Reading Notes of *Causal Inference*[*]

Gao Fangshu

2019.5.7

Causal inference without models

1 A definition of causal effect

1.1 Individual causal effects

Causal effect for individual i:

$$Y_i^{a=1} \neq Y_i^{a=0}$$

the treatment A has a causal effect on an individual's outcome Y if $Y^{a=1} \neq Y^{a=0}$ for the individual.

That is, for individual i, $Y^{a=1}$ (is a random variable, read Y under treatment $a = 1$), the outcome variable that would have been observed under the treatment value $a = 1$ is not equal to the outcome variable that would have been observed under the treatment value $a = 0$ ($Y^{a=0}$). The variables $Y^{a=1}$ and $Y^{a=0}$ are referred to as *potential outcomes* or as *counterfactual outcomes*. In economics, we often refer "counterfactual outcomes" to outcomes that had not happend. But here, both $Y^{a=0}$ and $Y^{a=1}$ are counterfactual outcomes, no matter auctually $a = 0$ or $a = 1$.

The counterfactual outcomes that corresponds to the treatment value that the individual actually received is *actually factual*.

Consistency:

$$\text{if } A_i = a, \text{ then } Y_i^a = Y^{A_i} = Y_i$$

an individual with observed treatment $A = a$, has observed outcome Y equal to his counterfactual outcome Y^a.

Consistency is that the observed outcome is equal to what (we think that) would have been observed.

In general, individual causal effects cannot be identified – that is, cannot be expressed as a function of the observed data – because of missing data.

1.2 Average causal effects

Average causal effect in population:

$$\mathrm{E}\left[Y^{a=1}\right] \neq \mathrm{E}\left[Y^{a=0}\right]$$

[*] Hernán MA, Robins JM (2019). Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming.

Absence of an average causal effect does not imply absence of individual effects. When there is no causal effect for any individual in the population, i.e., $Y^{a=1} = Y^{a=0}$ for all individuals, we say that the *sharp causal null hypothesis* is true.

Average causal effects can sometimes be identified from data, even if individual causal effects cannot.

1.3 Measures of causal effect

We can represent the causal null by: (1) causal risk difference $\Pr\left[Y^{a=1} = 1\right] - \Pr\left[Y^{a=0} = 1\right] = 0$; (2) causal risk ratio $\frac{\Pr\left[Y^{a=1}=1\right]}{\Pr\left[Y^{a=0}=1\right]} = 1$; (3) causal odds ratio $\frac{\Pr\left[Y^{a=1}=1\right]/\Pr\left[Y^{a=1}=0\right]}{\Pr\left[Y^{a=0}=1\right]/\Pr\left[Y^{a=0}=0\right]} = 1$.

These causal parameters quantify the strength of the same causal effect on different scales. Because the causal risk difference, risk ratio, and odds ratio (and other summaries) measure the causal effect, we refer to them as *effect measures*.

1.4 Random variability

In causal inference, random error derives from sampling variability, nondeterministic counterfactuals, or both.

1^{st} source of random error is *sampling variability*. When we only have a random sample from a much larger, near-infinite population, $\Pr\left[Y^{a=0} = 1\right]$ cannot be directly computed (instead, we have $\widehat{\Pr}\left[Y^{a=0} = 1\right]$ from the sample, sometimes it is a consistent estimator of $\Pr\left[Y^{a=0} = 1\right]$).

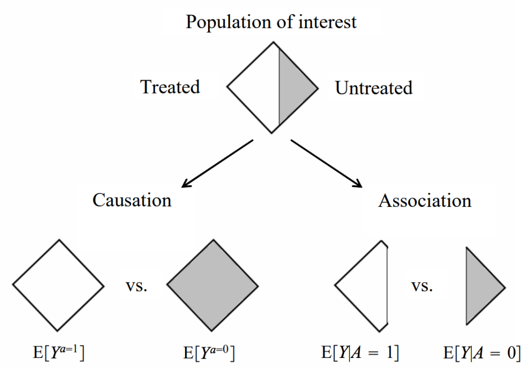
2^{nd} source of random error is *nondeterministic counterfactuals*. Counterfactual outcomes may be stochastic or nondeterministic.

TODO: Technical Point 1.2 **Nondeterministic counterfactuals** needs to be added.

1.5 Causation versus association

Some equivalent definitions of independence are: (1) associational risk difference $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0] = 0$; (2) associational risk ratio $\frac{\Pr[Y=1|A=1]}{\Pr[Y=1|A=0]} = 1$; (3) associational odds ratio $\frac{\Pr[Y=1|A=1]/\Pr[Y=0|A=1]}{\Pr[Y=1|A=0]/\Pr[Y=0|A=0]} = 1$. The left hand sides measure the association on different scales, and we refer to them as *association measures*.

For a continuous outcome Y we define *mean independence* between treatment and outcome as: $\mathrm{E}[Y|A = 1] = \mathrm{E}[Y|A = 0]$. Independence and mean independence are the same concept for dichotomous outcomes.



As the figure shows, *association* is defined by a different risk in two disjoint subsets of the population determined by the individuals' actual treatment value ($A = 1$ or $A = 0$), whereas *causation* is defined by a different risk in the same population under two different treatment values

($a = 1$ or $a = 0$). In economics, “selection bias” may be a problem for association. In this book, the discrepancy between association and causation is referred as *confounding*.

2 Randomized experiments

2.1 Randomization

In ideal randomized experiments, association is causation.

Exchangeability:

$$Y^a \perp A \text{ for all } a$$

That is, the risk under the potential treatment value a among the treated, $\Pr[Y^a = 1|A = 1]$, equals the risk under the potential treatment value a among the untreated, $\Pr[Y^a = 1|A = 0]$, for both $a = 0$ and $a = 1$. $\Pr[Y^a = 1|A = 1]$ means that we choose people by $A = 1$, and observe their outcome with treatment value a.

Randomization is so highly valued because it is expected to produce exchangeability. When the treated and the untreated are exchangeable, we sometimes say that treatment is exogenous, and thus *exogeneity* is commonly used as a synonym for exchangeability.

Independence between the counterfactual outcome and the observed treatment $Y^a \perp A$ does not imply independence between the observed outcome and the observed treatment $Y \perp A$. For example, suppose there is a causal effect on some individuals so that $Y^{a=1} \neq Y^{a=0}$. Since $Y = Y^A$, then Y^a with a evaluated at the observed treatment A is the observed Y^A, which depends on A and thus will not be independent of A.

It is possible that a study is a randomized experiment even if exchangeability does not hold in infinite samples. There may be two reasons: (1) the sample size is too small, random fluctuations arising from sampling variability could explain almost anything; (2) randomized experiments with more than one randomization step.

2.2 Conditional randomization

We call the experiments using a single unconditional (marginal) randomization probability that is common to all individuals as *marginally randomized experiments*. And we call the experiments using several randomization probabilities that depend (are conditional) on the values of the variables as *conditionally randomized experiments*.

Conditional randomization does not guarantee unconditional (or marginal) exchangeability $Y^a \perp A$, it guarantees *conditional exchangeability* $Y^a \perp A|L$ within levels of the variable L.

Conditional exchangeability:

$$Y^a \perp A|L \ (Y^a \perp A|L = l \text{ holds for all values } l) \text{ for all } a$$

We can compute the average causal effect in each of these subsets of strata of the population, for example, $\Pr\left[Y^{a=1} = 1|L = 1\right] / \Pr\left[Y^{a=0} = 1|L = 1\right]$ (however, we can compute the average causal effect $\Pr\left[Y^{a=1} = 1\right] / \Pr\left[Y^{a=0} = 1\right]$ in the entire population, if we do not expect to have information on L for future individuals). We refer to this method to compute stratum-specific causal effects as *stratification*. Stratumspecific causal risk ratio in the subset $L = 1$ may differ from the causal risk ratio in $L = 0$. In that case, we say that the effect of treatment is modified by L, or that there is *effect modification* by L.

2.3 Standardization

3 Observational studies

4 Effect modification

5 Interaction

6 Graphical representation of causal effects

7 Confounding

8 Selection bias

9 Measurement bias

10 Random variability

Causal inference with models

11 Why model?

11.1 Data cannot speak for themselves

We cannot always let the data "speak for themselves" to obtain a consistent estimate. Rather, we often need to supplement the data with a model. For example, we would not compute an estimator $\widehat{\mathrm{E}}[Y|A=a]$ of $\mathrm{E}[Y|A=a]$ when A were a truly continuous variable, the sample average would be undefined for nearly all treatment levels.

11.2 Parametric estimators

A model is an a priori restriction on the distribution of the data. A parametric model is like adding information that is not in the data to compensate for the lack of sufficient information in the data themselves.

Model-based causal inference relies on the condition of (approximately) *no model misspecification.* Because parametric models are rarely, if ever, perfectly specified, certain degree of model misspecification is almost always expected.

11.3 Nonparametric estimators

Whenever the number of parameters in the model is equal to the number of population quantities that can be estimated by using the model, then the model is *saturated.* A saturated model has the same number of unknowns in both sides of the equal sign.

When a model has only a few parameters but it is used to estimate many population quantities, we say that the model is *parsimonious.*

For causal inference, identifiability assumptions are the assumptions that we would have to make even if we had an infinite amount of data. Modeling assumptions are the assumptions that we have to make precisely because we do not have an infinite amount of data.

11.4 The bias-variance trade-off

In general, the larger the number of parameters in the model, the fewer restrictions the model imposes; the less smooth the model, the more protection afforded against bias from model misspecification. Although less smooth models may yield a less biased estimate, they also result in a larger variance, i.e., wider 95% confidence intervals around the estimate.

Technical Point 11.1 (A taxonomy of commonly used models): Linear regression models are a subset of larger class of models: Generalized Linear Models. GLMs have three components: a linear functional form $\sum_{i=0}^{p} \theta_i X_i$, a link function $g\{\}$ such that $g\{\mathrm{E}[Y|X]\} = \sum_{i=1}^{p} \theta_i X_i$, and a distribution for the Y conditional on X. If we do not model the distribution of Y conditional on X, we refer to the model as a conditional mean model. We can restrict outcomes with different link functions (identity link, log link, logit link, etc.). There are interesting introductions, that we omit here, about generalized estimating equation (GEE) models, generalized additive models (GAMs) and kernel regression models.

12 IP weighting and marginal structural models

13 Standardization and the parametric g-formula

14 G-estimation of structural nested models

15 Outcome regression and propensity scores

Outcome regression and propensity scores are commonly used but have limited applicability for complex longitudinal data.

15.1 Outcome regression

15.2 Propensity scores

If the distribution of $p(L)$ were the same for the treated $A = 1$ and the untreated $A = 0$, then there would be no confounding due to L, i.e., there would be no open path from L to A on a causal diagram.

Individuals with same propensity score $p(L)$ will generally have different values of some covariates L. For example, two individuals with $p(L) = 0.2$ may differ with respect to smoking intensity and exercise, and yet they may be equally likely to quit smoking given all the variables in L. That is, both individuals have the same conditional probability of ending up in the treated group $A = 1$. If we consider all individuals with a given value of $p(L)$ in the superpopulation, this group will include individuals with different values of L (e.g., different values of smoking intensity and exercise), but the distribution of L will be the same in the treated and the untreated, that is, $A \perp L|p(L)$. We say the propensity score balances the covariates between the treated and the untreated.

15.3 Propensity stratification and standardization

15.4 Propensity matching

Defining closeness in propensity matching entails a bias-variance trade-off.

Using propensity scores to detect the overlapping range of the treated and the untreated may be useful, but simply restricting the study population to that range is a lazy way to ensure positivity. The automatic positivity ensured by propensity matching needs to be weighed against the difficulty of assessing transportability when restriction is solely based on the value of the estimated propensity scores.

15.5 Propensity models, structural models, predictive models

Propensity models are models for the probability of treatment A given the variables L used to try to achieve conditional exchangeability.

The dual use of outcome regression in both causal inference method and in prediction has led to many misunderstandings. One of the most important misunderstandings has to do with variable selection procedures. When the interest lies exclusively on outcome prediction, investigators want to select any variables that, when included as covariates in the model, improve its predictive ability. Many well-known variable selection procedures — e.g., forward selection, backward elimination, stepwise selection — and more recent developments in machine learning are used to enhance prediction. These are powerful tools for investigators who are interested in prediction, especially when dealing with very high-dimensional data. A possible result of this mismatch is the inclusion of superfluous — or even harmful — covariates in propensity models and structural models. Specifically, the application of predictive algorithms to causal inference models may result in inflated variances.

Propensity models do not need to predict treatment very well. They just need to include the variables L that guarantee exchangeability. Covariates that are strongly associated with treatment, but are not necessary to guarantee exchangeability, do not help reduce bias. If these covariates were included in L, adjustment can actually result in estimates with very large variances. Only variables that influence simultaneously the participation decision and the outcome variable should be included. It should also be clear that only variables that are unaffected by participation (or the anticipation of it) should be included in the model. To ensure this, variables should either be fixed over time or measured before participation.(Caliendo & Kopeinig, 2008)

Besides variance inflation, a predictive attitude towards variable selection for causal inference models — both propensity models and outcome regression models — may also result in self-inflicted bias.

16 Instrumental variable estimation

17 Causal survival analysis

18 Variable selection for causal inference