Propensity Score Methods, Models and Adjustment

Dr David A. Stephens

Department of Mathematics & Statistics
McGill University
Montreal, QC, Canada.

david.stephens@mcgill.ca
www.math.mcgill.ca/dstephens/SISCER2021/

The Key Challenge

Key issue: causal conclusions from observational data

- Experimental studies:
  - Treatment assigned by the researcher, independent of confounding factors;
  - Causal statements possible.
- Observational studies:
  - Treatment assignment dependent on confounding factors;
  - Causal statements not possible ?

Module Content

1. The need for adjustment: confounding in observational studies.
2. Manufacturing balance: the propensity score.
3. Statistical tools utilizing the propensity score.
4. Examples and extensions.

Part 1: Introduction

1.1 The central causal question

1.2 Notation

1.3 Causal estimands

1.4 Basics of estimation

1.5 The Monte Carlo paradigm

1.6 Collapsibility

1.7 The randomized study

1.8 Confounding

1.9 Statistical modelling

Part 2: The Propensity Score

2.1 Manufacturing balance

2.2 The propensity score for binary exposures

2.3 Matching via the propensity score

2.4 The Generalized Propensity Score

2.5 Propensity score regression

2.6 Adjustment by weighting

2.7 Augmentation and double robustness

Part 3: Implementation and Computation

3.1 Statistical tools

3.2 Key considerations

Part 4: Extensions

4.1 Longitudinal studies

4.2 The Marginal Structural Model (MSM)

Part 5: New Directions

5.1 New challenges

5.2 Bayesian approaches

Part 1

Introduction

The central causal question

In many research domains, the objective of an investigation is to quantify the effect on a measurable outcome of changing one of the conditions under which the outcome is measured.

- in a health research setting, we may wish to discover the benefits of a new therapy compared to standard care;
- in economics, we may wish to study the impact of a training programme on the wages of unskilled workers;
- in transportation, we may attempt to understand the effect of embarking upon road building schemes on traffic flow or density in a metropolitan area.

The central statistical challenge is that, unless the condition of interest is changed independently, the inferred effect may be subject to the influence of other variables.

The central causal question

Example: The effect of nutrition on health

In a large cohort, the relationship between diet and health status is to be investigated. Study participants are queried on the nutritional quality of their diets, and their health status in relation to key indicators is assessed via questionnaires.

For a specific outcome condition of interest, incidence of cardiovascular disease (CVD), the relation to a specific dietary component, vitamin E intake, is to be assessed.

In the study, both incidence of disease and vitamin E intake were dichotomized

- Exposure: Normal/Low intake of vitamin E.
- Outcome: No incidence/Incidence of CVD in five years from study initiation.

The central causal question

Example: The effect of nutrition on health

		Outcome	
		CVD	No CVD
Exposure	Normal	27	8020
	Low	86	1879

Question: does a diet lower in vitamin E lead to higher chance of developing CVD ? More specifically, is this a *causal* link ?

- that is, if we were to *intervene* to change an individual's exposure status, by how much would their risk of CVD change ?

The language of causal inference

We seek to quantify the effect on an *outcome* of changes in the value of an *exposure* or *treatment*.

- Outcome: could be
  - binary;
  - integer-valued;
  - continuous-valued.
- Exposure: could be
  - binary;
  - integer-valued;
  - continuous-valued.
- Study: could be
  - cross-sectional (single time point);
  - longitudinal (multiple time points), with single or multiple exposures.

We consider an *intervention* to change exposure status.

Notation

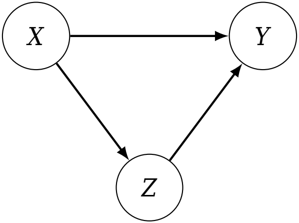
We adopt the following notation: let

- i index individuals included in the study;
- Y_i denote the *outcome* for individual i;
- Z_i denote the *exposure* for individual i;
- X_i denote the values of other *predictors* (or *covariates*).

For a cross-sectional study, Y_i and Z_i will be scalar-valued; for the longitudinal case, Y_i and Z_i may be vector valued. X_i is typically vector-valued at each measurement time point.

We will treat these variables as *random* quantities, and regard them as samples from an infinite population, rather than a finite population.

Graphical representation



Directed Acyclic Graph (DAG) for basic confounding set up in observational studies.

DAGs are commonly used to clarify causal thinking and assumptions.

Graphical representation

- an inbound arrow indicates a causal relationship
  - X is a *direct cause* of Y and Z;
  - Z is a direct cause of Y, but also a *mediator* of the *indirect cause* of X on Y;
- a variable (node) that has no inbound arrows can be considered a '*founder*' variable;
- we must consider *paths* from the exposure Z to the outcome Y; there are two
  - the *direct* path $Z \longrightarrow Y$,
  - the *indirect* path $Z \longrightarrow X \longrightarrow Y$

Structural modelling

We can think of the DAG as encapsulating the following equations:

$$Z = g_Z(X, \varepsilon_Z)$$

$$Y = g_Y(X, Z, \varepsilon_Y)$$

where ε_Z and ε_Y are independent random perturbations, and g_Z and g_Y are mapping functions.

That is,

- we take X and ε_Z and combine them through g_Z to obtain Z;
- we combine Z with X and ε_Y through g_Y to obtain Y.

Structural modelling

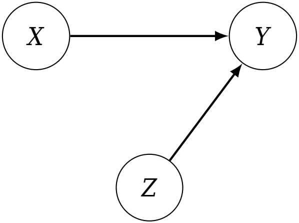
For example

$$Z = X + \varepsilon_Z$$

$$Y = 2X + 5Z + 3XZ + \varepsilon_Y$$

Causal goal

Our goal is to understand the *unconfounded* effect of Z on Y, that is, where X is *not treated as a cause* of Z.



DAG with no confounding.

Causal goal

In the structural model, we imagine Z being fixed to some value, z say, not generated by its structural model.

$$Y = 2X + 5\mathsf{z} + 3X\mathsf{z} + \varepsilon_Y$$

Counterfactual or Potential Outcomes

In order to phrase causal questions of interest, it is useful to consider certain hypothetical outcome quantities that represent the possible outcomes under different exposure alternatives.

We denote by

$$Y_i(\mathsf{z})$$

the outcome for individual i if we *intervene* to set exposure to z.

$Y_i(\mathsf{z})$ is termed a *counterfactual* or *potential outcome*.

Counterfactual or Potential Outcomes

If exposure is binary, the pair of potential outcomes

$$\{Y_i(0), Y_i(1)\}$$

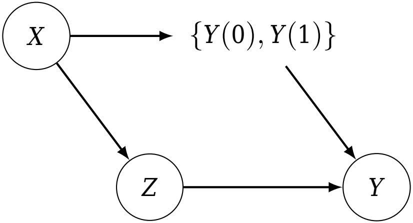
represent the outcomes that would result for individual i if that subject was not exposed, or exposed, respectively.

The observed outcome, Y_i, may be written in terms of the potential outcomes and the observed exposure, Z_i, as

$$Y_i = (1 - Z_i)Y_i(0) + Z_iY_i(1).$$

Counterfactual or Potential Outcomes

That is, $Y(0)$ and $Y(1)$ are (potentially) caused by X, but not Z.



DAG with potential outcomes

Counterfactual or Potential Outcomes

If exposure is multi-valued, the potential outcomes

$$\{Y_i(\mathsf{z}_1), Y_i(\mathsf{z}_2), \ldots, Y_i(\mathsf{z}_d)\}$$

represent the outcomes that would result for individual i if that subject exposed to exposure level $\mathsf{z}_1, \mathsf{z}_2, \ldots, \mathsf{z}_d$ respectively.

Counterfactual or Potential Outcomes

The observed outcome, Y_i, may then be written in terms of the potential outcomes and the observed exposure, Z_i, as

$$Y_i = \sum_{j=1}^{d} \mathbb{1}_{\{\mathsf{z}_j\}}(Z_i) Y_i(\mathsf{z}_j).$$

where $\mathbb{1}_{\mathcal{A}}(Z)$ is the *indicator*[1] for the set $\mathcal{A}$, with $\mathbb{1}_{\mathcal{A}}(Z) = 1$ if $Z \in \mathcal{A}$, and zero otherwise.

[1] or ‘pick-off’ function

Counterfactual or Potential Outcomes

If exposure is continuous-valued, the potential outcomes

$$\{Y_i(\mathsf{z}), \mathsf{z} \in \mathcal{Z}\}$$

represent the outcomes that would result for individual i if that subject exposed to exposure level z which varies in the set $\mathcal{Z}$.

Counterfactual or Potential Outcomes

Note 1.

It is rare that we can ever observe more than one of the potential outcomes for a given subject in a given study, that is, for binary exposures it is rare that we will be able to observe both

$$Y_i(0) \quad \text{and} \quad Y_i(1)$$

in the same study.

In the previous example, we cannot observe the CVD outcome under both the assumption that the subject *did* and simultaneously *did not* have a low vitamin E diet.

This is the first fundamental challenge of causal inference.

Causal Estimands

The central question of causal inference relates to comparing the (expected) values of different potential outcomes.

We consider the causal effect of exposure to be defined by *differences* in potential outcomes corresponding to *different* exposure levels.

Note 2.

This is a statistical, rather than necessarily mechanistic, definition of causality.

Binary Exposures

For a binary exposure, we define the causal effect of exposure by considering contrasts between $Y_i(0)$ and $Y_i(1)$; for example, we might consider

- Additive contrasts

$$Y_i(1) - Y_i(0)$$

- Multiplicative contrasts

$$Y_i(1)/Y_i(0)$$

Continuous Exposures

For a continuous exposure, we might consider the path tracing how $Y_i(z)$ changes as z changes across some relevant set of values.

This leads to a *causal dose-response* function.

Example: Occlusion Therapy for Amblyopia

We might seek to study the effect of occlusion therapy (patching) on vision improvement of amblyopic children. Patching 'doses' are measured in terms of time for which the fellow (normal functioning) eye is patched.

As time is measured continuously, we may consider how vision improvement changes for any relevant dose of occlusion.

Expected counterfactuals

In general, we are interested in *population* causal effects based on *expected* potential outcomes

$$\mathbb{E}[Y_i(\mathsf{z})]$$

or contrasts of these quantities.

We might also consider *subgroup-specific* expected quantities

$$\mathbb{E}[Y_i(\mathsf{z})|i \in \mathcal{I}]$$

where $\mathcal{I}$ is some stratum of interest in the general population.

Expected counterfactuals: binary exposure

For a binary exposure, we might consider the average effect of exposure (or *average treatment effect*, ATE) defined as

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

If the outcome is also binary, we note that

$$\mathbb{E}[Y_i(\mathsf{z})] \equiv \Pr[Y_i(\mathsf{z}) = 1]$$

so may also consider odds or odds ratios quantities

$$\frac{\Pr[Y_i(\mathsf{z}) = 1]}{\Pr[Y_i(\mathsf{z}) = 0]} \qquad\qquad \frac{\Pr[Y_i(1) = 1]/\Pr[Y_i(1) = 0]}{\Pr[Y_i(0) = 1]/\Pr[Y_i(0) = 0]}.$$

Expected counterfactuals: binary exposure

We may also consider quantities such as the

average treatment effect on the treated, ATT

defined as

$$\mathbb{E}[Y_i(1) - Y_i(0)|Z_i = 1]$$

although such quantities can be harder to interpret.

Example: antidepressants and autism

Example:

Antidepressants are quite widely prescribed for a variety of mental health concerns. However, pregnant women may be reluctant to embark on a course of antidepressants during pregnancy.

We might wish to investigate, in a population of users (and potential users) of antidepressants, the incidence of autism-spectrum disorder in early childhood and to assess the possibility of causal influence of antidepressant use on this incidence.

Example: antidepressants and autism

Example:

- Outcome: binary, recording the a diagnosis of autism-spectrum disorder in the child by age 5;
- Exposure: antidepressant use during 2nd or 3rd trimester of pregnancy.

Then we may wish to quantity

$$\mathbb{E}\big[Y_i(\text{antidepressant}) - Y_i(\text{no antidepressant})\big|\text{Antidep. actually used}\big].$$

Estimation of average potential outcomes

We wish to obtain estimates of causal quantities of interest based on the available data, which typically constitute a random sample from the target population.

Typically, we will use sample mean type quantities: for a random sample of size n, the sample mean

$$\frac{1}{n}\sum_{i=1}^{n} Y_i$$

is an estimator of the population mean and so on.

Estimation of average potential outcomes

In a typical causal setting, we wish to perform estimation of

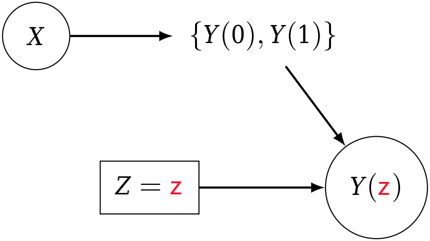
average potential outcome

(APO) values.

Consider first the situation where all subjects in a random sample receive a given exposure $\mathtt{z}$; we wish to estimate $\mathbb{E}[Y(\mathtt{z})]$.

Estimation of average potential outcomes

The intervention to set $Z = z$ is done independently of X, so the arrow $X \rightarrow Z$ is *removed*.



DAG with exposure intervention $Z = z$

Estimation of average potential outcomes

As a mathematical calculation, we write the expected outcome as

$$\mathbb{E}[Y(z)] = \int y \, f_{Y(z)}(y) \; \mathrm{d}y$$

which we read as

“average the collection of possible y values weighted by their probability of being observed".

The quantity $f_{Y(z)}(y)$ is the hypothetical distribution of the potential outcome $Y(z)$.

Estimation of average potential outcomes

We may also write this as

$$\mathbb{E}[Y(z)] = \int y \, f_{Y(z),X}(y, x) \;\; \mathrm{d}y \;\; \mathrm{d}x$$

which recognizes that in the population, the values of the predictors X also vary randomly according to some probability distribution.

Estimation of average potential outcomes

The quantity $f_{Y(\mathsf{z}),X}(y,x)$ is the hypothetical joint distribution of the potential outcome $Y(\mathsf{z})$ and X

- this describes how these two quantities vary together.
  - $f_{Y(\mathsf{z})}(y)$ is the distribution of the potential outcome $Y(.)$ when we set the exposure to z.
  - $f_{Y(\mathsf{z}),X}(y,x)$ is the joint distribution of $(Y(.),X)$ in the population where we set the exposure to z.

Estimation of average potential outcomes

Note that we may also write

$$\mathbb{E}[Y(\mathsf{z})] = \int y \mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y(\mathsf{z}),X}(y, x) \;\, \mathrm{d}y \;\, \mathrm{d}z \;\, \mathrm{d}x$$

assuming an exposure distribution that sets $z = \mathsf{z}$ with probability one.

- the data are considered to be sampled from the distribution

$$\mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y(\mathsf{z}),X}(y, x) = \mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y(\mathsf{z})|X}(y|x) f_X(x).$$

Estimation of average potential outcomes

However, remembering the DAG for our intervention (p. 38) we can deduce that

$$f_{Y(\mathsf{z})|X}(y|x) = f_{Y|Z,X}(y|\mathsf{z}, x)$$

so that the population distribution becomes

$$\mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y|Z,X}(y|z,x) f_X(x).$$

Thus, for the APO we have

$$\mathbb{E}[Y(\mathsf{z})] = \int y\, \mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y(\mathsf{z}),X}(y|x) f_X(x) \;\, \mathrm{d}y \;\, \mathrm{d}z \;\, \mathrm{d}x$$

Estimation of average potential outcomes

Now, in our hypothetical sample, we have observed n data points

$$\{(x_i, y_i, z_i), i = 1, \ldots, n\}$$

from the joint distribution

$$\mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y(\mathsf{z}),X}(y|x) f_X(x)$$

so that $z_i = \mathsf{z}$ for all i. We may *estimate* the relevant APO $\mathbb{E}[Y(\mathsf{z})]$ by

$$\widehat{\mathbb{E}}[Y(\mathsf{z})] = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}.$$

Estimation of average potential outcomes

Note 3.

To estimate functions of the sample mean, we may use simple transformations of the estimator; for example, if the outcome is binary, we estimate the odds

$$\frac{\Pr[Y_i(z) = 1]}{\Pr[Y_i(z) = 0]} \qquad \text{by} \qquad \frac{\bar{y}}{1 - \bar{y}}.$$

Side Note: Monte Carlo methods

Causal quantities are typically *average* measures across a given population, hence we often need to consider integrals with respect to probability distributions.

For any function $g(.)$, we have

$$\mathbb{E}[g(Y)] = \int g(y)\, f_Y(y)\ \mathrm{d}y$$

$$= \int g(y)\, f_{Y,X}(y,x)\ \mathrm{d}y\ \mathrm{d}x$$

Rather than performing this calculation using integration, we approximate it numerically using *Monte Carlo*.

Side Note: Monte Carlo methods

Monte Carlo calculations proceed as follows:

- generate a sample of size n from the density

  $$f_Y(y)$$

  to yield $y_1, \ldots, y_n$; there are standard techniques to achieve this.
- approximate $\mathbb{E}[g(Y)]$ by

  $$\widehat{\mathbb{E}}[g(Y)] = \frac{1}{n}\sum_{i=1}^{n} g(y_i).$$

- For large n, $\widehat{\mathbb{E}}[g(Y)]$ provides a good approximation to $\mathbb{E}[g(Y)]$.

Side Note: Monte Carlo methods

Note 4.

This calculation is *at the heart of frequentist methods in statistics*:

- we collect a sample of *data* of size n,
- form *estimates* based on this sample (which often correspond to sample averages),
- if our sample is large enough, we are confident in our results.

Side Note: Monte Carlo methods

We have that

$$\mathbb{E}[g(Y)] = \int g(y)\, f_Y(y)\ \mathrm{d}y = \int g(y)\, \frac{f_Y(y)}{f_Y^*(y)} f_Y^*(y)\ \mathrm{d}y$$

where $f_Y^*(y)$ is some other density. Thus

$$\mathbb{E}_{f_Y}[g(Y)] \equiv \mathbb{E}_{f_Y^*}\left[g(Y)\frac{f_Y(Y)}{f_Y^*(Y)}\right].$$

Side Note: Monte Carlo methods

This is known as *importance sampling*: we

- generate a sample of size n from the density

$$f_Y^*(y)$$

  to yield $y_1, \ldots, y_n$;

- approximate $\mathbb{E}[g(Y)]$ by

$$\widehat{\mathbb{E}}[g(Y)] = \frac{1}{n}\sum_{i=1}^{n} g(y_i)\frac{f_Y(y_i)}{f_Y^*(y_i)}.$$

Side Note: Monte Carlo methods

This means that even if we do *not* have a sample from the distribution of interest, f_Y, we can still compute averages with respect to f_Y if we have access to a sample from a related distribution, f_Y^*.

Clearly, for the importance sampling computation to work, we need that

$$\frac{f_Y(y_i)}{f_Y^*(y_i)}$$

is *finite* for the required range of Y, which means that we must have

$$f_Y^*(y) > 0 \quad \text{whenever} \quad f_Y(y) > 0.$$

Side Note: Marginal and conditional measures of effect

Many of the causal measures described above are *marginal* measures.

That is, they involve *averaging* over the distribution of X: as we have seen

$$\mathbb{E}[Y(\mathsf{z})] = \int y \, f_{Y|Z,X}(y|\mathsf{z}, x) f_X(x) \;\, \mathrm{d}y \;\, \mathrm{d}x.$$

This is sometimes known as a *G-computation* formula.

Side Note: Marginal and conditional measures of effect

Marginal measures are not typically the same as the equivalent measure defined for the *conditional* model

$$f_{Y|Z,X}(y|z,x)$$

Marginal measures that do not have the same interpretation in the conditional model are termed *non-collapsible*.

Side Note: Marginal and conditional measures of effect

Example: Logistic regression

Consider the binary response, binary exposure regression model, where

$$\Pr[Y = 1|Z = z, X = x] = \frac{\exp\{\beta_0 + \beta_1 z + \beta_2 x\}}{1 + \exp\{\beta_0 + \beta_1 z + \beta_2 x\}} = \mu(x, z; \beta)$$

say. We then have that in this *conditional* (on x) model, the parameter

$$\beta_1 = \log\left(\frac{\Pr[Y = 1|Z = 1, X = x] / \Pr[Y = 0|Z = 1, X = x]}{\Pr[Y = 1|Z = 0, X = x] / \Pr[Y = 0|Z = 0, X = x]}\right)$$

is the log odds ratio comparing outcome probabilities with for $Z = 1$ and $Z = 0$ respectively.

Side Note: Marginal and conditional measures of effect

Example: Logistic regression

In the *marginal* model, we wish to consider

$$\Pr[Y = 1|Z = z]$$

directly, and from the specified conditional model we have

$$\Pr[Y = 1|Z = z] = \int \Pr[Y = 1|Z = z, X = x] f_X(x) \;\, \mathrm{d}x$$

assuming that Z and X are *independent*. Explicitly,

$$\Pr[Y = 1|Z = z] = \int \mu(x, z; \beta) f_X(x) \;\, \mathrm{d}x$$

Side Note: Marginal and conditional measures of effect

Example: Logistic regression

Typically, the integral that defines $\Pr[Y = 1|Z = z]$ in this way is not tractable. However, as, Y is binary, we may still consider a logistic regression model in the marginal distribution, say parameterized as

$$\Pr[Y = 1|Z = z] = \frac{\exp\{\theta_0 + \theta_1 z\}}{1 + \exp\{\theta_0 + \theta_1 z\}}$$

where θ_1 is the marginal log odds ratio.

In general, $\beta_1 \neq \theta_1$.

The randomized study

The approach that intervenes to set exposure equal to z for all subjects, however, does not facilitate comparison of APOs for different values of z.

Therefore consider a study design based on *randomization*; consider from simplicity the binary exposure case. Suppose that a random sample of size $2n$ is obtained, and split into two equal parts.

- the *first* group of n are assigned the exposure and form the '*exposed*' or '*treated*' sample,
- the *second* group are left '*untreated*'.

The randomized study

For both the treated and untreated groups we may use the previous logic, and estimate the ATE

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

by the difference in means in the two groups, that is

$$\frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=n+1}^{2n} y_i.$$

The key idea here is that the two halves of the original sample are *exchangeable* with respect to their properties:

- the only systematic difference between them is *due to exposure assignment*.

The randomized study

In a slightly modified design, suppose that we obtain a random sample of size n from the study population, but then assign exposure *randomly* to subjects in the sample: subject i receives treatment with probability p.

- if $p = 1/2$, there is an equal chance of receiving treatment or not;
- we may choose any value of $0 < p < 1$.

In the final sample, the number treated, n_1, is a realization of a random variable N_1 where

$$N_1 \sim \text{Binomial}(n, p).$$

The randomized study

This suggests the estimators[2]

$$\widehat{\mathbb{E}}[Y(z)] = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i) Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i)} \qquad z = 0, 1 \tag{1}$$

where the indicator $\mathbb{1}_{\{z\}}(Z_i)$ identifies individuals that received treatment z.

[2] Formula (1) just says to take the mean in each treatment group !

The randomized study

Note that for the denominator,

$$\sum_{i=1}^{n} \mathbb{1}_{\{1\}}(Z_i) \sim \text{Binomial}(n, p)$$

so we may consider replacing the denominators by their expected values

$$np \qquad \text{and} \qquad n(1-p)$$

respectively for $z = 0, 1$. This yields the estimators

$$\widehat{\mathbb{E}}[Y(1)] = \frac{1}{np} \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(Z_i) Y_i \qquad \widehat{\mathbb{E}}[Y(0)] = \frac{1}{n(1-p)} \sum_{i=1}^{n} \mathbb{1}_{\{0\}}(Z_i) Y_i. \tag{2}$$

The randomized study

Note 5.

The estimators in (1) are *more efficient* than the estimators in (2), that is, they have *lower variances*.

It is more efficient to use an estimated value of p

$$\widehat{p} = \frac{n_1}{n}$$

than p itself.

The randomized study

We have that

$$\mathbb{E}[Y(\mathsf{z})] = \frac{\displaystyle\int y\,\mathbb{1}_{\{\mathsf{z}\}}(z)\,f_{Y|Z,X}(y|z,x)f_X(x)f_Z(z)\ \mathrm{d}y\ \mathrm{d}z\ \mathrm{d}x}{\displaystyle\int \mathbb{1}_{\{\mathsf{z}\}}(z)f_Z(z)\ \mathrm{d}z}$$

and have data which are a random sample from the joint density

$$f_{Y|Z,X}(y|z,x)f_X(x)f_Z(z)$$

which demonstrates that the estimators in (1) are akin to *Monte Carlo* estimators.

The challenge of confounding

The second main challenge of causal inference is that for *observational* (or *non-experimental*) studies, *exposure is not necessarily assigned independently of other variables*.

- it may be that exposure is assigned dependent on one or more of the measured predictors;
- if these predictors also predict outcome, then there is the possibility of *confounding* of the causal effect of exposure by those other variables;
- this is the set up in the DAG on p. 15.

The challenge of confounding

Specifically, in terms of densities, if predictor(s) X

- predicts outcome Y in the presence of Z:

$$f_{Y|Z,X}(y|z,x) \neq f_{Y|Z}(y|z)$$

and

- predicts exposure Z:

$$f_{Z|X}(z|x) \neq f_Z(z)$$

then X is a *confounder*.

Confounding: example

Example: The effect of nutrition on health: revisited

The relationship between low vitamin E diet and CVD incidence may be confounded by socio-economic status (SES); poorer individuals may have worse diets, and also may have higher risk of cardiovascular incidents via mechanisms other than those determined by diet:

- smoking;
- pollution;
- access to preventive measures/health advice.

Confounding

Confounding is a central challenge as it renders the observed sample unsuitable for causal comparisons unless adjustments are made:

- in the binary case, if confounding is present, the treated and untreated groups are *not directly comparable*;
- the effect of confounder X on outcome is potentially *different* in the treated and untreated groups.
- direct comparison of sample means *does not* yield valid insight into average treatment effects;

Causal inference is fundamentally about comparing exposure subgroups on an *equal footing*, where there is no residual influence of the other predictors. This is possible in the randomized study as randomization breaks the association between Z and X.

Confounding and collapsibility

Note 6.

Confounding is not the same as non-collapsibility.

- Non-collapsibility concerns the measures of effect being reported, and the parameters being estimated; parameters in a marginal model do not in general correspond to parameters in a conditional model.

  Non-collapsibility is a property of the model, not the study design. It may be present even for a randomized study.

- Confounding concerns the inter-relationship between outcome, exposure and confounder. It is not model-dependent, and does depend on the study design.

Simple confounding example

Suppose that Y, Z and X are all binary variables. Suppose that the true (structural) relationship between Y and (Z, X) is given by

$$\mathbb{E}[Y|Z=z, X=x] = \Pr[Y=1|Z=z, X=x] = 0.2 + 0.2z - 0.1x$$

with $\Pr[X=1] = q$. Then, by iterated expectation

$$\mathbb{E}[Y(\mathsf{z})] = 0.2 + 0.2\,\mathsf{z} - 0.1q$$

and

$$\mathbb{E}[Y(1) - Y(0)] = 0.2.$$

Simple confounding example

Suppose also that in the population from which the data are drawn

$$\Pr[Z = 1|X = x] = \begin{cases} p_0 & x = 0 \\ p_1 & x = 1 \end{cases} = (1 - x)p_0 + xp_1.$$

in which case

$$\Pr[Z = 1] = (1 - q)p_0 + qp_1.$$

Simple confounding example

If we consider the estimators in (2)

$$\widehat{\mathbb{E}}[Y(1)] = \frac{1}{np}\sum_{i=1}^{n} \mathbb{1}_{\{1\}}(Z_i)Y_i \qquad \widehat{\mathbb{E}}[Y(0)] = \frac{1}{n(1-p)}\sum_{i=1}^{n} \mathbb{1}_{\{0\}}(Z_i)Y_i$$

and set $p = (1-q)p_0 + qp_1$, we see that for the first term

$$\begin{aligned} \mathbb{E}_{Y,Z,X}[\mathbb{1}_{\{1\}}(Z)Y] &= \mathbb{E}_{Z,X}[\mathbb{1}_{\{1\}}(Z)\mathbb{E}_{Y|Z,X}[Y|Z,X]] \\ &= \mathbb{E}_{Z,X}[\mathbb{1}_{\{1\}}(Z)(0.2 + 0.2Z - 0.1X)] \\ &= 0.2\mathbb{E}_X[\mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)|X]] \\ &\quad + 0.2\mathbb{E}_X[\mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)Z|X]] \\ &\quad - 0.1\mathbb{E}_X[X\mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)|X])] \end{aligned}$$

Simple confounding example

Now

$$\begin{aligned}\mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)|X] &= \mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)Z|X] \\ &\equiv \Pr[Z=1|X] = (1-X)p_0 + Xp_1\end{aligned}$$

and

$$\mathbb{E}_X[\mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)|X]] = (1-q)p_0 + qp_1 = p$$

$$\mathbb{E}_X[\mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)Z|X]] = (1-q)p_0 + qp_1 = p$$

$$\mathbb{E}_X[X\mathbb{E}_{Z|X}[\mathbb{1}_{\{1\}}(Z)|X])] = qp_1$$

and therefore

$$\mathbb{E}_{Y,Z,X}[\mathbb{1}_{\{1\}}(Z)Y] = 0.4p - 0.1qp_1.$$

Simple confounding example

$$\therefore \quad \mathbb{E}\left[\frac{1}{np}\sum_{i=1}^{n}\mathbb{1}_{\{1\}}(Z_i)Y_i\right] = \frac{0.4p - 0.1p_1}{p}$$

By a similar calculation, as $\mathbb{1}_{\{0\}}(Z) = 1 - \mathbb{1}_{\{1\}}(Z)$,

$$\mathbb{E}_X[\mathbb{E}_{Z|X}[\mathbb{1}_{\{0\}}(Z)|X]] = 1 - p$$

$$\mathbb{E}_X[\mathbb{E}_{Z|X}[\mathbb{1}_{\{0\}}(Z)Z|X]] = 0$$

$$\mathbb{E}_X[X\mathbb{E}_{Z|X}[\mathbb{1}_{\{0\}}(Z)|X])] = q(1 - p_1)$$

so

$$\mathbb{E}\left[\frac{1}{n(1-p)}\sum_{i=1}^{n}\mathbb{1}_{\{0\}}(Z_i)Y_i\right] = \frac{0.2(1-p) - 0.1q(1-p_1)}{1-p}$$

Simple confounding example

Finally, therefore ATE estimator

$$\widehat{\mathbb{E}}[Y(1)] - \widehat{\mathbb{E}}[Y(0)]$$

has expectation

$$\frac{0.4p - 0.1qp_1}{p} - \frac{0.2(1-p) - 0.1q(1-p_1)}{1-p}$$

which equals

$$0.2 - 0.1q\left\{\frac{p_1}{p} - \frac{1-p_1}{1-p}\right\}$$

and therefore the unadjusted estimator based on (2) is *biased*.

Simple confounding example

The bias is caused by the fact that the two subsamples with

$$Z = 0 \qquad \text{and} \qquad Z = 1$$

are *not directly comparable* - they have a different profile in terms of X; by *Bayes theorem*

$$\Pr[X = 1|Z = 1] = \frac{p_1 q}{p} \qquad \Pr[X = 1|Z = 0] = \frac{(1 - p_1)q}{1 - p}$$

so, here, conditioning on $Z = 1$ and $Z = 0$ in turn in the computation of (2), leads to a different composition of X values in the two subsamples.

As X influences Y, the resulting Y values not directly comparable.

Instruments

If predictor $\widetilde{Z}$ predicts Z, but does not predict Y in the presence of Z, then $\widetilde{Z}$ is termed an *instrument*.

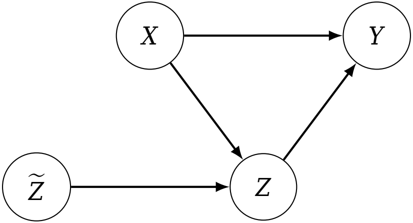
Example: Non-compliance

In a randomized study of a binary treatment, if Z_i records the treatment actually *received* by individual i, suppose that there is *non-compliance* with respect to the treatment; that is, if $\widetilde{Z}_i$ records the treatment *assigned* by the experimenter, then possibly

$$\widetilde{z}_i \neq z_i.$$

Then $\widetilde{Z}_i$ predicts Z_i, but is not associated with outcome Y_i given Z_i.

Instruments



DAG with instrument $\widetilde{Z}$.

Instruments

Instruments are *not* confounders as they do not predict outcome once the influence of the exposure has been accounted for.

Suppose in the previous confounding example, we had

$$\mathbb{E}[Y|Z=z,X=0] = \Pr[Y=1|Z=z,X=1] = 0.2 + 0.2z$$

for the structural model, but

$$\Pr[Z=1|X] = (1-X)p_0 + Xp_1.$$

Then X influences Z, and there is still an imbalance in the two subgroups indexed by Z with respect to the X values, *but* as X does not influence Y, there is *no bias* if the ATE estimator based on (2) is used.

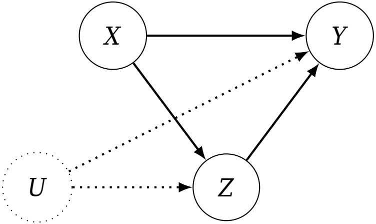
Critical Assumption

An important assumption that is commonly made is that of

No unmeasured confounding

that is, the measured predictors X include (possibly as a subset) all variables that confound the effect of Z on Y.

Critical Assumption



DAG with unmeasured confounder U.

Critical Assumption

We must assume that all variables that simultaneously influence exposure and outcome have been measured in the study.

- This is a strong (and possibly unrealistic) assumption in practical applications;
- It is the assumption made in standard regression analysis !
- It may be relaxed, and the influence of unmeasured confounders studied in sensitivity analyses.

Model-based analysis

So far, estimation based on the data via (1) and (2) has proceeded in a *non-parametric* or model-free fashion.

- models such as

  $$f_{Y(\mathsf{z}),X}(y,x)$$

  have been considered, but not modelled parametrically.

We now consider *semiparametric* specifications, where *parametric* models for example for

$$\mathbb{E}[Y(\mathsf{z})|X]$$

are considered but no distributional assumptions are made.

Correct model specification

We propose an *outcome mean model*

$$\mathbb{E}[Y|X, Z] = \mu(X, Z)$$

that may be parametric in nature, say

$$\mathbb{E}[Y|X, Z; \beta] = \mu(X, Z; \beta)$$

The importance of 'no unmeasured confounders'

An important consequence of the no unmeasured confounders assumption is that we have the *equivalence* of the conditional mean *structural* and *observed-data* outcome models, that is

$$\mathbb{E}[Y(\mathsf{z})|X] \quad \text{and} \quad \mathbb{E}[Y|X, Z = \mathsf{z}]$$

when this model is *correctly specified*.

Inference under correct specification

We might (optimistically) assume that the model $\mathbb{E}[Y|Z,X]$ is *correctly specified*, and captures the true relationship.

If this is, in fact, the case, then

No special techniques are needed to estimate the causal effect.

We may simply use *regression* of Y on (X,Z) using mean model $\mathbb{E}[Y|X,Z]$.

Inference under correct specification

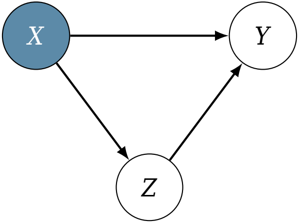
To estimate the APO, we simply set

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n} \sum_{i=1}^{n} \mu(X_i, z) \tag{3}$$

and derive other estimates from this: if $\mu(x, z)$ correctly captures the relationship of the outcome to the exposure and confounders, then the estimator of the APO in (3) is *consistent* (gives the correct answer as the sample size increase to infinity).

Inference under correct specification

By conditioning on X in the regression model, we *block* the indirect (confounding) path between Z and Y:



DAG with confounding path $Z \to X \to Y$ blocked by conditioning on X

Inference under correct specification

The third challenge of causal inference is that

correct specification cannot be guaranteed.

- we may not capture the relationship between Y and (Z, X) correctly.

Part 2

The Propensity Score

Constructing a balanced sample

Recall the randomized trial setting in the case of a binary exposure.

- we obtain a random sample of size n of individuals from the target population, and measure their X values;
- according to some random assignment procedure, we *intervene* to assign treatment Z to individuals, and measure their outcome Y;
- the link between X and Z is *broken* by the random allocation.

Recall that this procedure led to the valid use of the estimators of the ATE based on (1) and (2).

Constructing a balanced sample

The important feature of the randomized study is that we have, for confounders X (indeed all predictors)

$$f_{X|Z}(x|1) \equiv f_{X|Z}(x|0) \quad \text{for all } x,$$

or equivalently, in the case of a binary confounder,

$$\Pr[X = 1|Z = 1] = \Pr[X = 1|Z = 0].$$

The distribution of X is *balanced* across the two exposure groups; this renders direct comparison of the outcomes possible.

Probabilistically, X and Z are independent.

Constructing a balanced sample

In an *observational* study, there is a possibility that the two exposure groups are systematically *not balanced*

$$f_{X|Z}(x|1) \neq f_{X|Z}(x|0) \quad \text{for some } x,$$

or in the binary case

$$\Pr[X = 1|Z = 1] \neq \Pr[X = 1|Z = 0].$$

If X influences Y also, then this imbalance renders direct comparison of outcomes in the two groups impossible.

Constructing a balanced sample

Whilst *global* balance may not be present, it may be that '*local*' balance, within certain *strata* of the sample, may be present.

- Let $\mathcal{S}$ be some identified stratum in the sample space for X;
- suppose for $x \in \mathcal{S}$, we have *balance*; that is, within $\mathcal{S}$, X is independent of Z;

$$f_{X|Z:\mathcal{S}}(x|1 : x \in \mathcal{S}) = f_{X|Z}(x|0 : x \in \mathcal{S});$$

- for individuals who have X values in $\mathcal{S}$, there is the possibility of *direct comparison* of the treated and untreated groups.

We might then restrict attention to causal statements within stratum $\mathcal{S}$.

Constructing a balanced sample

Note 7.

In an extreme yet trivial case, consider a confounder X that takes only a single value, x_0 say, for all individuals.

Then it is clear that any systematic differences in outcomes *must* be due to exposure.

Constructing a balanced sample

For *discrete* confounders,

- we can consider defining strata where the X values are *precisely matched*,
- then compare the outcomes for treated and untreated individuals *within* those strata;
- we can then extend this comparison to *multiple* strata, and combine.

Constructing a balanced sample

Consider matching strata $\mathcal{S}_1, \ldots, \mathcal{S}_K$. We would then be able to compute the ATE by noting that

$$\mathbb{E}[Y(1) - Y(0)] = \sum_{k=1}^{K} \mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k] \Pr[X \in \mathcal{S}_k]$$

- $\mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k]$ may be estimated non-parametrically from the data by using (1) or (2) for data restricted to have $x \in \mathcal{S}_k$.
- $\Pr[X \in \mathcal{S}_k]$ may be estimated using the empirical proportion of x that lie in $\mathcal{S}_k$.

Constructing a balanced sample

For *continuous* confounders, we might consider the same strategy: consider matching strata $\mathcal{S}_1, \ldots, \mathcal{S}_K$. Then the formula

$$\mathbb{E}[Y(1) - Y(0)] = \sum_{k=1}^{K} \mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k] \Pr[X \in \mathcal{S}_k]$$

still holds. However

- we must assume a model for how $\mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k]$ varies with x for $x \in \mathcal{S}_k$.

In both cases, inference is restricted to the set of X space contained in

$$\bigcup_{k=1}^{K} \mathcal{S}_k.$$

Constructing a balanced sample

In the continuous case, the above calculations depend on the assumption that the treatment effect is similar for x values that lie '*close together*' in predictor (confounder) space. However

I. Unless we can achieve *exact* matching, then the term 'close together' needs careful consideration.

II. If X is *moderate* or *high-dimensional*, there may be insufficient data to achieve adequate matching to facilitate the estimation of

$$\mathbb{E}[Y(1) - Y(0) | X \in \mathcal{S}_k];$$

recall that we need a large enough sample of treated and untreated subjects in stratum $\mathcal{S}_k$.

Nevertheless, matching in this fashion is an important tool in causal comparison.

Balance via the propensity score

We now introduce the important concept of the propensity score that facilitates causal comparison via a balancing approach.

Recall that our goal is to mimic the construction of the randomized study that facilitates direct comparison between treated and untreated groups. We may not be able to achieve this globally, but possibly can achieve it locally in strata of X space.

The question is how to define these strata.

Balance via the propensity score

Recall that in the binary exposure case, balance corresponds to being able to state that within $\mathcal{S}$, X is *independent* of Z:

$$f_{X|Z:\mathcal{S}}(x|1 : x \in \mathcal{S}) = f_{X|Z}(x|0 : x \in \mathcal{S})$$

This can be achieved if $\mathcal{S}$ is defined in terms of a *statistic*, $\mathsf{e}(X)$[3] say. That is, we consider the conditional distribution

$$f_{X|Z,\mathsf{e}(X)}(x|z, e)$$

so that, *given* $\mathsf{e}(X) = e$, *Z is independent of X*, so that within strata of $\mathsf{e}(X)$, the treated and untreated groups are directly comparable.

[3] note the sans serif font $\mathsf{e}(.)$, distinct from e which indicates a numerical value.

Balance via the propensity score

For conditional independence, we require that

$$f_{X|Z,\mathsf{e}(X)}(x|z,e) = f_{Z|\mathsf{e}(X)}(z|e) \qquad \text{for all } x, z, e. \tag{4}$$

Now, as Z is binary, we must be able to write

$$f_{Z|\mathsf{e}(X)}(z|e) = p(e)^z(1-p(e))^{1-z} \qquad z \in 0, 1$$

where $p(e)$ is a probability, and a function of the fixed value e.

Balance via the propensity score

But $\mathsf{e}(X)$ is a function of X, so automatically we have that

$$f_{Z|X,\mathsf{e}(X)}(z|x,e) \equiv f_{Z|X}(z|x) \qquad \text{provided } e = \mathsf{e}(x).$$

Therefore, we require that

$$f_{Z|X}(z|x) = f_{Z|X,\mathsf{e}(X)}(z|x,e) = p(e)^z(1-p(e))^{1-z}$$

for all relevant z, x, with $e = \mathsf{e}(x)$.

Balance via the propensity score

This can be achieved by choosing the statistic[4]

$$\mathsf{e}(x) = f_{Z|X}(1|x) = \Pr_{Z|X}[Z = 1|X = x]$$

and setting $p(.)$ to be the identity function, so that

$$f_{Z|X}(z|x) = e^z(1-e)^{1-z} \quad z = 0, 1, e = \mathsf{e}(x).$$

The random variable $\mathsf{e}(X)$ defines the strata via which the causal calculation can be considered.

[4] Choosing $\mathsf{e}(x)$ to be some monotone transform of $f_{Z|X}(1|x)$ would also achieve the same balance.

Balance via the propensity score

The function $\mathsf{e}(x)$ defined in this way is the *propensity score*[5]. It has the following important properties:

(i) it is a balancing score; conditional on $\mathsf{e}(X)$, X and Z are independent;

(ii) it is a *scalar* quantity, irrespective of the dimension of X;

(iii) in noting that for balance we require that

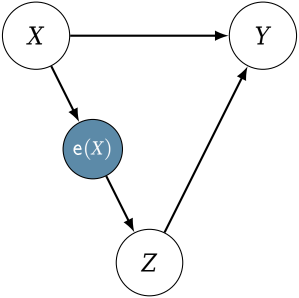
$$f_{Z|X}(z|x) \equiv f_{Z|\mathsf{e}(X)}(z|e),$$

the above construction demonstrates that if $\widetilde{\mathsf{e}}(X)$ is another balancing score, then $\mathsf{e}(X)$ is a function of $\widetilde{\mathsf{e}}(X)$;

- that is, $\mathsf{e}(X)$ is the '*coarsest*' balancing score.

[5] see Rosenbaum & Rubin (1983), Biometrika

Balance via the propensity score



DAG with confounding path $Z \rightarrow X \rightarrow Y$ blocked by conditioning on $\mathsf{e}(X)$

Evaluating the propensity score

To achieve balance we must ensure that

$$\mathsf{e}(X) = \Pr[Z = 1|X]$$

is *correctly specified*.

- If X comprises entirely *discrete* components, then we may be able to estimate $\Pr[Z = 1|X]$ entirely non-parametrically, and satisfactorily if the sample size is large enough.
- If X has *continuous* components, it is common to use parametric modelling, with
$$\mathsf{e}(X;\alpha) = \Pr[Z = 1|X;\alpha].$$
Balance then depends on *correct specification* of this model.

Unconfoundedness given the propensity score

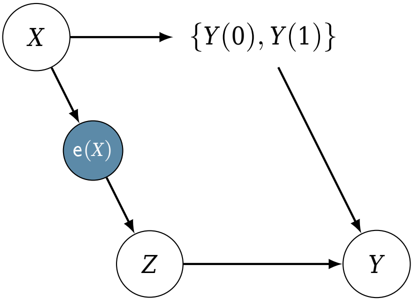
The assumption of ‘no unmeasured confounders’ amounts to assuming that the potential outcomes are jointly *independent* of exposure assignment given the confounders, that is

$$\{Y(0), Y(1)\} \perp Z \mid X$$

that is, in terms of densities

$$\begin{aligned} f_{Y(\mathsf{z}),Z|X}(y, z|x) &= f_{Y(\mathsf{z})|X}(y|x) f_{Z|X}(z|x) \\ &= f_{Y|Z,X}(y|\mathsf{z}, x) f_{Z|X}(z|x). \end{aligned}$$

Unconfoundedness given the propensity score



Directed Acyclic Graph (DAG) with potential outcomes and $\mathsf{e}(X)$

Unconfoundedness given the propensity score

We have by factorization that

$$f_{Y(\textsf{z}),Z|\mathsf{e}(X)}(y,z|e) = \frac{1}{f_{\mathsf{e}(X)}(e)} \int_{\mathcal{S}_e} f_{Y(\textsf{z}),Z,X}(y,z,x) \ \mathrm{d}x$$

where $\mathcal{S}_e$ is the set of x values

$$\mathcal{S}_e \equiv \{x : \mathsf{e}(x) = e\}$$

that yield a propensity score value equal to the value e.

Unconfoundedness given the propensity score

Now we have by unconfoundness given X that

$$f_{Y(\mathsf{z}),Z,X}(y,z,x) = f_{Y(\mathsf{z})|X}(y|x)f_{Z|X}(z|x)f_X(x)$$

and on the set $\mathcal{S}_e$, we have

$$f_{Z|X}(z|x) = e^z(1-e)^{1-z} \equiv f_{Z|\mathsf{e}(X)}(z|e).$$

Unconfoundedness given the propensity score

Therefore, recalling the $\mathcal{S}_e$ is defined via the fixed constant e,

$$\begin{aligned}\int_{\mathcal{S}_e} f_{Y(\mathsf{z}),Z,X}(y,z,x)\ \mathrm{d}x &= \int_{\mathcal{S}_e} f_{Y(\mathsf{z})|X}(y|x)e^z(1-e)^{1-z}f_X(x)\ \mathrm{d}x\\ &= e^z(1-e)^{1-z}\int_{\mathcal{S}_e} f_{Y(\mathsf{z})|X}(y|x)f_X(x)\ \mathrm{d}x\\ &= f_{Z|\mathsf{e}(X)}(z|e)f_{Y(\mathsf{z})|\mathsf{e}(X)}(y|e).\end{aligned}$$

Hence

$$f_{Y(\mathsf{z}),Z|\mathsf{e}(X)}(y,z|e) = \frac{1}{f_{\mathsf{e}(X)}(e)}f_{Z|\mathsf{e}(X)}(z|e)f_{Y(\mathsf{z})|\mathsf{e}(X)}(y|e)$$

and so

$$Y(\mathsf{z}) \perp Z \mid \mathsf{e}(X) \qquad \text{for all } \mathsf{z}.$$

Estimation using the propensity score

We now consider the same stratified estimation strategy as before, but using $\mathsf{e}(X)$ instead X to stratify.

Consider strata $\mathcal{S}_1, \ldots, \mathcal{S}_K$ defined via $\mathsf{e}(X)$. In this case, recall that

$$0 < \mathsf{e}(X) < 1$$

so we might consider an equal quantile partition, say using quintiles.

Then we have

$$\mathbb{E}[Y(1) - Y(0)] = \sum_{k=1}^{K} \mathbb{E}[Y(1) - Y(0) | \mathsf{e}(X) \in \mathcal{S}_k] \Pr[\mathsf{e}(X) \in \mathcal{S}_k]$$

still holds approximately if the $\mathcal{S}_k$ are small enough.

Estimation using the propensity score

This still requires us to be able to estimate

$$\mathbb{E}[Y(1) - Y(0) | \mathsf{e}(X) \in \mathcal{S}_k]$$

so we need a sufficient number of treated and untreated individuals with $\mathsf{e}(X) \in \mathcal{S}_k$ to facilitate the ‘direct comparison’ within this stratum.

If the expected responses are constant across the stratum, the formulae (1) and (2) may be used.

Matching

The derivation of the propensity score indicates that it may be used to construct *matched* individuals or groups that can be compared directly.

- if two individuals have *precisely the same value* of $\mathtt{e}(x)$, then they are exactly matched;
- if one of the pair is treated and the other untreated, then their outcomes can be *compared directly*, as any imbalance between their measured confounder values has been removed by the fact that they are matched on $\mathtt{e}(x)$;
- this is conceptually identical to the standard procedure of matching in two-group comparison.

Matching

For an exactly matched pair (i_1, i_0), treated and untreated respectively, the quantity

$$y_{i_1} - y_{i_0}$$

is an unbiased estimate of the ATE

$$\mathbb{E}[Y(1) - Y(0)];$$

more typically we might choose m such matched pairs, usually with different $\mathsf{e}(x)$ values across pairs, and use the estimate

$$\frac{1}{m}\sum_{i=1}^{m}(y_{i_1} - y_{i_0})$$

Matching

Exact matching is difficult to achieve, therefore we more commonly attempt to achieve approximate matching

- May match one treated to M untreated ($1 : M$ matching)
- caliper matching;
- nearest neighbour/kernel matching;
- matching with replacement.

Most standard software packages have functions that provide automatic matching using a variety of methods.

Beyond binary exposures

The theory developed above extends beyond the case of binary exposures.

Recall that we require *balance* to proceed with causal comparisons; essentially, with strata defined using X or $\mathtt{e}(X)$, the distribution of X should not depend on Z.

We seek a scalar statistic such that, conditional on the value of that statistic, X and Z are independent. In the case of general exposures, we must consider balancing scores that are functions of *both* Z and X.

Beyond binary exposures

For a balancing score $\mathsf{b}(Z, X)$[6], we require that

$$X \perp Z \mid \mathsf{b}(Z, X).$$

We denote $B = \mathsf{b}(Z, X)$ for convenience.

Consider the conditional distribution $f_{Z|X,B}(z|x, b)$: we wish to demonstrate that

$$f_{Z|X,B}(z|x, b) = f_{Z|B}(z|b) \qquad \text{for all } z, x, b.$$

That is, we require that B completely characterizes the conditional distribution of Z given X.

[6] note the sans serif font $\mathsf{b}(.)$, distinct from b which indicates a numerical value.

Beyond binary exposures

This can be achieved by choosing the statistic

$$\mathsf{b}(z, x) = f_{Z|X}(z|x)$$

in line with the choice in the binary case.

The balancing score defined in this way is termed the

Generalized Propensity Score

which is a balancing score for general exposures.

Beyond binary exposures

Note, however, that this choice that mimics the binary exposure case is not the only one that we might make. The requirement

$$f_{Z|X,B}(z|x,b) = f_{Z|B}(z|b)$$

for all relevant z, x is met if we define $\mathsf{b}(Z,X)$ to be *any* sufficient statistic that characterizes the conditional distribution of Z given X.

It may be possible, for example, to choose functions purely of X.

Beyond binary exposures

Example: Normally distributed exposures

Suppose that continuous valued exposure Z is distributed as

$$Z|X = x \sim \text{Normal}(x\alpha, \sigma^2)$$

for row-vector confounder X. We have that

$$f_{Z|X}(z|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(z - x\alpha)^2\right\}$$

Beyond binary exposures

Example: Normally distributed exposures

We might therefore choose

$$\mathsf{b}(Z, X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Z - X\alpha)^2\right\}.$$

However, the linear predictor

$$\mathsf{b}(X; \alpha) = X\alpha$$

also characterizes the conditional distribution of Z given X; if we know that $x\alpha = b$, then

$$Z|X = x \equiv Z|B = b \sim \text{Normal}(b, \sigma^2).$$

In both cases, parameters α are to be estimated.

Beyond binary exposures

The generalized propensity score inherits all the properties of the standard propensity score;

- it induces balance;
- if the potential outcomes and exposure are independent given X under the unconfoundeness assumption, they are also independent given $\mathsf{b}(Z, X)$.

However, how exactly to use the generalized propensity score in causal adjustment for continuous exposures is not clear.

Propensity Score Regression

Up to this point we have considered using the propensity score for stratification, that is, to produce directly comparable groups of treated and untreated individuals.

Causal comparison can also be carried out using regression techniques: that is, we consider building an estimator of the APO by *regressing* the outcome on a function of the exposure and the propensity score.

Regressing on the propensity score is a means of controlling the confounding.

Propensity Score Regression

If we construct a model

$$\mathbb{E}[Y|X = x, Z = z, \mathsf{b}(Z, X) = b] = \mu(x, z, b)$$

then by the unconfoundedness result that

$$\mathbb{E}[Y(\mathsf{z})] = \mathbb{E}_X[\mathbb{E}[Y|X, Z = \mathsf{z}, \mathsf{b}(\mathsf{z}, X)] = \mathbb{E}_X[\mu(X, \mathsf{z}, \mathsf{b}(\mathsf{z}, X))].$$

Propensity Score Regression

That is, to estimate the APO, we might

- fit the propensity model $\mathsf{b}(Z,X)$ by regressing Z on X;
- fit the conditional outcome model $\mu(x,z,b)$ using the fitted values $\widehat{\mathsf{b}}(z_i,x_i)$;
- for each z of interest, estimate the APO by

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}(x_i,\mathsf{z},\widehat{\mathsf{b}}(\mathsf{z},x_i)).$$

If, more simply, we have $\mathsf{b}(Z,X)\equiv\mathsf{b}(X)$ (as for the propensity score) we proceed as above.

Propensity Score Regression

Example: Binary exposure

- $\mathsf{e}(x;\alpha) = \Pr[Z = 1|X = x;\alpha]$ then regress Z on X to obtain $\widehat{\alpha}$ and fitted values $\widehat{\mathsf{e}}(x) \equiv \mathsf{e}(x;\widehat{\alpha})$.
- $\mathbb{E}[Y|X = x, Z = z, \mathsf{e}(X) = e;\beta] = \mu(x, z, e;\beta)$ and estimate this model by regressing y_i on z_i and $e_i = \widehat{\mathsf{e}}(x_i)$.

  For example, we might have that

$$\mathbb{E}[Y|X_i = x_i, Z = z_i, \mathsf{e}(X_i) = e_i;\beta] = \beta_0 + \beta_1 z_i + \beta_2 e_i.$$

We then average the model predictions to obtain the APO estimate

$$\widehat{\mathbb{E}}[Y(\mathsf{z})] = \frac{1}{n}\sum_{i=1}^{n} \mu(x_i, \mathsf{z}, \widehat{\mathsf{e}}(x_i);\widehat{\beta}).$$

Propensity Score Regression

Example: Continuous exposure

We propose a parametric probability density for the exposure

$$\mathsf{b}(z, x; \alpha) = f_{Z|X}(z|x; \alpha)$$

for which we estimate α by regressing Z on X to obtain $\widehat{\alpha}$ and fitted values $\widehat{\mathsf{b}}(z, x) \equiv \mathsf{b}(z, x; \widehat{\alpha})$. Then we specify

$$\mathbb{E}[Y|X = x, Z = z, \mathsf{b}(X, Z) = b; \beta] = \mu(x, z, b; \beta)$$

and estimate this model by regressing y on z and $\widehat{\mathsf{b}}(z, x)$.

For example,

$$\mathbb{E}[Y|X_i = x_i, Z = z_i, \mathsf{b}(z_i, x_i) = b_i; \beta] = \beta_0 + \beta_1 z_i + \beta_2 b_i.$$

Propensity Score Regression

Example: Continuous exposure

We then compute the predictions under this model, and average them to obtain the APO estimate

$$\widehat{\mathbb{E}}[Y(\mathsf{z})] = \frac{1}{n}\sum_{i=1}^{n} \mu(x_i, \mathsf{z}, \widehat{\mathsf{b}}(\mathsf{z}, x_i); \widehat{\beta}).$$

Note that here the propensity terms that enter into μ are computed at the target z values

not the observed exposure values.

Propensity Score Regression

These procedures require us to make two modelling choices:

- the propensity model, $\mathsf{b}(z, x)$ or $\mathsf{b}(x)$;
- the outcome mean model $\mu(x, z, b)$.

For consistent inference for the ATE, we need

- the propensity model, *and*
- the dependence of the outcome mean model on z

to be correctly specified.

Propensity Score Regression

Example: Binary exposure

Suppose that the true (data generating) model can be written

$$\mathbb{E}[Y|X = x, Z = z] = \widetilde{\mu}(x, z) = \widetilde{\mu}_0(x) + z\widetilde{\mu}_1(x).$$

Then the propensity score regression model

$$\mathbb{E}[Y|X = x, Z = z, \mathsf{b}(X) = b] = \mu_0(x) + z\widetilde{\mu}_1(x) + b\widetilde{\mu}_1(x)$$

is sufficient to give consistent estimation of the ATE

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\widetilde{\mu}_1(X)].$$

Propensity Score Regression

Example: Binary exposure

That is, we may mis-specify the component

$$\widetilde{\mu}_0(x)$$

provided we correctly specify the propensity model $\mathsf{b}(x)$.

Weighting approaches

We focus on the APO

$$\mathbb{E}[Y(z)] = \int y\, f_{Y(z),X}(y, x)\ \mathrm{d}y\, \mathrm{d}x$$

and utilize the propensity model in a different fashion;

- instead of accounting for confounding by balancing through matching or regression, we aim to achieve balance via *weighting*

Average potential outcome

Recall that intervening to set $Z = \mathsf{z}$ leads to the calculation

$$\mathbb{E}[Y(\mathsf{z})] = \int y \mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y(\mathsf{z}),X}(y, x) \ \mathrm{d}y \ \mathrm{d}z \ \mathrm{d}x.$$

We take a random sample from the population with density

$$\mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y(\mathsf{z}),X}(y, x) \equiv \mathbb{1}_{\{\mathsf{z}\}}(z)\, f_{Y|Z,X}(y|z, x) f_X(x).$$

and construct the usual estimator

$$\widehat{\mathbb{E}}[Y(\mathsf{z})] = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

as $Z_i = \mathsf{z}$ for all i.

Average potential outcome: Experimental study

In a randomized (*experimental*) study, suppose that exposure $Z = z$ is assigned with probability determined by $f_Z(z)$.

Then we have the estimators

$$\widehat{\mathbb{E}}[Y(z)] = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i)Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i)} \quad \text{or} \quad \widehat{\mathbb{E}}[Y(z)] = \frac{1}{nf_Z(z)} \sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i)Y_i.$$

Average potential outcome: Observational study

Denote by $P_{\mathcal{E}}$ the probability distribution for samples drawn under the *experimental* design corresponding to the density

$$f^{\mathcal{E}}_{Y|Z,X}(y|z,x)f^{\mathcal{E}}_X(x)f^{\mathcal{E}}_Z(z).$$

If the data arise from the *observational* (non-experimental) distribution $P_{\mathcal{O}}(\,\mathrm{d}y,\,\mathrm{d}z,\,\mathrm{d}x)$. We have by the *importance sampling* argument

$$\begin{aligned}\mathbb{E}[Y(\mathsf{z})] &= \frac{1}{f^{\mathcal{E}}_Z(\mathsf{z})}\int y\mathbb{1}_{\{\mathsf{z}\}}(z)\,P_{\mathcal{E}}(\,\mathrm{d}y,\,\mathrm{d}z,\,\mathrm{d}x)\\ &= \frac{1}{f^{\mathcal{E}}_Z(\mathsf{z})}\int y\mathbb{1}_{\{\mathsf{z}\}}(z)\,\underbrace{\frac{P_{\mathcal{E}}(\,\mathrm{d}y,\,\mathrm{d}z,\,\mathrm{d}x)}{P_{\mathcal{O}}(\,\mathrm{d}y,\,\mathrm{d}z,\,\mathrm{d}x)}}_{①}P_{\mathcal{O}}(\,\mathrm{d}y,\,\mathrm{d}z,\,\mathrm{d}x).\end{aligned}$$

Average potential outcome: Observational study

In terms of densities $\textcircled{1}$ becomes

$$\frac{f^{\mathcal{E}}_{Y|Z,X}(y|z,x)f^{\mathcal{E}}_{Z}(z)f^{\mathcal{E}}_{X}(x)}{f^{\mathcal{O}}_{Y|Z,X}(y|z,x)f^{\mathcal{O}}_{Z|X}(z|x)f^{\mathcal{O}}_{X}(x)} = \frac{f^{\mathcal{E}}_{Y|Z,X}(y|z,x)}{f^{\mathcal{O}}_{Y|Z,X}(y|z,x)} \times \frac{f^{\mathcal{E}}_{Z}(z)}{f^{\mathcal{O}}_{Z|X}(z|x)} \times \frac{f^{\mathcal{E}}_{X}(x)}{f^{\mathcal{O}}_{X}(x)}$$

- for the first term, we have that

  $$\frac{f^{\mathcal{E}}_{Y|Z,X}(y|z,x)}{f^{\mathcal{O}}_{Y|Z,X}(y|z,x)} = 1 \qquad \text{for all } y, z, x;$$

  under the *no unmeasured confounders* assumption.
- the third term equals 1 by assumption.

Experimental vs observational sampling

The second term

$$\frac{f_Z^{\mathcal{E}}(z)}{f_{Z|X}^{\mathcal{O}}(z|x)}$$

constitutes a *weight* that appears in the integral that yields the desired APO; the term

$$\frac{1}{f_{Z|X}^{\mathcal{O}}(z|x)}$$

accounts for the *imbalance* that influences the confounding and measures the difference between the *observed* sample and a hypothetical idealized *randomized* sample.

Estimation

This suggests the (non-parametric) estimators

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_i)Y_i}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)} \tag{IPW0}$$

which is unbiased, or

$$\widehat{\mathbb{E}}[Y(z)] = \frac{\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_i)Y_i}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}}{\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}} \tag{IPW}$$

which is consistent, each provided $f^{\mathcal{O}}_{Z|X}(.|.)$ *correctly specifies* the conditional density of Z given X for all (z, x).

Inverse weighting and the propensity score

Note 8.

Inverse weighting constructs a pseudo-population in which there are no imbalances on confounders between the exposure groups. The pseudo-population is balanced, as required for direct comparison of treated and untreated groups.

Note 9.

The term in the denominator, $f^{\mathcal{O}}_{Z|X}(z_i|x_i)$, is the *exposure model*. If Z_i is binary, this essentially reduces to

$$\mathsf{e}(x_i)^{z_i}(1 - \mathsf{e}(x_i))^{1-z_i}$$

where $\mathsf{e}(.)$ is the propensity score as defined previously.

Positivity

Note 10.

We must have

$$f^{\mathcal{O}}_{Z|X}(z|x) > 0$$

for all x, z.

This is termed the *positivity* assumption or

experimental treatment assignment

assumption.

Estimation via augmentation

We may write

$$\mathbb{E}[Y(z)] = \mathbb{E}[Y(z) - \mu(X, z)] + \mathbb{E}[\mu(X, z)]$$

where $\mu(x, z) = \mathbb{E}[Y|X = x, Z = z]$.

We then have the alternate estimator

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_i)(Y_i - \mu(X_i, Z_i))}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)} + \frac{1}{n}\sum_{i=1}^{n} \mu(X_i, z) \quad \text{(AIPW)}$$

Estimation via augmentation

Then, if both

$$f_{Z|X}^{\mathcal{O}}(z|x) \qquad \text{and} \qquad \mu(x,z)$$

are correctly specified, we have

$$\mathrm{Var}_{\mathrm{AIPW}} \leq \mathrm{Var}_{\mathrm{IPW}}.$$

Furthermore, (AIPW) is *doubly robust*

- *consistent* even if one of $f_{Z|X}^{\mathcal{O}}(z|x)$ and $\mu(x,z)$ is *mis-specified*.

Properties under mis-specification

Suppose that, in reality, the correct specifications are

$$\widetilde{f}_{Z|X}(z|x) \qquad \widetilde{\mu}(x,z).$$

Then the *bias* of (AIPW) is

$$\mathbb{E}\left[\frac{(f^{\mathcal{O}}_{Z|X}(\mathsf{z}|X) - \widetilde{f}_{Z|X}(\mathsf{z}|X))(\mu(X,\mathsf{z}) - \widetilde{\mu}(X,\mathsf{z}))}{f^{\mathcal{O}}_{Z|X}(\mathsf{z}|X)}\right] \tag{5}$$

which is zero if

$$f^{\mathcal{O}}_{Z|X} \equiv \widetilde{f}_{Z|X} \qquad \text{or} \qquad \mu(x,z) \equiv \widetilde{\mu}(x,z).$$

Properties under mis-specification

Asymptotically, for estimators that are sample averages, the variance of the estimator converges to zero under standard conditions.

Therefore in large samples it is the magnitude of the bias as given by (5) that determines the quality of the estimator.

- equation (5) demonstrates that mis-specification in the functions $\mu(x, z)$ and $f^{\mathcal{O}}_{Z|X}$ play equal roles in the bias.

Parametric modelling: two-stage approach

In the formulation, parametric models

$$f^{\mathcal{O}}_{Z|X}(z|x;\alpha) \qquad \mu(x,z;\beta)$$

are typically used.

Parameters (α,β) are estimated from the observed data by regressing

- Stage I: Z on X using $(z_i,x_i), i=1,\ldots,n$,
- Stage II: Y on (Z,X) using $(y_i,z_i,x_i), i=1,\ldots,n$

and using plug-in version of (IPW) and (AIPW).

The estimated propensity score

Note 11.

It is possible to conceive of situations where the propensity-type model

$$f^{\mathcal{O}}_{Z|X}(z|x) \qquad \text{or} \qquad f^{\mathcal{O}}_{Z|X}(z|x;\alpha)$$

is known precisely and does not need to be estimated.

This is akin to the randomized study where the allocation probabilities are fixed by the experimenter. It can be shown that using *estimated* quantities

$$\widehat{f}^{\mathcal{O}}_{Z|X}(z|x) \qquad \text{or} \qquad f^{\mathcal{O}}_{Z|X}(z|x;\widehat{\alpha})$$

yields *lower variances* for the resulting estimators than if the *known* quantities are used.

Alternative view of augmentation

We may write the estimating equation yielding (AIPW) as

$$\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}(Y_i - \mu(X_i, Z_i)) + \sum_{i=1}^{n} \{\mu(X_i, z) - \mu(z)\} = 0$$

Alternative view of augmentation

The first summation is a component of the score obtained when performing OLS regression for Y with mean function

$$\mu(x, z) = \mu_0(x, z) + \epsilon \frac{\mathbb{1}_{\{\mathsf{z}\}}(z)}{f^{\mathcal{O}}_{Z|X}(z|x)}$$

and $\mu_0(x, z)$ is a conditional mean model, and ϵ is a regression coefficient associated with the derived predictor

$$\frac{\mathbb{1}_{\{\mathsf{z}\}}(z)}{f^{\mathcal{O}}_{Z|X}(z|x)}.$$

Alternative view of augmentation

Therefore, an estimator equivalent to (AIPW) can be obtained by regressing Y on (X, Z) for fixed z using $\mu(x, z)$, and forming the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mu_0(X_i, Z_i) + \widehat{\epsilon}\frac{\mathbb{1}_{\{\mathsf{z}\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}\right\}.$$

In a parametric model setting, this becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\mu_0(X_i, Z_i; \widehat{\beta}) + \widehat{\epsilon}\frac{\mathbb{1}_{\{\mathsf{z}\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i; \widehat{\alpha})}\right\}$$

where α is estimated from Stage (I), and β is estimated along with ϵ in the secondary regression.

Augmentation and contrasts

The equivalent to (AIPW) for estimating the ATE for binary treatment

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

is merely $\widehat{\mathbb{E}}[Y(1)] - \widehat{\mathbb{E}}[Y(0)]$ or

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\mathbb{1}_1(Z_i)}{f^{\mathcal{O}}_{Z|X}(1|X_i)} - \frac{\mathbb{1}_0(Z_i)}{f^{\mathcal{O}}_{Z|X}(0|X_i)}\right](Y_i - \mu(X_i, Z_i)) + \frac{1}{n}\sum_{i=1}^{n}\delta(X_i)$$

where

$$\delta(x) = \mu(x,1) - \mu(x,0).$$

Augmentation and contrasts

Therefore we can repeat the above argument and base the contrast estimator on the regression of Y on (X, Z) using the mean specification

$$\mu(x,z) = \mu_0(x,z) + \epsilon \left[\frac{\mathbb{1}_1(z)}{f^{\mathcal{O}}_{Z|X}(1|x)} - \frac{\mathbb{1}_0(z)}{f^{\mathcal{O}}_{Z|X}(0|x)} \right]$$

or

$$\mu(x,z) = \mu_0(x,z) + \left[\epsilon_1 \frac{\mathbb{1}_1(z)}{f^{\mathcal{O}}_{Z|X}(1|x)} - \epsilon_0 \frac{\mathbb{1}_0(z)}{f^{\mathcal{O}}_{Z|X}(0|x)} \right].$$

Part 3

Implementation and Computation

Statistical modelling tools

Causal inference typically relies on reasonably standard statistical tools:

1. **Standard distributions:**
   - Normal;
   - Binomial;
   - Time-to-event distributions (Exponential, Weibull etc.)
2. **Regression tools:**
   - linear model/ordinary least squares;
   - generalized linear model, typically linear regression;
   - survival models.

Pooled logistic regression

For a survival outcome, *pooled logistic regression* is often used.

The usual continuous survival time outcome is replaced by a discrete, binary outcome;

- this is achieved by partitioning the outcome space into short intervals,

  $$(0, t_1], (t_1, t_2], \ldots$$

  and assuming that the failure density is approximately constant in each interval.

- using a hazard parameterization, we have that

  $$\Pr[\text{Failure in } (t_{j-1}, t_j] | \text{No failure before } t_{j-1}] = q_j$$

  which converts each single failure time outcome into a series of binary responses, with 0 recording ‘no failure’ and 1 recording ‘failure’.

Semiparametric estimation

Semiparametric models based on *estimating equations* are typically used:

- such models make no parametric assumptions about the distributions of the various quantities, but instead make moment restrictions;
- resulting estimators inherit good asymptotic properties;
- variance of estimators typically estimated in a ‘robust’ fashion using the sandwich estimator of the asymptotic variance.

Key considerations

In light of the previous discussions, in order to facilitate causal comparisons, there are several key considerations that practitioners must take into account.

1. **The importance of no unmeasured confounding.**

   When considering the study design, it is essential for valid conclusions to have measured and recorded all confounders.

Key considerations

2. **Model construction for the outcome regression.**
    - ideally, the model for the expected value of Y given Z and X, $\mu(x, z)$, should be correctly specified, that is, correctly capture the relationship between outcome and the other variables.
    - if this can be done, then no causal adjustments are necessary.
    - conventional model building techniques (variable selection) can be used; this will prioritize predictors of outcome and therefore will select all confounders;
    - however, in finite sample, this method may omit weak confounders that may lead to bias.

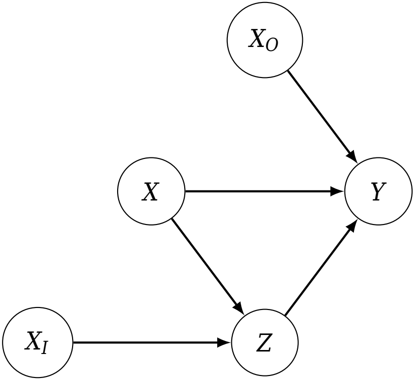
Key considerations

3. **Model construction for the propensity score.**

   Ideally, the model for the (generalized) propensity score, $\mathsf{e}(x)$ or $\mathsf{b}(z, x)$, should be correctly capture the relationship between the exposure and the confounders. We focus on

   - identifying the *confounders*;
   - *ignoring* the *instruments*: instruments do not predict the outcome, therefore cannot be a source of bias (unless there is unmeasured confounding) - however they can increase the variability of the resulting propensity score estimators.
   - the need for the specified propensity model to induce *balance*;
   - ensuring *positivity*: strata constructed from the propensity score must have sufficient data within them to facilitate comparison;
   - effective model selection.

Key considerations



DAG with predictors classified by their effects.

X are *confounders*; X_I are *instruments*; X_O are *pure predictors of outcome*.

Key considerations

Note 12.

Conventional model selection techniques (stepwise selection, selection via information criteria, sparse selection) *should not be used* when constructing the propensity score.

This is because such techniques prioritize the accurate prediction of exposure conditional on the other predictors; however, this is *not* the goal of the analysis.

These techniques may merely select strong instruments and omit strong predictors of outcome that are only weakly associated with exposure.

Key considerations

Note 13.

An apparently conservative approach is to build rich (highly parameterized) models for both $\mu(x, z)$ and $\mathsf{e}(x)$.

This approach prioritizes

bias elimination

at the cost of

variance inflation.

Key considerations

4. **The required measure of effect.**
   Is the causal measure required
   - a risk difference ?
   - a risk ratio ?
   - an odds ratio ?
   - an ATT, ATE or APO ?

Key considerations

Example: NHANES Analysis

See knitr sheet.

Example: Simulation study

Comparison of different adjustment methods.

Part 4

Extensions

Longitudinal studies

It is common for studies to involve multiple longitudinal measurements of exposure, confounders and outcomes.

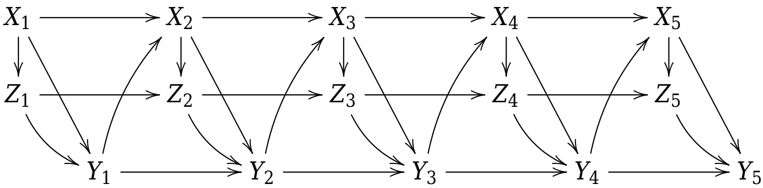
In this case, the possible effect of confounding of the exposure effect by the confounders is more complicated.

Furthermore, we may be interested in different types of effect:

- the *direct* effect: the effect of exposure in any given interval on the outcome in that interval, or the final observed outcome;
- the *total* effect: the effect of exposure aggregated across intervals final observed outcome;

Illustration

Possible structure across five intervals:



Mediation and time-varying confounding

- The effect of exposure on later outcomes may be *mediated* through variables measured at intermediate time points
  - for example, the effect of exposure Z_1 may have a direct effect on Y_1 that is confounded by X_1; however, the effect of Z_1 on Y_2 may also be non-negligible. This effect is mediated via X_2.

- There may be *time-varying* confounding;

Multivariate versions of the propensity score

The propensity score may be generalized to the multivariate setting. We consider for $j = 1, \ldots, m$,

- exposure: $\widetilde{Z}_{ij} = (Z_{i1}, \ldots, Z_{ij})$;
- outcome: $\widetilde{Y}_{ij} = (Y_{i1}, \ldots, Y_{ij})$;
- confounders: $\widetilde{X}_{ij} = (X_{i1}, \ldots, X_{ij})$.

Sometimes the notation

$$Z_{1:m} = (Z_1, \ldots, Z_m)$$

will be useful.

Multivariate versions of the propensity score

We consider vectors of potential outcomes corresponding to these observed quantities, that is, we consider a potential sequence of interventions up to time j

$$\widetilde{\mathsf{z}}_{ij} = (\mathsf{z}_{i1}, \ldots, \mathsf{z}_{ij})$$

and then the corresponding sequence of potential outcomes

$$\widetilde{Y}(\widetilde{\mathsf{z}}_{ij}) = (Y(\mathsf{z}_{i1}), \ldots, Y(\mathsf{z}_{ij})).$$

Multivariate versions of the propensity score

We define the *multivariate (generalized) propensity score* by

$$\mathsf{b}_j(z,x) = f_{Z_j|X_j,\widetilde{Z}_{j-1},\widetilde{X}_{j-1}}(z|x,\widetilde{z}_{j-1},\widetilde{x}_{j-1})$$

that is, using the conditional distribution of exposure at interval j, given the confounder at interval j, and the historical values of exposures and confounders.

Under the sequential generalizations of the *no unmeasured confounders* and *positivity* assumptions, this multivariate extension of the propensity score provides the required balance, and provides a means of estimating the *direct effect* of exposure.

The use of mixed models

The multivariate generalization above essentially builds a joint model for the sequence of exposures, and embeds this in a full joint distribution for all measured variables.

An alternative approach uses *mixed* (or *random effect*) models to capture the joint structure.

- such an approach is common in longitudinal data analysis;
- here we consider building a model for the longitudinal exposure data that encompasses a random effect.

The use of mixed models

Suppose first we have a continuous exposure: we consider the mixed effect model where for time point j

$$Z_{ij} = \widetilde{X}_{ij}\alpha + \widetilde{Z}_{i,j-1}\vartheta + \xi_i + \epsilon_{ij}$$

where

- $\widetilde{X}_{ij}\alpha$ captures the fixed effect contribution of past and current confounders;
- $\widetilde{Z}_{i,j-1}\vartheta$ captures the fixed effect contribution of past exposures;
- ξ_i is a subject specific *random effect*;
- ϵ_{ij} is a residual error.

The use of mixed models

The random effect ξ_i helps to capture unmeasured time-invariant confounding.

The distributional assumption made about ϵ_{ij} determine the precise form of a generalized propensity score that can again be used to estimate the direct effect of exposure.

The use of mixed models

For binary or other discrete exposures, the random effect model is built on the linear predictor scale, with say

$$\eta_{ij} = \widetilde{X}_{ij}\alpha + \widetilde{Z}_{i,j-1}\vartheta + \xi_i$$

determining the required conditional mean for the exposure at interval j.

Full-likelihood based inference may be used, but also generalized estimating approaches may be developed.

Estimation of Total Effects

The estimation of the total effect of exposure is more complicated as the need to acknowledge mediation and time-varying confounding renders standard likelihood-based approaches inappropriate.

The *Marginal Structural Model* is a semiparametric inverse weighting methodology designed to estimate total effects of functions of aggregate exposures that generalizes conventional inverse weighting.

The Marginal Structural Model

We observe for each individual i a sequence of exposures

$$Z_{i1}, Z_{i2}, \ldots, Z_{im}$$

and confounders

$$X_{i1}, X_{i2}, \ldots, X_{im}$$

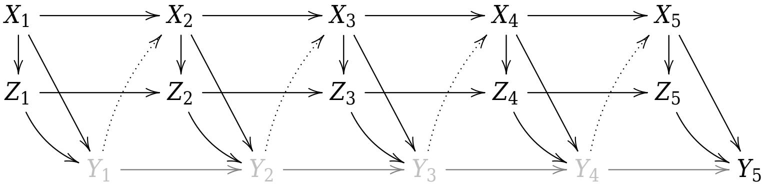
along with outcome $Y_i \equiv Y_{im}$ measured at the end of the study.

Intermediate outcomes $Y_{i1}, Y_{i2}, \ldots, Y_{i,m-1}$ also possibly available.

We might also consider individual level *frailty* variables $\{\upsilon_i\}$, which are determinants of both the outcome and the intermediate variables, but can be assumed conditionally independent of the exposure assignments.

The Marginal Structural Model

For example, with $m = 5$:



Common example: pooled logistic regression

- discrete time survival outcome
- outcome is binary, intermediate outcomes monotonic
- length of follow-up is random, or event time is censored.

The Marginal Structural Model

We seek to quantify the causal effect of exposure pattern

$$\widetilde{z} = (z_1, z_2, \cdots, z_m)$$

on the outcome. If the outcome is binary, we might consider[7]

$$\log\left(\frac{f(Y_{im} = 1|\widetilde{z}; \theta)}{f(Y_{im} = 0|\widetilde{z}; \theta)}\right) = \theta_0 + \theta_1 \sum_{j=1}^{m} z_j$$

as the *true* (structural) model. Note that this is a *marginal* model.

[7] We might also consider structural models in which the influence of covariates/confounders is recognized.

The Marginal Structural Model

However, this model is expressed for data presumed to be collected under an *experimental* design, $\mathcal{E}$.

In reality, it is necessary to adjust for the influence of

- *time-varying confounding* due to the observational nature of exposure assignment
- *mediation* as past exposures may influence future values of the confounders, exposures and outcome.

The adjustment can be achieved using *inverse weighting* via a *marginal structural model*.

The Marginal Structural Model

Causal parameter θ may be estimated via the weighted pseudo-likelihood

$$\mathcal{L}(\theta; \tilde{x}, y, \tilde{z}, \gamma, \alpha) \equiv \prod_{i=1}^{n} f(y_i \mid \tilde{z}_i; \theta)^{w_i},$$

where

$$w_i = \frac{\prod_{j=1}^{m} f(z_{ij} \mid \tilde{z}_{i(j-1)}; \alpha_j)}{\prod_{j=1}^{m} f(z_{ij} \mid \tilde{z}_{i(j-1)}, \tilde{x}_{ij}; \gamma_j)}$$

defines *stabilized* inverse weights.

The Marginal Structural Model: The logic

- Inference is required under *hypothetical* population $\mathcal{E}$;
  - in population $\mathcal{E}$, the *conditional independence* $z_{ij} \perp \tilde{x}_{ij} \mid \tilde{z}_{i(j-1)}$ holds true.
- Samples from *observational* population $\mathcal{O}$ are available.
- The weights w_i convey information on how much $\mathcal{O}$ resembles $\mathcal{E}$: this information is contained in the parameters γ.
- $\mathcal{E}$ has the *same marginal exposure assignment distribution* as $\mathcal{O}$.
- Inference using the weighted likelihood typically proceeds using robust (sandwich) variance estimation, or the bootstrap.

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

Antiretroviral therapy (ART) has reduced morbidity and mortality due to nearly all HIV-related illnesses, apart from mortality due to end-stage liver disease, which has increased since ART treatment became widespread.

In part, this increase may be due to improved overall survival combined with Hepatitis C virus (HCV) associated hepatic liver fibrosis, the progress of which is accelerated by immune dysfunction related to HIV-infection.

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

The Canadian Co-infection Cohort Study is one of the largest projects set up to study the role of ART on the development of end-stage liver disease in HIV-HCV co-infected individuals.

Given the importance of ART in improving HIV-related immunosuppression, it is hypothesized that liver fibrosis progression in co-infected individuals may be partly related to adverse consequences of ART interruptions.

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

Study comprised

- $N = 474$ individuals with at least one follow-up visit (scheduled at every six months) after the baseline visit,
- 2066 follow-up visits in total (1592 excluding the baseline visits).
- The number of follow-up visits m_i ranged from 2 to 16 (median 4).

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

We adopt a *pooled logistic regression* approach:

- a single binary outcome (death at study termination)
- longitudinal binary exposure (adherence to ART)
- possible confounders
  - *baseline covariates:* female gender, hepatitis B surface antigen (HBsAg) test and baseline APRI, as well as
  - *time-varying covariates:* age, current intravenous drug use (binary), current alcohol use (binary), duration of HCV infection, HIV viral load, CD4 cell count, as well as ART interruption status at the previous visit.
- need also a model for informative censoring.

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

- Analysis includes co-infected adults who were not on HCV treatment and did not have liver fibrosis at baseline.
- The outcome event was defined as aminotransferase-to-platelet ratio index (APRI), a surrogate marker for liver fibrosis, being at least 1.5 in any subsequent visit.
- Included visits where the individuals were either on ART or had interrupted therapy ($Z_{ij} = 1$), based on self-reported medication information, during the 6 months before each follow-up visit.

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

- Individuals suspected of having spontaneously cleared their HCV infection (based on two consecutive negative HCV viral load measurements) were excluded as they are not considered at risk for fibrosis progression.
- In the treatment assignment model all time-varying covariates (x_{ij}), including the laboratory measurements (HIV viral load and CD4 cell count), were lagged one visit.
- Individuals starting HCV medication during the follow-up were censored.

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

We considered the structural model

$$\log\left(\frac{f(Y_{ij}=1|\widetilde{\mathsf{z}}_{ij};\theta)}{f(Y_{ij}=0|\widetilde{\mathsf{z}}_{ij};\theta)}\right)=\theta_0+\theta_1\mathsf{z}_j$$

θ_1 measures the total effect of exposure in the most recent interval, allowing for mediation.

Real Data Example

Example: ART interruption in HIV/HCV co-infected individuals

Results:

Estimator	$\hat{\theta}_1$	SE	z
Unadjusted	4.616	0.333	13.853
MSM	0.354	0.377	0.937
Bootstrap	0.308	0.395	0.780

After adjustment for confounding and effects of mediation, we can conclude that the marginal effect of exposure is *non-significant*.

Part 5

New Challenges and Approaches

New challenges

The main challenge for causal adjustments using the propensity score is the nature of the observational data being recorded.

The data sets and databases being collected are increasingly complex and typically originate from different sources. The benefits of 'Big Data' come with the costs of more involved computation and modelling.

There is always an important trade off between the sample size *n* and the dimension of the confounder (and predictor) set.

Examples

- pharmacoepidemiology;
- electronic health records and primary care decision making;
- real-time health monitoring;

Data synthesis and combination

For observational databases, the choice of inclusion/exclusion criteria for analysis can have profound influence on the ultimate results:

- different databases can lead to different conclusions for the same effect of interest purely because of the methodology used to construct the raw data, irrespective of modelling choices.
- the key task of the statistician is to report uncertainty in a coherent fashion, ensuring that all sources of uncertainty are reflected. This should include uncertainty introduced due to lack of compatibility of data sources.

Classic challenges

Modern quantitative health research also has conventional challenges:

- *missing data*: many causal procedures are adapted forms of procedures developed for handling *informative missingness* (especially inverse weighting);
- *length-bias and left truncation in prevalent case studies*: selection of prevalent cases is also a form of ‘selection bias’ that causes bias in estimation if unadjusted;
- *non-compliance*: in randomized and observational studies there is the possibility of non- or partial compliance which is again a potential source of selection bias.

The Bayesian version

The *Bayesian* paradigm also provides a framework for decision-making under uncertainty.

Much of the reasoning on causal inference, and many of the modelling choices we must make for causal comparison and adjustment, are identical under Bayesian and classical (frequentist, semiparametric) reasoning.

The advantages of Bayesian thinking

With increasingly complex data sets in high dimensions, Bayesian methods can be useful as they

- provide a means of informed and coherent decision making in the presence of uncertainty;
- yield interpretable variability estimates in finite sample at the cost of interpretable modelling assumptions;
- allow the statistician to impose structure onto the inference problem that is helpful when information is sparse;
- naturally handle prediction, hierarchical modelling, data synthesis, and missing data problems.

Typically, these advantages come at the cost of more involved computation.

Bayesian causal inference: recent history

- D.B. Rubin formulated the modern foundations for causal inference from a largely Bayesian (missing data) perspective:
  - revived potential outcome concept to define causal estimand
  - inference through Bayesian (model-based) predictive formulation
  - focus on matching
- Semiparametric frequentist formulation pre-dominant from mid 80s
- Recent Bayesian approaches largely mimic semiparametric approach, but with explicit probability models.

Bayesian inference for two-stage models

- Full Bayes: full likelihood in two parametric models
  - needs correct specification;
  - two component models are treated independently.
- Quasi-Bayes: use semiparametric estimating equation approach for Stage II, with Stage I parameters treated in a fully Bayesian fashion.
  - possibly good frequentist performance;
  - difficult to understand frequentist properties.
- Pseudo-Bayes: use amended likelihood to avoid feedback between Stage I and Stage II
  - not fully Bayesian, no proper probability model

Five Considerations

1. The causal contrast
2. Do we really need potential outcomes ?
3. ‘Observables’ implies ‘Prediction’
4. The Fundamental Theory of Bayesian Inference.
5. The Bayesian Causal Specification

Part 6

Conclusions

Conclusions

- Causal inference methods provide answers to important questions concerning the impact of hypothetical exposures;
- Standard statistical methods are used;
- Balance is the key to accounting for confounding;
- The propensity score is a tool for achieving balance;
- The propensity score can be used for
  - matching,
  - weighting, and
  - as part of regression modelling.
- Bayesian methods are not widely used, but are generally applicable.

Key remaining challenges

- Model selection;
- Scale and complexity of observational data;

Collaborators & Acknowledgments

- McGill: Erica Moodie, Michael Wallace, Marina Klein
- Toronto: Olli Saarela
- Imperial College London: Dan Graham, Emma McCoy

NSERC
CRSNG

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

Reading List

References

- The propensity score
  - Rosenbaum and Rubin (1983): The introduction of the propensity score, gives basic definitions and properties.
- Applications
  - Austin (2011)
- Extensions beyond binary treatments
  - Hirano and Imbens (2004)
  - Imai and van Dyk (2004)
- Propensity score regression
  - Robins et al. (1992)
- Weighting
  - Lunceford and Davidian (2004)
  - Bang and Robins (2005)
- The marginal structural model
  - Hernán et al. (2000)
  - Hernán et al. (2001)

References

- Model selection
  - Brookhart et al. (2006)
- Longitudinal studies
  - Moodie and Stephens (2012)
  - Graham et al. (2014)
- High-dimensional settings
  - Schneeweiss et al. (2009)
- Bayesian methods
  - Rubin (1978)
  - McCandless et al. (2009)
  - An (2010)
  - Saarela et al. (2015)

References

An, W. (2010). Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. Sociological Methodology **40,** 151–189.

Austin, P. C. (2011). A tutorial and case study in propensity score analysis: An application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivariate Behavioral Research **46,** 119–151.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics **61,** 962–972.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Sturmer, T. (2006). Variable selection for propensity score models. American Journal of Epidemiology **163,** 1149–1156.

Graham, D. J., McCoy, E. J., and Stephens, D. A. (2014). Quantifying causal effects of road network capacity expansions on traffic volume and density via a mixed model propensity score estimator. Journal of the American Statistical Association **109,** 1440–1449.

Hernán, M. A., Brumback, B., and Robins, J. M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology **11,** 561–570.

Hernán, M. A., Brumback, B., and Robins, J. M. (2001). Marginal structural models to estimate the joint causal effect of nonrandomized treatments. Journal of the American Statistical Association **96,** 440–448.

Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. Applied Bayesian modeling and causal inference from incomplete-data perspectives pages 73–84.

References

Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association **99,** 854–866.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Statistics in Medicine **23,** 2937–2960. DOI: 10.1002/sim.1903.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. Statistics in Medicine **28,** 94–112.

Moodie, E. E. M. and Stephens, D. A. (2012). Estimation of dose-response functions for longitudinal data using the generalised propensity score. Statistical Methods in Medical Research .

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics **48,** 479–495.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika **70,** 41–55.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. The Annals of Statistics **6,** pp. 34–58.

Saarela, O., Stephens, D. A., Moodie, E. E. M., and Klein, M. B. (2015). On Bayesian estimation of marginal structural models. Biometrics **71,** 279–288.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology **20,** 512–522. doi:10.1097/EDE.0b013e3181a663cc.