estimation of causal effect, we need to compare observed outcomes, and because there is only one realized potential outcome per unit, we will need to consider multiple units.

3 The Stable Unit Treatment Value Assumption

We must observe multiple units, some exposed to the active treatment, some exposed to the alternative (control) treatment. The treatments are subject to a set of assumptions. The first is SUTVA.

> The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

No interference The no-inference part of SUTVA is that the treatment applied to one unit does not affect the outcome for other units. There exist settings in which the no-inference part of SUTVA is controversial. Economists refer the following as *general equilibrium* effects.

- Job training programs. The outcomes for one individual may well be affected by the number of people trained when that number is sufficiently large to create increased competition for certain jobs.
- The causal effect of your immunization versus no immunization will surely depend on the immunization of others: if everybody else is already immunized with a perfect vaccine, and others can therefore neither get the disease nor transmit it.

No hidden variations of treatments The second component of SUTVA requires that an individual receiving a specific treatment level cannot receive different forms of that treatment. The causal effect of aspirin requires that one of the aspirin tablets is old and no longer contains a fully effective dose.

[3] Whether taking aspirin or not, out conclusion on causal effect is not violated.

SUTVA also requires there is no difference in the method of administering the treatment matter. The effect of taking a drug for a particular individual may differ depending on whether the individual was assigned to receive it or chose to take it.

4 Causal Estimands

We start with a population of units, indexed by $i = 1, \dots, N$. Each unit in this population can be exposed to one of a set of treatments. Let $\mathbb{T}_i$ denotes the set of treatments

$$\mathbb{T}_i = \mathbb{T} = \{0, 1\}$$

where 0 means control treatment, 1 means active treatment.

Unit-level causal effects For each unit i, and for each treatment in the common set of treatments, $\mathbb{T} = \{0, 1\}$, there are corresponding potential outcome, $Y_i(0)$ and $Y_i(1)$. The *unit-level causal effects* is the comparisons of $Y_i(1)$ and $Y_i(0)$

$$Y_i(1) - Y_i(0), \text{ or } \frac{Y_i(1)}{Y_i(0)}.$$

Summarizing such unit-level causal effects for the finite sample or for subpopulations, we get a *causal estimand* of average difference of potential outcomes,

$$\tau_{\text{fs}} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i(1) - Y_i(0)\right),$$

where the subscript “fs” indicates that we average over the finite sample[4].

Average over subpopulation The subpopulation may be defined in terms of different sets of variables.

- In terms of pretreatment variables, or covariates,

  $$\tau_{\text{fs}} = \frac{1}{N(f)} \sum_{i:X_i=f} (Y_i(1) - Y_i(0)),$$

  Here $X_i \in f, m$ is an indicator for being in a subpopulation (e.g. female), and $N(f) = \sum_{i=1}^{N} \mathbf{1}_{X_i=f}$ is the number of units in that subpopulation.

- In terms of treatment,

  $$\tau_{\text{fs},t} = \frac{1}{N_t} \sum_{i:W_i=1} \left(Y_i(1) - Y_i(0)\right),$$

  where N_t is the number of units exposed to the active treatment.

[4] Why emphasize this?

- In terms of potential outcomes,

$$\tau_{\text{fs,pos}} = \frac{1}{N_{\text{pos}}} \sum_{i:Y_i(0)>0,Y_i(1)>0} \left(Y_i(1) - Y_i(0)\right),$$

  where $N_{\text{pos}} = \sum_{i=1}^{N} \mathbf{1}_{Y_i(0)>0,Y_i(1)>0}$. For example, one may be interested in the average effect of a job-training program on hourly wages, averaged on only over those individuals who would have been employed.

More general functions of potential outcomes The second generalization of the average treatment effect—general functions of potential outcomes. For example,

- We may be interested in the median of $Y_i(1)$ versus the median of $Y_i(0)$.
- One may also be interested in the median of the difference $Y_i(1) - Y_i(0)$.

In all cases with $\mathbb{T} = \{0, 1\}$, we can write the causal estimand as a row-exchangeable function of all potential outcomes for all units, all treatments, and pretreatment variables:

$$\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W}).$$

The $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ are the N-component column vectors of potential outcomes with ith elements equal to $Y_i(0)$ and $Y_i(1)$, $\mathbf{W}$ is the N-component column vector of treatment assignments, and X is the $N \times K$ matrix of covariates.

5 Regression methods

5.1 Linear Regression with No Covariates

We specify a linear regression function for the observed outcome Y_i^{obs} as

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + \varepsilon_i,$$

where W_i is the indicator for the receipt of treatment, ε_i captures unobserved determinants of the outcome.

The ordinary least squares (OLS) estimator for τ is based on minimizing the sum of squared residuals over α and τ,

$$(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}) = \arg\min_{\tau,\alpha} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2,$$

with solutions

$$\hat{\tau}^{\text{obs}} = \frac{\sum_{i=1}^{N}(W_i - \bar{W}) \cdot (Y_i^{\text{obs}} - \bar{Y}^{\text{obs}})}{\sum_{i=1}^{N}\left(W_i - \bar{W}\right)^2,} \text{ and } \hat{\alpha}^{\text{ols}} = \bar{Y}^{\text{obs}} - \hat{\tau}^{\text{ols}} \cdot \bar{W},$$

where

$$\bar{Y}^{\text{obs}} = \frac{1}{N} \sum_{i=1}^{N} Y_i^{\text{obs}} \text{ and } \bar{W} = \frac{1}{N} \sum_{i=1}^{N} W_i = \frac{N}{N}.$$

The OLS estimator $\hat{\tau}^{\text{ols}}$ is identical to the difference if average outcomes by treatment status

$$\hat{\tau}^{\text{ols}} = \bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}} = \hat{\tau}^{\text{dif}},$$

where $\hat{Y}_t^{\text{obs}}$ and $\hat{Y}_c^{\text{obs}}$ are the averages of the observed outcomes in the treatment and control groups respectively.

The least squares estimate of τ is the causal effect of the treatment in randomized experiments, if The residuals ε_i are independent of, or at least uncorrelated with, the treatment indicator W_i. In current context, $\hat{\tau}^{\text{ols}}$ is identical to $\bar{Y}_t^{\text{obs}} - \bar{Y}_c^{\text{obs}}$, which is unbiased for the finite-sample average treatment effect as well as for the super-population average treatment effect (Imbens and Rubin, 2015, Chapter 6).

5.2 The causal interpretation of $\hat{\tau}^{\text{ols}}$

Let α be the population average outcome under the control, $\alpha = \mu_c = \mathbb{E}_{\text{sp}}[Y_i(0)]$, τ_{sp} is the super-population average treatment effect, $\tau_{\text{sp}} = \mu_t - \mu_c = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)]$. Now *define* the residual ε_i in terms of the population parameters, treatment indicator, and the potential outcomes as

$$\varepsilon_i = Y_i(0) - \alpha + W_i \cdot \big(Y_i(1) - Y_i(0) - \tau_{\text{sp}}\big) = \begin{cases} Y_i^{\text{obs}} - \alpha, & \text{if } W_i = 0 \\ Y_i^{\text{obs}} - \alpha - \tau_{\text{sp}}, & \text{if } W_i = 1. \end{cases}$$

Then we can write

$$\varepsilon_i = Y_i^{\text{obs}} - (\alpha + \tau_{\text{sp}} \cdot W_i),$$

and thus we can write the observed outcome as

$$Y_i^{\text{obs}} = \alpha + \tau_{\text{sp}} \cdot W_i + \varepsilon_i.$$

In a randomized setting, the potential outcomes are random variables. If the assignment is random, which implies that assignment is independent of the potential outcomes

$$\Pr\big(W_i = 1 \mid Y_i(0), Y_i(1)\big) = \Pr(W_i = 1),$$

or

$$W_i \perp\!\!\!\perp (Y_i(0), Y_i(1)).$$

The definition of the residual, in combination with random assignment and random sampling from a super-population, implies that the residual has mean zero conditional on the treatment indicator in the population:

$$\mathbb{E}_{\text{sp}}[\varepsilon_i \mid W_i = 0] = \mathbb{E}_{\text{sp}}[Y_i(0) - \alpha \mid W_i = 0] = \mathbb{E}_{\text{sp}}[Y_i(0) - \alpha] = 0,$$

and

$$\begin{aligned} \mathbb{E}_{\text{sp}}[\varepsilon_i \mid W_i = 1] &= \mathbb{E}_{\text{sp}}[Y_i(1) - \alpha - \tau_{\text{sp}} \mid W_i = 1] \\ &= \mathbb{E}_{\text{sp}}[Y_i(1) - \alpha - \tau_{\text{sp}} \mid W_i = 1] \\ &= 0, \end{aligned}$$

so that

$$\mathbb{E}_{\text{sp}}[\varepsilon_i \mid W_i = w] = 0,$$

for $w = 0, 1$.

The conditional mean of ε_i given W_i is zero, so the least squares estimator, $\hat{\tau}^{\text{ols}}$, is unbiased for τ_{sp}. The assumptions about residuals in least squares analyses actually follow from random sampling and random assignment.

Another way of deriving this result. Define the estimators as

$$(\hat{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}}) = \arg\min_{\alpha,\tau} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2.$$

Theses estimators converge to the population limits (α^*, τ^*) that minimize the expected value of the sum of squares

$$(\alpha^*, \tau^*) = \arg\min_{\alpha,\tau} \mathbb{E}_{\text{sp}} \left[\frac{1}{N} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i\right)^2 \right]$$

This implies that the population limit is

$$\tau^* = \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}} \mid W_i = 1] - \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}} \mid W_i] = 0.$$

Random assignment of W_i implies

$$\mathbb{E}_{\text{sp}}[Y_i^{\text{obs}} \mid W_i = 1] - \mathbb{E}_{\text{sp}}[Y_i^{\text{obs}} \mid W_i = 0] = \mathbb{E}_{\text{sp}}[Y_i(1) - Y_i(0)] = \tau_{\text{sp}},$$

so that the population limit of the least squares estimator is equal to the population average treatment effect, $\tau^* = \tau_{\text{sp}}$.

For the time being, this note is leaving the inference (sampling variance and confidence intervals) of the least squares methods to the future work. The distinctions between homoskedasiticity and heteroskedasticity in the sampling variance will be mentioned here.

5.3 Linear Regression with Additional Covariates

The regression function is specified as

$$Y_i^{\text{obs}} = \alpha + \tau \cdot W_i + X_i\beta + \varepsilon,$$

where X_i is a row vector of covariates (i.e., pretreatment variables). We estimate the regression coefficients using least squares

$$(\hat{\tau}^{\text{ols}}, \hat{\alpha}^{\text{ols}}, \hat{\beta}^{\text{ols}}) = \arg\min_{\tau,\alpha,\beta} \sum_{i=1}^{N} \left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta\right)^2.$$

Under some regularity conditions, when the sample gets large, $(\hat{\alpha}^{\text{ols}}, \hat{\tau}^{\text{ols}}, \hat{\beta}^{\text{ols}})$ converge to $(\alpha^*, \tau^*, \beta^*)$, defined as

$$(\alpha^*, \tau^*, \beta^*) = \arg\min_{\alpha,\beta,\tau} \mathbb{E}\left[\left(Y_i^{\text{obs}} - \alpha - \tau \cdot W_i - X_i\beta\right)\right].$$

In this case with additional predictors, it is no longer true that $\hat{\tau}^{\text{ols}}$ is unbiased for τ_{sp} in finite samples. However, irrespective of whether the regression function is truly linear in the covariates in the population, the least squares estimate $\hat{\tau}^{\text{ols}}$ is unbiased in large samples for the population average treatment effect, τ_{sp}.

No matter how non-linear the conditional expectations of the potential outcomes given the covariates are in the super-population, simple least square regression is consistent for estimating the population average treatment effect.

By randomizing treatment assignment, the super-population correlation between the treatment indicator and the covariates is zero. Even though in finite samples the actual correlation may differ from zero, in large samples this correlation will vanish, and as a result the inclusion of the covariates does not matter for the limiting values of the estimator.

There are more related topics:

- linear regression with covariates and interactions;
- transformations of the outcome variable;
- the limits on increases in precision due to covariates;
- testing for the presence of treatment effects.

For more content, please refer to Imbens and Rubin (2015, Chapter 7).

6 Model-based Inference for Completely Randomized Experiments

There are four approaches to causal inference,

- Fisher's exact p-value approach,
- Neyman's repeated sampling approach,
- regression methods,
- and model based approach.

The model-based approach is very flexible compared to others. It can easily accommodate a wide variety of estimands—we may be interested not only in average treatment effects but also in quantiles, or in measures of dispersion of the distributions of potential outcomes. We can conduct causal inference for any causal estimand

$$\tau = \tau\big(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}, \mathbf{W}\big).$$

The model-based approach can be extended to observational studies, where the assignment mechanism is (partially) unknown. Although the resulting inference may be more sensitive to the modeling assumptions, and thus less credible than in randomized experiments, the basic approach is the same as in classical randomized experiments.

The potential outcomes, and thus the causal estimands, are well defined irrespective of the stochastic model for either the treatment assignment. In many cases, at least in large samples, estimates for the average treatment effect are unbiased from Neyman's repeated sampling perspective.

7 Bayesian Model-Based Imputation

The primary goal of this approach is to build a model for the missing potential outcomes, given the observed data

$$f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}).$$

With this model, we can derive the distribution for the estimand of interest

$$\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}),$$

using the estimand in terms of observed and missing potential outcomes as

$$\tau = \tau(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}).$$

The conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ depends intricately on the joint distribution of the potential outcomes, $\big(\mathbf{Y}(0), \mathbf{Y}(1)\big)$, and on the assignment mechanism. The former requires scientific knowledge, the assignment mechanism here is assumed to be a completely randomized experiment.

The joint distribution of two potential outcomes $\big(\mathbf{Y}(0), \mathbf{Y}(1)\big)$ is

$$f(\mathbf{Y}(0), \mathbf{Y}(1)).$$

Under row (unit) exchangeablility of the matrix $\big(\mathbf{Y}(0), \mathbf{Y}(1)\big)$, and by an appeal to Finetti's theorem, we can model this join distribution as the integral over the product of iid unit-level distributions

$$f(\mathbf{Y}(0), \mathbf{Y}(1)) = \int \prod_{i=1}^{N} f(Y_i(0), Y_i(1) \mid \theta) \cdot p(\theta) d\theta,$$

where θ is an unknown, finite-dimensional parameter of $f(Y_i(0), Y_i(1) \mid \theta)$, which lies in a parameter space Θ, and $p(\theta)$ is its marginal distribution. The unknown parameters θ is specified by subject-matter knowledge.

Specifying the second input, the prior distribution of θ, $p(\theta)$.

7.1 The Four Steps of the Bayesian Approach to Model-Based Inference

1. Deriving $f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta)$.

2. Posterior distribution for the parameter θ, $f(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$.

3. Combing the conditional distribution $f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta)$, and the posterior distribution $f(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$ to obtain the conditional distribution of the missing data given the observed data.

4. Using the estimand $\tau = \tau\big(\mathbf{Y}(0), \mathbf{Y}(1)\big)$, and the conditional distribution $f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$ to obtain the conditional distribution of the estimand given the observed values, $f(\tau \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$.

7.1.1 The First Step

Combine the conditional distribution of the vector of assignments given the potential outcomes

$$\Pr(\mathbf{W} \mid \mathbf{Y}(0), \mathbf{Y}(1)),$$

with the model for the joint distribution of the potential outcomes given θ,

$$f(\mathbf{Y}(0), \mathbf{Y}(1) \mid \theta),$$

to get the joint distribution[5] of $\big(\mathbf{W}, \mathbf{Y}(0), \mathbf{Y}(1)\big)$ given θ,

$$f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W} \mid \theta) = \Pr(\mathbf{W} \mid \mathbf{Y}(0), \mathbf{Y}(1), \theta) \cdot f(\mathbf{Y}(0), \mathbf{Y}(1) \mid \theta).$$

Derive the conditional distribution of the potential outcomes given the vector of assignments and the parameter θ

$$f(\mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{W}, \theta) = \frac{f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W} \mid \theta)}{\Pr(\mathbf{W} \mid \theta)} = \frac{f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W} \mid \theta)}{\int f(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W} \mid \theta)\, \mathrm{d}\mathbf{Y}(0)\, \mathrm{d}\mathbf{Y}(1)}.$$

In the context of randomized experiment, $\mathbf{W}$ is independent of $(\mathbf{Y}(0), \mathbf{Y}(1))$, so

$$f(\mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{W}, \theta) = f(\mathbf{Y}(0), \mathbf{Y}(1) \mid \theta).$$

Transform the distribution of $\mathbf{Y}(0)$ and $\mathbf{Y}(1)$ given $\mathbf{W}$ and θ into distribution for $\mathbf{Y}^{\text{mis}}$ given $\mathbf{Y}^{\text{obs}}, \mathbf{W}$, and θ. Recall that we can express the pair $(Y_i^{\text{mis}}, Y_i^{\text{obs}})$ as functions of $(Y_i(0), Y_i(1), W_i)$

$$Y_i^{\text{obs}} = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1, \end{cases} \text{ and } Y_i^{\text{mis}} = \begin{cases} Y_i(0) & \text{if } W_i = 1, \\ Y_i(1) & \text{if } W_i = 0. \end{cases}$$

Hence we can write $(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}})$ as a transformation of $(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$ like

$$(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}) = g(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W}).$$

Thus we can derive

$$f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta) = \frac{f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}} \mid \mathbf{W}, \theta)}{f(\mathbf{Y}^{\text{obs}} \mid \mathbf{W}, \theta)} = \frac{f(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}} \mid \mathbf{W}, \theta)}{\int_{\mathbf{y}^{\text{mis}}} f(\mathbf{y}^{\text{mis}}, \mathbf{Y}^{\text{obs}} \mid \mathbf{W}, \theta)\, \mathrm{d}\mathbf{y}^{\text{mis}}}.$$

7.1.2 The Second Step

Combine the prior distribution on θ, $p(\theta)$, with the distribution of the observed data given θ to derive the posterior distribution of θ, $p(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$. The likelihood function is

$$\mathcal{L}(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}) \equiv f(\mathbf{Y}^{\text{obs}}, \mathbf{W} \mid \theta) = \int_{\mathbf{y}^{\text{mis}}} f(\mathbf{y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W} \mid \theta)\, \mathrm{d}\mathbf{y}^{\text{mis}}.$$

[5] The joint probability density function $f_{X,Y}$ for continuous random variables is equal to

$$f_{X,Y}(x, y) = f_{Y|X}(y \mid x) f_X(x) = f_{X|Y}(x \mid y) f_Y(y).$$

Combine the likelihood function with the prior distribution $p(\theta)$, we obtain the posterior distribution of the parameters

$$\begin{aligned} p(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}) &= \frac{p(\theta) \cdot \mathcal{L}(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})}{f(\mathbf{Y}^{\text{obs}}, \mathbf{W})} \\ &= \frac{p(\theta) \cdot f(\mathbf{Y}^{\text{obs}}, \mathbf{W} \mid \theta)}{f(\mathbf{Y}^{\text{obs}}, \mathbf{W})}, \end{aligned}$$

where

$$f(\mathbf{Y}^{\text{obs}}, \mathbf{W}) = \int_\theta p(\theta) \cdot \mathcal{L}(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}) \, \mathrm{d}\theta$$

is the marginal distribution of $(\mathbf{Y}, \mathbf{W})$.

7.1.3 The Third Step

Combine the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta)$ and the posterior distribution of θ to derive the joint distribution of $(\mathbf{Y}^{\text{mis}}, \theta)$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$

$$f(\mathbf{Y}^{\text{mis}}, \theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}) = f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta) \cdot p(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}).$$

Integrate over θ to derive the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$

$$f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}) = \int_\theta f(\mathbf{Y}^{\text{mis}}, \theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}) \, \mathrm{d}\theta.$$

7.1.4 The Final Step

We use the conditional distribution of the missing data given the observed data $f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$ and the observed data $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$ to obtain the distribution of the estimand of interest given the observed data.

The general form of the estimand is $\tau = \tau(\mathbf{Y}(0), \mathbf{Y}(1), \mathbf{W})$. We write τ in terms of observed and missing potential outcomes and the treatment assignment

$$(\mathbf{Y}(0), \mathbf{Y}(1)) = h(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}).$$

Thus, we get

$$\tilde{\tau}(\mathbf{Y}^{\text{mis}}, \mathbf{Y}^{\text{obs}}, \mathbf{W}).$$

Combined with the conditional distribution of $\mathbf{Y}^{\text{mis}}$ given $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$, we derive the conditional distributional distribution of τ given the observed data $(\mathbf{Y}^{\text{obs}}, \mathbf{W})$, that is the posterior distribution of τ:

$$f(\tau \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}).$$

We can derive the posterior mean, standard deviation and any other feature of the posterior distribution of the causal estimand.

7.1.5 Summary

It is complicate to model the missing potential outcomes given the observed data with three know inputs:

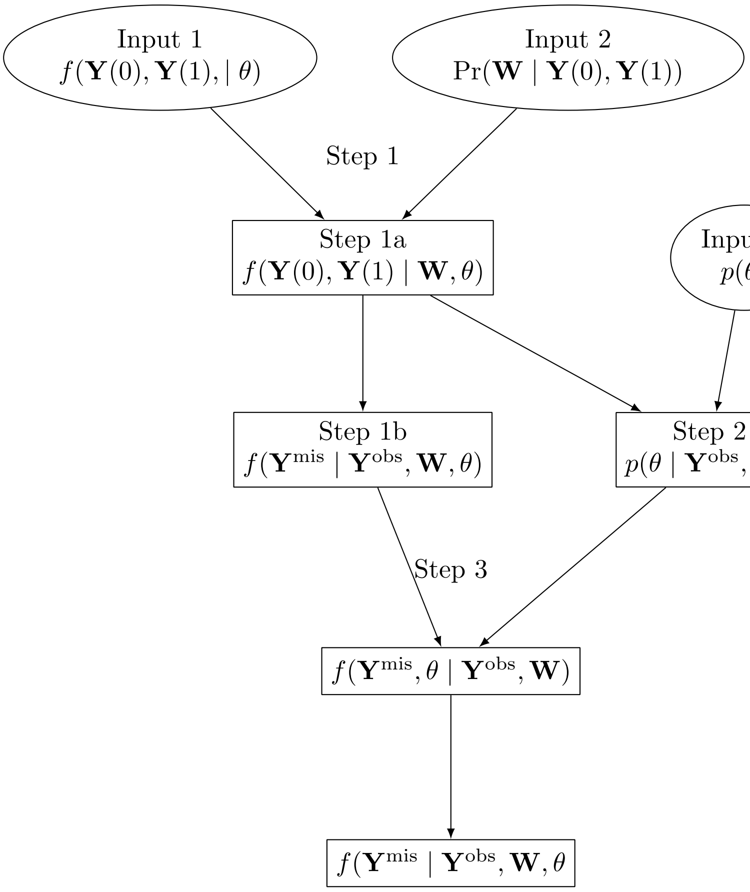
Input 1
$f(\mathbf{Y}(0), \mathbf{Y}(1), \mid \theta)$
Input 2
$\Pr(\mathbf{W} \mid \mathbf{Y}(0), \mathbf{Y}(1))$
Step 1
Step 1a
$f(\mathbf{Y}(0), \mathbf{Y}(1) \mid \mathbf{W}, \theta)$
Input 3
$p(\theta)$
Step 1b
$f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta)$
Step 2
$p(\theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$
Step 3
$f(\mathbf{Y}^{\text{mis}}, \theta \mid \mathbf{Y}^{\text{obs}}, \mathbf{W})$
$f(\mathbf{Y}^{\text{mis}} \mid \mathbf{Y}^{\text{obs}}, \mathbf{W}, \theta$

Figure 1: Model-based approach

- conditional distribution of the potential outcomes given θ

$$f(\mathbf{Y}(0), \mathbf{Y}(1) \mid \theta);$$

- the prior distribution of θ

$$p(\theta);$$

- the assignment mechanism

$$\Pr(\mathbf{W} \mid \mathbf{Y}(0), \mathbf{Y}(1)).$$

Of the four steps to derive the posterior distribution of causal estimand, the fundamental one is the third step, which derives the posterior distribution of missing potential outcomes. The two steps before it prepares the intermediates. In Figure 1, we illustrate the whole procedure.

[5] A randomized experiment does not use this input.

References

Imbens, Guido and Donald B. Rubin (2015). *Causal inference for statistics, social, and biomedical sciences: an introduction*. New York: Cambridge University Press. 625 pp. ISBN: 978-0-521-88588-1.