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/SISCER2026/



The Key Challenge

Key objective: 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.
5. New directions.

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 G-estimation
2.7 Adjustment by weighting
2.8 Augmentation and double robustness
2.9 Continuous treatments

Part 3: Implementation and Computation

3.1 Statistical tools
3.2 Variance Estimation
3.3 Key practical considerations
3.4 Examples

Part 4: Extensions

4.1 Longitudinal studies

4.2 The Marginal Structural Model (MSM)

Part 5: Practical challenges

5.1 Classic challenges

5.2 Data complexity

5.3 Optimal Treatment Strategies

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 want to understand the effect of embarking upon road building schemes on traffic flow or density.

In such cases, 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

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.

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.

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).

We consider an *intervention* to change exposure status.

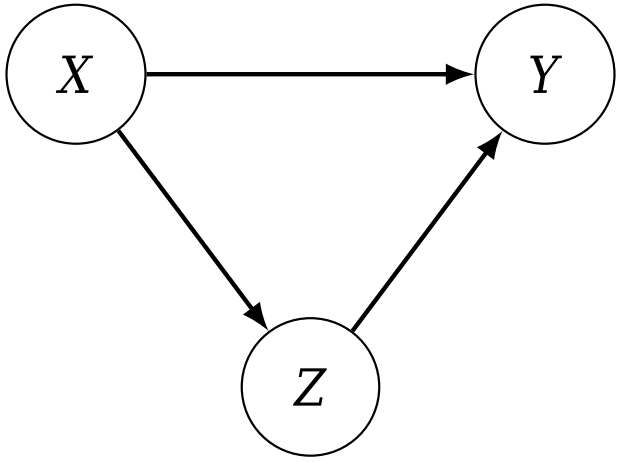
Notation

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.

Graphical representation



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

DAGs are commonly used to clarify causal thinking and assumptions.

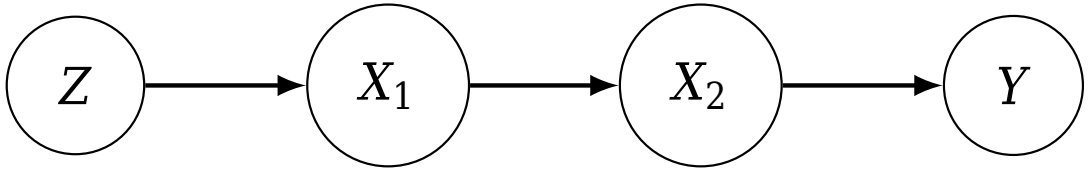
Probability & Causal DAGs recap

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 Z 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
  - the *direct* path $Z \rightarrow Y$,
  - the *indirect* path $Z \rightarrow X \rightarrow Y$

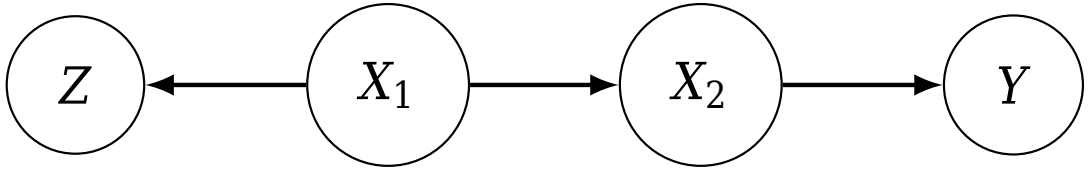
Graphical representation

- paths from Z to Y may be
  - *directed* if all the arrows point in the same direction,



Directed path

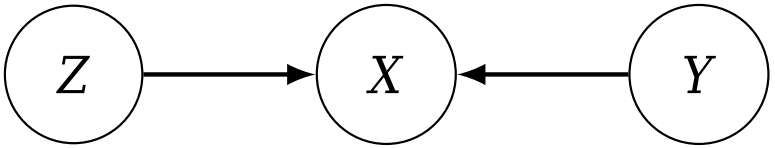
  - *undirected*



Unirected path

Graphical representation

- paths are *open* unless they contain a *collider*



Collider at X

- open paths may be closed by *conditioning*
- conditioning on a collider *opens* an otherwise closed path

Graphical representation

- causal effects 'flow' down *open directed* paths from Z to Y,
- *confounding* occurs if there are *open undirected* paths from Z to Y,
- the key task in a causal statistical analysis is to block open undirected paths by conditioning; conditioning could involve
  - stratification,
  - matching,
  - including variables in regression models.

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

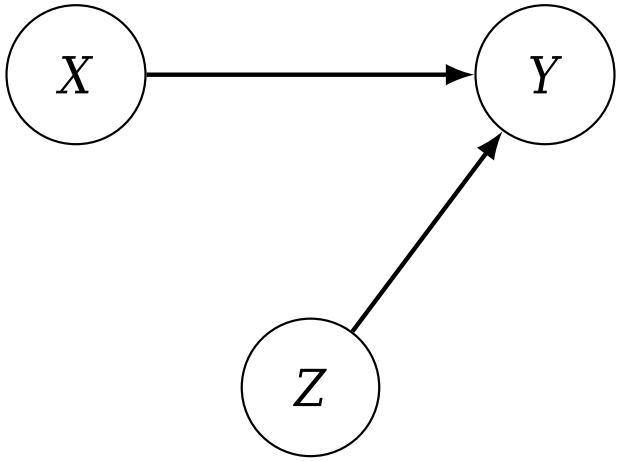
For example

$$Z = X + \varepsilon_Z$$

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

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.

$$Y = 2X + 5z + 3Xz + \varepsilon_Y$$

Counterfactual or Potential Outcomes

We denote by

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

the hypothetical outcome for individual i if we were to *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

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

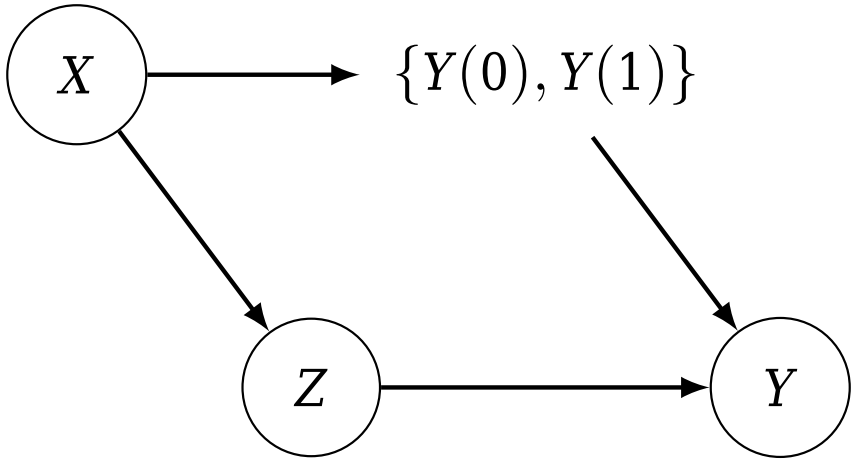
represent the outcomes that would result for individual i if that subject was not exposed, or exposed, respectively.

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

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

Counterfactual or Potential Outcomes

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



DAG with potential outcomes

If we know the value of X, then we know the distribution of both $Y(0)$ and $Y(1)$ without needing to know anything about the *actual* treatment Z.

Counterfactual or Potential Outcomes

If exposure is *multi-valued*, the potential outcomes

$$\{Y_i(z_1), Y_i(z_2), \ldots, Y_i(z_d)\}$$

represent the outcomes that would result for individual i if that subject exposed to exposure level $z_1, z_2, \ldots, z_d$ respectively.

Then for the observed Y_i,

$$Y_i = Y_i(z_j) \qquad \Longleftrightarrow \qquad Z_i = z_j.$$

Counterfactual or Potential Outcomes

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

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

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

$$\mathbb{1}_{\{z_j\}}(z) = \begin{cases} 1 & z = z_j \\ 0 & z \neq z_j \end{cases}$$

Counterfactual or Potential Outcomes

If exposure is *continuous-valued*, the potential outcomes

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

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.

Causal Estimands

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

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

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.

Expected counterfactuals

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

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

or contrasts of these quantities. We might also consider

- *subgroup-specific* expected quantities

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

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

- *conditional* expected quantities

$$\mathbb{E}[Y_i(\mathbf{z})|X_i = x]$$

  as x varies.

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

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

so may also consider odds or odds ratios quantities

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

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

JAMA Pediatr. 2016 Feb;170(2):117-24. doi: 10.1001/jamapediatrics.2015.3356.

Antidepressant Use During Pregnancy and the Risk of Autism Spectrum Disorder in Children

Takoua Boukhris [1], Odile Sheehy [2], Laurent Mottron [3], Anick Bérard [1]

Affiliations

Affiliations

1 Faculty of Pharmacy, University of Montréal, Montréal, Québec, Canada2Research Unit on Medications and Pregnancy, Research Center, CHU Sainte-Justine, Montréal, Québec, Canada.
2 Research Unit on Medications and Pregnancy, Research Center, CHU Sainte-Justine, Montréal, Québec, Canada.
3 Centre d'excellence en Troubles Envahissants du Développement de l'Université de Montréal, Montréal, Québec, Canada4Département de Psychiatrie, Hôpital Rivière-des-Prairies, Montréal, Québec, Canada5Centre de Recherche de l'Institut Universitaire de Psych.

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

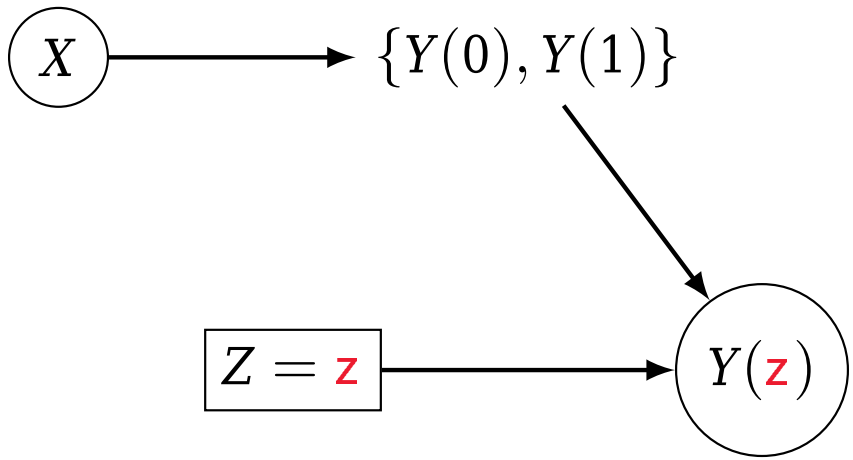
(APO) values.

Consider first the situation where all subjects in a random sample receive a given exposure z; we wish to estimate

$$\mu(z) = \mathbb{E}[Y(z)].$$

Estimation of average potential outcomes

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



DAG with exposure intervention $Z = z$

Estimation of average potential outcomes

As a mathematical calculation, we write the expected outcome as

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

where $f_{Y(\mathsf{z})}(y)$ is the distribution of the potential outcome $Y(\mathsf{z})$.

We read this calculation as

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

Estimation of average potential outcomes

We may also write this as

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

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

We know from the DAG on p. 41 that

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

Estimation of average potential outcomes

Note that we may also write

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

assuming an exposure distribution that sets $z = z$ with probability one.

- the data are considered to be sampled from the distribution

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

Thus, for the APO we have

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

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

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

Side Note: Monte Carlo methods

Causal quantities are typically *average* measures across a given population which we write as *integrals* with respect to probability distributions.

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

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

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 and *estimates* based on this sample (eg sample averages),
- if our sample is large enough, the estimate will be close to the true expected value.

Side Note: Monte Carlo methods

Importance sampling is based on forming *weighted* averages

- we *re-weight* obtained samples from another distribution f_Y^* so that we can estimate quantities relating to f_Y
- this is like '*standardization*' (eg standardized mortality rate) in epidemiology.

Monte Carlo recap

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

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

This is sometimes known as a *G-computation* formula. It essentially arises by assuming X and Z are *independent*, and then studying the *marginal* (over X) distribution

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

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*.

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:

- they only differ *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.

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

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

We may write

$$N_z = \sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i).$$

as the count of the number of individuals receiving $Z = z$.

The randomized study

This suggests the estimators[1]

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

that is

$$\widehat{\mathbb{E}}[Y(0)] = \frac{1}{N_0} \sum_{i=1}^{n} (1 - Z_i) Y_i \qquad \widehat{\mathbb{E}}[Y(1)] = \frac{1}{N_1} \sum_{i=1}^{n} Z_i Y_i.$$

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

Note that for the denominators,

$$N_0 \sim \text{Binomial}(n, 1-p) \qquad N_1 \sim \text{Binomial}(n, p)$$

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

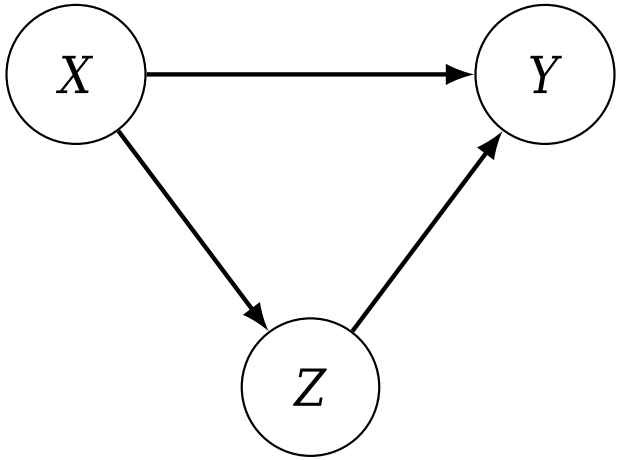
$$\widehat{\mathbb{E}}[Y(0)] = \frac{1}{n(1-p)} \sum_{i=1}^{n} (1 - Z_i) Y_i \qquad \widehat{\mathbb{E}}[Y(1)] = \frac{1}{np} \sum_{i=1}^{n} Z_i Y_i. \quad (2)$$

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



X is a confounder

- X predicts outcome Y in the presence of Z:

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

- X predicts exposure Z:

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

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;

Confounding

Causal inference is fundamentally about comparing exposure subgroups on an *equal footing*: there is no influence of the other predictors.

- This is possible in the randomized study as randomization breaks the association between Z and X.
- It is *not* directly possible in the presence of confounding.

Confounding and collapsibility

Note 5.

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

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

and

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

Simple confounding example

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

$$\Pr[Z = 1 | X = x] = \begin{cases} p_0 & x = 0 \\ p_1 & x = 1 \end{cases}$$

$$= (1 - x)p_0 + xp_1$$

so that

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

say.

Simple confounding example

It can be shown that ATE estimator

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

has expectation

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

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

Example

Simple confounding example

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

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

are *not directly comparable* - they have a different profile in terms of X.

Simple confounding example

By *Bayes theorem*

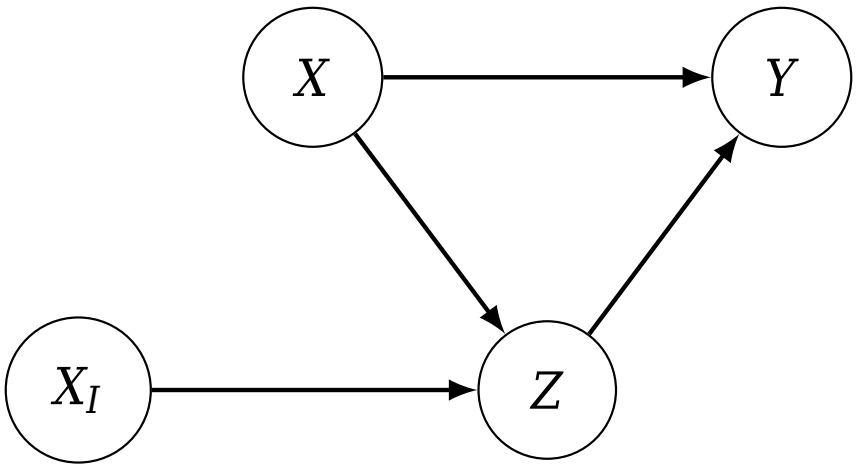
$$\Pr[X = 1 | Z = 1] = \frac{p_1 q}{p} \qquad \Pr[X = 1 | Z = 0] = \frac{(1 - p_1) q}{1 - p}$$

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 X_I predicts Z, but does not predict Y in the presence of Z, then X_I is termed an *instrument*.



DAG with instrument X_I: X_I predicts Z, but is not associated with outcome Y if we know Z.

Example: Non-compliance

In a randomized study of a binary treatment, if Z records the treatment actually *received*, suppose that there is *non-compliance* with respect to the treatment; that is, if X_I records the treatment *assigned* by the experimenter, then possibly

$$x_I \neq z.$$

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.

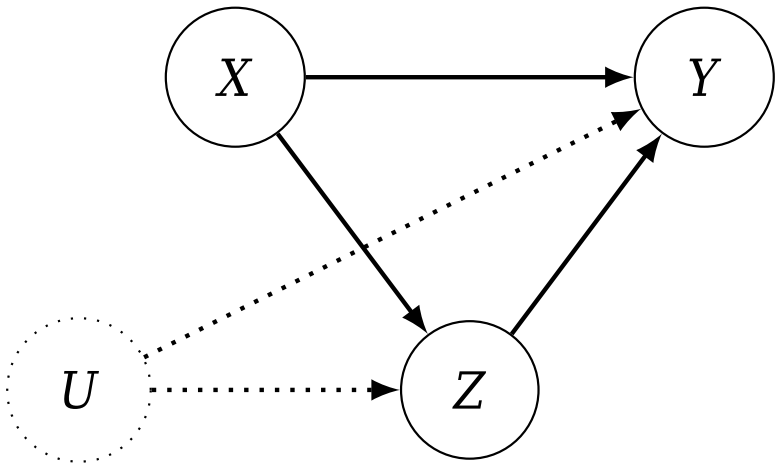
Critical Assumption

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

Suppose that the *true outcome mean* is given by

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

which may be parametric in nature, say

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

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

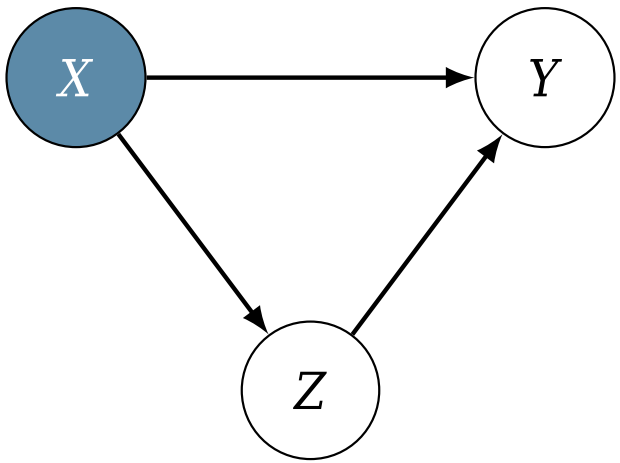
To estimate the APO, we simply set

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

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,
- we may mistakenly use a model $m(x, z)$ instead of $\mu(x, z)$.

We seek statistical methods that can overcome this.

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 6.

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

- define strata where the X values are *precisely matched*, then
- compare the outcomes for treated and untreated individuals *within* those strata, then
- perform 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 again consider K matching strata

$$\mathcal{S}_1, \ldots, \mathcal{S}_K.$$

The formula for combining strata still holds, but

- 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, conclusions are restricted to the region covered by the strata.

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*, $\mathsf{e}(X)$[2] 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.

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

Balance via the propensity score

For the conditional independence $X \perp\!\!\!\perp Z|\mathsf{e}$, we require that

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

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

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[3]

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

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.

[3] 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*[4]. 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

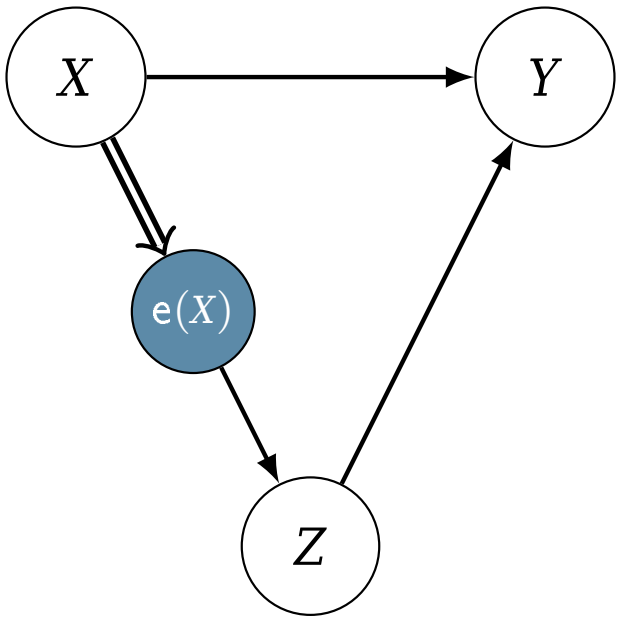
$$f_{Z|X}(z|x) \equiv f_{Z|\mathsf{e}(X)}(z|e),$$

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.

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

Balance via the propensity score



DAG with confounding path $Z \to X \to Y$ blocked by conditioning on $e(X)$. Double arrow $X \Rightarrow e(X)$ indicates a deterministic relationship.

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

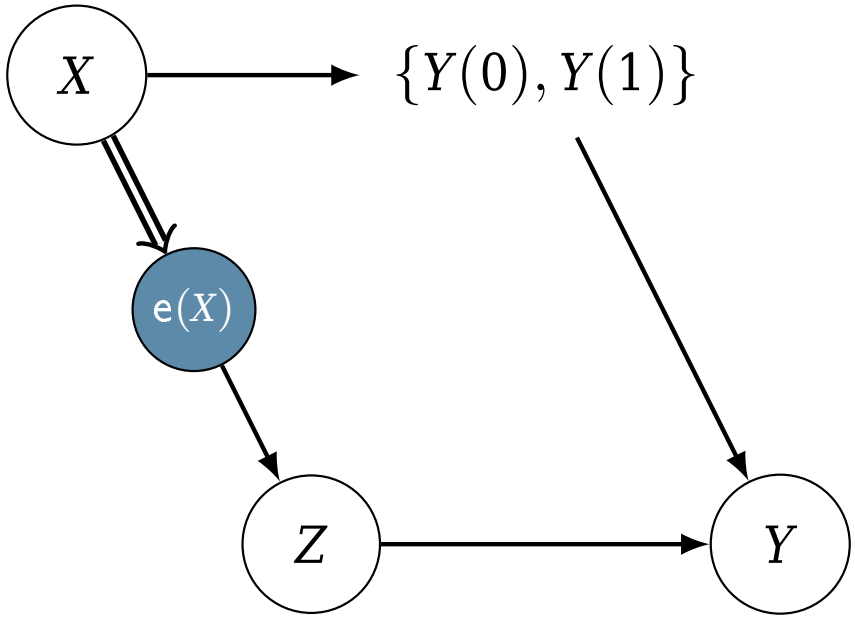
The assumption of 'no unmeasured confounders' amounts to assuming that the potential outcomes are jointly *independent* of exposure assignment given the confounders, that is

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

that is, in terms of densities

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

Unconfoundedness given the propensity score



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

It is clear from the DAG that

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

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 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 $\mathsf{e}(x)$, 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 $\mathsf{e}(x)$;
- 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,

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

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 $\mathsf{e}(X)$, 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 may consider balancing scores that are functions of *both* Z and X.

Beyond binary exposures

For a balancing score $\mathsf{b}(Z, X)$[5], we require that

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

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.

[5] 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.

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

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

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.

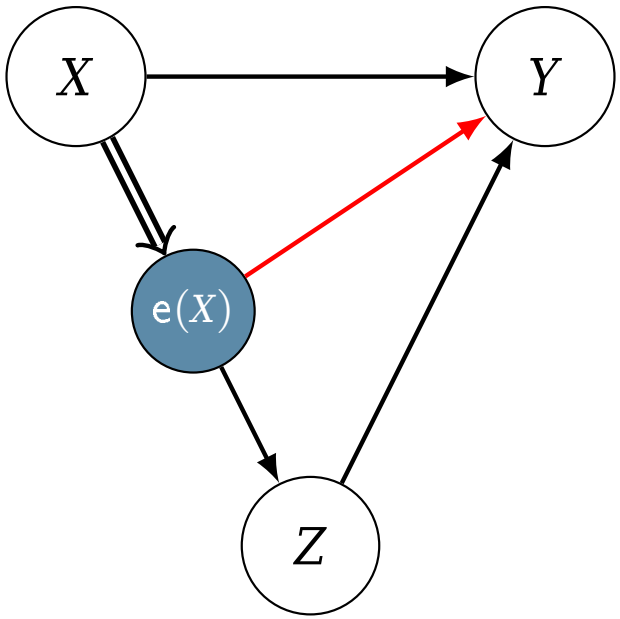
Propensity Score Regression

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



Conditioning on $e(X)$ achieved using regression

Propensity Score Regression

We may build a regression model

$$\mathbb{E}_{Y|X,Z}[Y|X, \mathsf{e}(X), Z]$$

which, as

$$X \perp\!\!\!\perp Z \mid \mathsf{e}(X)$$

has the advantage that it will be more robust to possible mis-specification when a parametric model is proposed.

Propensity Score Regression

For example, we may specify

$$\mathbb{E}[Y|X, \mathsf{e}(X), Z] = \beta_0 + \psi_0 Z + \phi_0 \mathsf{e}(X)$$

and carry out OLS estimation to estimate ψ_0 as the ATE parameter.

Propensity Score Regression

If we have the *true* conditional mean[6] is

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

then by the unconfoundedness result that

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

[6] this is very particular assumption!

Propensity Score Regression

That is, to estimate the APO, we might

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

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

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 $\widehat{\mathsf{e}}(x) \equiv \mathsf{e}(x;\widehat{\alpha})$.
- For $\mu(x, z, e; \theta)$, estimate θ by regressing y_i on z_i and $e_i = \widehat{\mathsf{e}}(x_i)$.

  For example, if $\theta = (\beta_0, \psi_0, \phi_0)$

$$\mathbb{E}[Y | X_i = x_i, Z = z_i, \mathsf{e}(X_i) = e_i; \theta] = \beta_0 + \psi_0 z_i + \phi_0 e_i.$$

Example: Binary exposure

We then average the model predictions to obtain the APO estimate

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

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

Such procedures require us to make two modelling choices:

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

For consistent inference for the ATE, we need

- the propensity model, *and*
- the outcome mean model

to be correctly specified.

However, in some circumstances, the second assumption can be relaxed to require that the dependence of the outcome mean model on z is correctly specified.

Propensity Score Regression

Example: Binary exposure

Suppose that the true (data generating) conditional mean can be written

$$\mathbb{E}[Y|X = x, Z = z] = \mu(x, z) = \mu_0(x) + z\mu_1(x)$$

but that the propensity score regression model

$$\mathbb{E}[Y|X = x, Z = z, \mathrm{e}(X) = e] = m_0(x) + z\mu_1(x) + e\mu_1(x)$$

is used.

Propensity Score Regression

Example: Binary exposure

This is sufficient to give consistent estimation of the ATE

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\mu_1(X)].$$

That is, we may mis-specify the 'treatment-free' component

$$\mu_0(x)$$

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

Propensity Score Regression

If we believe there is a treatment/predictor *interaction*, say

$$\mathbb{E}[Y|X,Z] = \beta_0 + Z(\psi_0 + \psi_1 X_1)$$

we should fit the propensity score regression model

$$\mathbb{E}[Y|X,\mathsf{e}(X),Z] = \beta_0 + Z(\psi_0 + \psi_1 X_1) + \mathsf{e}(X)(\phi_0 + \phi_1 X_1)$$

- for every term in Z, we include a *corresponding* term in $\mathsf{e}(X)$;
- we use predictions from this model in order to estimate the ATE.

PSR Example

Justification: G-estimation

In the binary treatment case, if the data generating model is

$$\mathbb{E}[Y|X = x, Z = z] = \mathbf{x}_0\beta_{\text{TRUE}} + z\,\mathbf{x}_2\psi = \mu(x, z; \beta_{\text{TRUE}}, \psi)$$

for row vectors $\mathbf{x}_0$ and $\mathbf{x}_2$, then the propensity score regression model

$$m(x, z; \beta, \psi, \phi) = \mathbf{x}_1\beta + z\mathbf{x}_2\psi + \mathsf{e}(x)\mathbf{x}_2\phi$$

for row vector $\mathbf{x}_1$ will return a consistent estimator of ψ even if the *treatment-free mean model* $\mathbf{x}_1\beta$ is mis-specified.

Justification: G-estimation

Consider the OLS estimation of (β, ψ, ϕ): we solve

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{i1}^\top \\ z_i \mathbf{x}_{i2}^\top \\ \mathsf{e}(x_i)\mathbf{x}_{i2}^\top \end{pmatrix} (y_i - \mathbf{x}_{i1}\beta - z_i\, \mathbf{x}_{i2}\psi - \mathsf{e}(x_i)\mathbf{x}_{i2}\phi) = \mathbf{0}$$

analytically using the usual approaches.

Justification: G-estimation

If $\mathbf{x}_1 = \mathbf{x}_0$ then the treatment-free model is correctly specified, and we still recover the correct ATE.

If the mean model is *mis-specified*, but

(i) the propensity score model $\mathsf{e}(x)$ is *correctly specified*;

(ii) the random quantity

$$\varepsilon_i = (Y_i - \mathbf{X}_{i1}\beta - Z_i\,\mathbf{X}_{i2}\psi - \mathsf{e}(X_i)\mathbf{X}_{i2}\phi)$$

is *independent* of Z_i, so that the effect of Z_i is correctly captured via

$$Z_i\,\mathbf{X}_{i2}\psi.$$

the solutions to the resulting estimating equation are still *consistent* for the true values.

Justification: G-estimation

We simply rearrange the OLS estimating equations to

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{i1}^\top \\ z_i \mathbf{x}_{i2}^\top \\ (z_i - \mathsf{e}(x_i))\mathbf{x}_{i2}^\top \end{pmatrix} (y_i - \mathbf{x}_{i1}\beta - z_i\, \mathbf{x}_{i2}\psi - \mathsf{e}(x_i)\mathbf{x}_{i2}\phi) = \mathbf{0}$$

to see this.

Justification: G-estimation

Inference for ψ is correct if *at least one* of

- the mean model, *or*
- the propensity score model

is correctly specified, *provided* the treatment effect model is correctly specified.

This is known as *double robustness*.

Justification: G-estimation

We may consider the reduced form

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{i1}^\top \\ (z_i - \mathsf{e}(x_i))\mathbf{x}_{i2}^\top \end{pmatrix} (y_i - \mathbf{x}_{i1}\beta - z_i\, \mathbf{x}_{i2}\psi) = \mathbf{0}.$$

This form still leads to double robustness.

Estimation based on this system is known as *G-estimation*.

Justification: G-estimation

The most basic form of the G-estimating equation arises from the model that omits the treatment-free component:

$$\sum_{i=1}^{n}(z_i - \mathsf{e}(x_i))\mathbf{x}_{i2}^{\top}(y_i - z_i\,\mathbf{x}_{i2}\psi) = 0$$

and in the simplest case with ψ one-dimensional

$$\sum_{i=1}^{n}(z_i - \mathsf{e}(x_i))(y_i - z_i\psi_0) = 0$$

say.

Justification: G-estimation

In this case we can solve explicitly to obtain

$$\widehat{\psi}_0 = \frac{\sum_{i=1}^{n}(z_i - \mathsf{e}(x_i))y_i}{\sum_{i=1}^{n} z_i(z_i - \mathsf{e}(x_i))}$$

with corresponding estimator

$$\frac{\sum_{i=1}^{n}(Z_i - \mathsf{e}(X_i))Y_i}{\sum_{i=1}^{n} Z_i(Z_i - \mathsf{e}(X_i))}.$$

Justification: G-estimation

These results extend to more complicated settings: for example, the *doubly robust* G-estimator takes the form

$$\frac{\sum_{i=1}^{n}(Z_i - \mathsf{e}(X_i))(Y_i - \mathbf{X}_{i1}\widehat{\beta})}{\sum_{i=1}^{n} Z_i(Z_i - \mathsf{e}(X_i))}.$$

Weighting approaches

We focus on the APO

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

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

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

Average potential outcome: Experimental study

Let $\mathcal{E}$ indicate the *experimental* design density

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

We have that

$$\mathbb{E}[Y(z)] \equiv \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z=z] = \frac{\mathbb{E}^{\mathcal{E}}_{X,Y,Z}[\mathbb{1}_{\{z\}}(Z)Y]}{\mathbb{E}^{\mathcal{E}}_{X,Y,Z}[\mathbb{1}_{\{z\}}(Z)]}$$

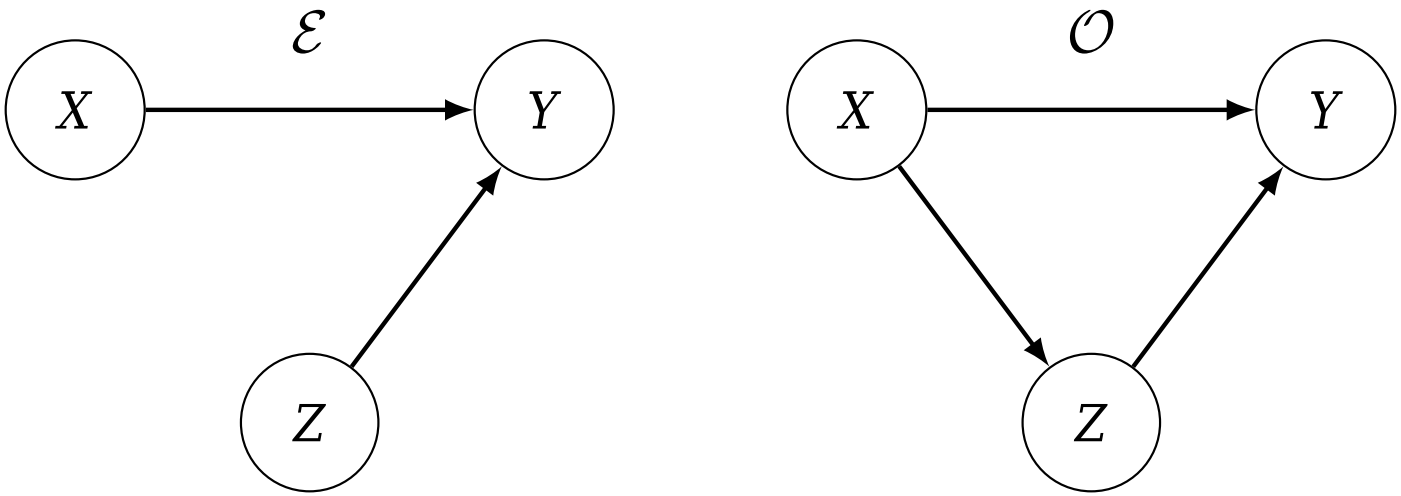
Average potential outcome: Observational study

However, the data arise from the *observational* (non-experimental) distribution $\mathcal{O}$ with density

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

and in order to perform estimation, we must re-write the expectations in terms of this density.

Average potential outcome: Observational study



DAGs for Experimental ($\mathcal{E}$) and Observational ($\mathcal{O}$) distributions

$$\mathcal{E} \; : \quad f^{\mathcal{E}}_{X,Y,Z}(x,y,z) = f^{\mathcal{E}}_{Y|Z,X}(y|z,x)\, f^{\mathcal{E}}_{Z}(z)\, f^{\mathcal{E}}_{X}(x)$$

$$\mathcal{O} \; : \quad f^{\mathcal{O}}_{X,Y,Z}(x,y,z) = f^{\mathcal{O}}_{Y|Z,X}(y|z,x)\, f^{\mathcal{O}}_{Z|X}(z|x)\, f^{\mathcal{O}}_{X}(x)$$

Reweighting

We may use re-weighting (importance sampling logic) to re-write

$$\mathbb{E}[Y(z)] = \frac{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[\mathbb{1}_{\{z\}}(Z) Y \, w(X,Y,Z)]}{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[\mathbb{1}_{\{z\}}(Z) \, w(X,Y,Z)]}$$

where

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

Reweighting

The function $w(x, y, z)$ can be re-written

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

- 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.

This suggests the (non-parametric) estimators

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

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 7.

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 8.

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 9.

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 $\mu(x, z) = \mathbb{E}[Y|X = x, Z = z]$ is the data generating conditional outcome mean.

We then have the alternate estimator

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

This is termed the *augmented IPW* (AIPW) estimator.

Estimation via augmentation

Then, if both

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

are correctly specified, we have

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

Furthermore, (AIPW) is *doubly robust*

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

Properties under mis-specification

Suppose that (possibly mis-specified) models

$$f(z|x) \qquad m(x,z).$$

are used to form the estimator

$$\widehat{\mathbb{E}}[Y(z)] = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_i)(Y_i - m(X_i,Z_i))}{f(Z_i|X_i)} + \frac{1}{n}\sum_{i=1}^{n} m(X_i,z) \tag{5}$$

$$= \sum_{i=1}^{n} W_{iz}(Y_i - m(X_i,Z_i)) + \frac{1}{n}\sum_{i=1}^{n} m(X_i,z)$$

Properties under mis-specification

Then the *bias* of the estimator is is

$$\mathbb{E}\left[\frac{(f(z|X) - f^{\mathcal{O}}_{Z|X}(z|X))(m(X,z) - \mu(X,z))}{f(z|X)}\right] \tag{6}$$

which is zero if

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

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 (6) that determines the quality of the estimator.

- equation (6) demonstrates how mis-specification in the functions $\mu(x, z)$ and $f^{\mathcal{O}}_{Z|X}$ contributes to the bias.

We proceed by assuming that $\mu(x, z)$ is represented by model $m(x, z)$, but that the propensity model $f^{\mathcal{O}}_{Z|X}(z|x)$ is correctly specified.

Parametric modelling: two-stage approach

In the formulation, parametric models for

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

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 10.

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.

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 (5) as

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

Alternative view of augmentation

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

$$m(x, z) = m_0(x, z) + \phi \frac{\mathbb{1}_{\{z\}}(z)}{f^{\mathcal{O}}_{Z|X}(z|x)}$$

and $m_0(x, z)$ is a conditional mean model, and ϕ is a regression coefficient associated with the derived predictor[7]

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

[7] This predictor is sometimes called the '*clever covariate*'

Alternative view of augmentation

Therefore, an estimator equivalent to (5) can be obtained by regressing Y on (X, Z) for fixed z using $m(x, z)$, and forming the estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left\{m_0(X_i, Z_i) + \widehat{\phi}\frac{\mathbb{1}_{\{z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}\right\}.$$

In a parametric model setting, this becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left\{m_0(X_i, Z_i; \widehat{\beta}) + \widehat{\phi}\frac{\mathbb{1}_{\{z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i; \widehat{\alpha})}\right\}$$

where α is estimated from Stage (I), and β is estimated along with ϕ 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

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

where

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

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

$$m(x,z) = m_0(x,z) + \phi \left[\frac{\mathbb{1}_1(z)}{f^{\mathcal{O}}_{Z|X}(1|x)} - \frac{\mathbb{1}_0(z)}{f^{\mathcal{O}}_{Z|X}(0|x)} \right]$$

or

$$m(x,z) = m_0(x,z) + \left[\phi_1 \frac{\mathbb{1}_1(z)}{f^{\mathcal{O}}_{Z|X}(1|x)} - \phi_0 \frac{\mathbb{1}_0(z)}{f^{\mathcal{O}}_{Z|X}(0|x)} \right].$$

Continuous treatments

For continuous treatments, if necessary we may carry out adjustment via the conditional *expectation*

$$\mathsf{b}(X) = \mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X]$$

rather than (for example) the propensity score which is based on the conditional *probability* model

$$f^{\mathcal{O}}_{Z|X}(Z|X).$$

Propensity score regression & G-estimation

The data generating model

$$Y = Z\psi + \mu_0(X;\beta) + \varepsilon$$

which forms the basis of the G-estimation procedure can be utilized if Z is *continuous*.

This relies on the construction of a model for $\mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X]$ which can be achieved using a linear model.

Propensity score regression & G-estimation

In the continuous setting we can consider other functions of Z in the treatment effect model, for example

$$\mathbf{x}_0\beta + \psi_0 Z + \psi_1 Z^2$$

or, in an *interaction* model

$$\mathbf{x}_0\beta + \psi_0 Z + \psi_1 Z^2 + \psi_2 Z X_1 + \psi_3 Z^2 X_1$$

Propensity score regression & G-estimation

In general, we need to construct models for all the terms in Z: for example

$$\mathsf{b}_1(X) = \mathbb{E}_{Z|X}[Z|X] \qquad \text{model for } Z$$

$$\mathsf{b}_2(X) = \mathbb{E}_{Z|X}[Z^2|X] \qquad \text{model for } Z^2$$

etc.

Propensity score regression & G-estimation

In propensity score regression, we again use the balancing scores to block the backdoor paths. For the model

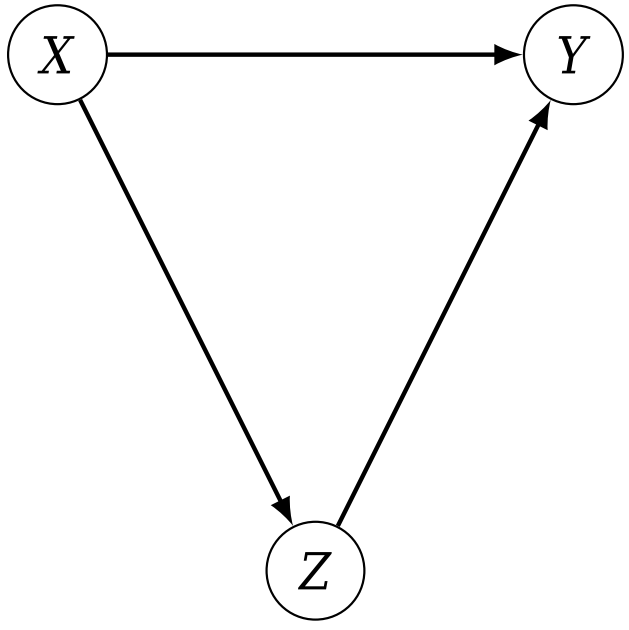
$$\mathbf{x}_0\beta + \psi_0 Z + \psi_1 Z^2$$

it is sufficient to use the PSR model

$$\beta_0 + \psi_0 Z + \psi_1 Z^2 + \phi_0 \mathsf{b}_1(X)$$

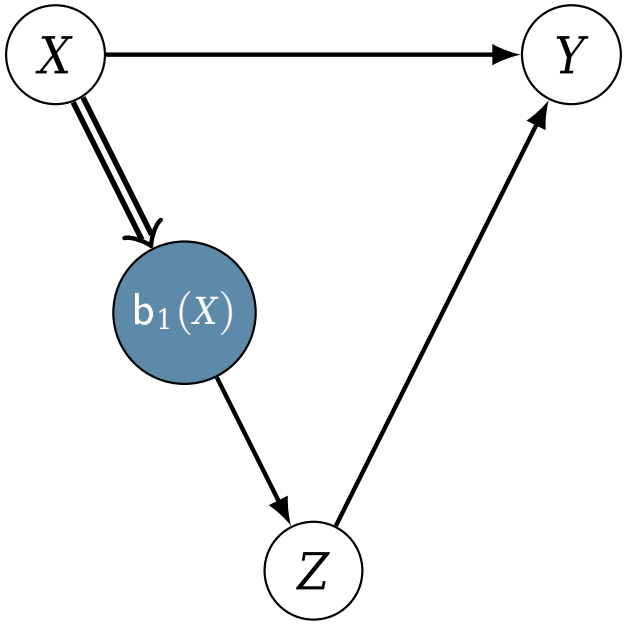
as there is only one backdoor path.

Propensity score regression & G-estimation



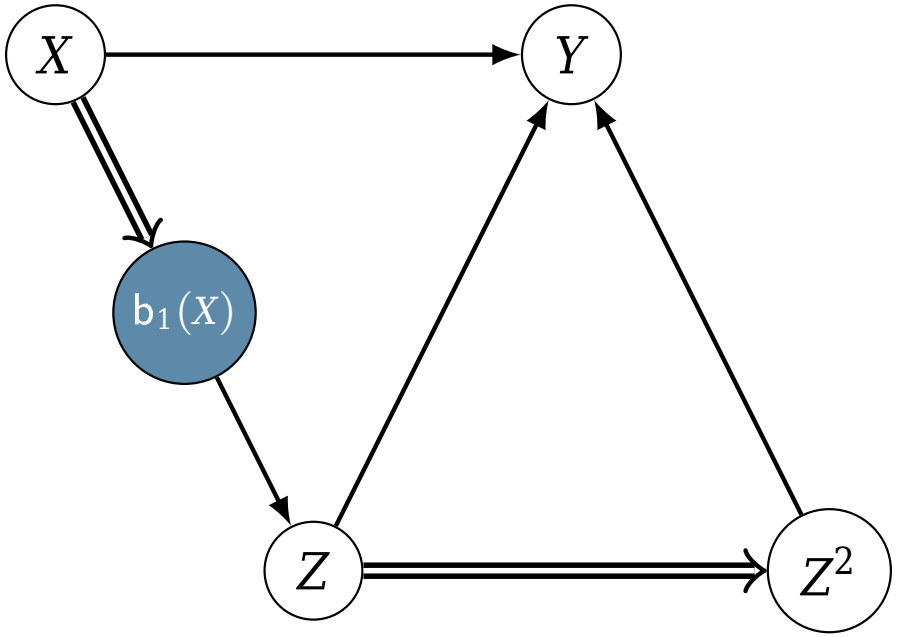
Confounding DAG

Propensity score regression & G-estimation



Confounding DAG with balancing score $b_1(X)$

Propensity score regression & G-estimation



Inclusion of quadratic term: confounding path still blocked by $b_1(X)$

However, for the model

$$\mathbf{x}_0\beta + \psi_0 Z + \psi_1 Z^2 + \psi_2 Z X_1 + \psi_3 Z^2 X_1$$

we must block *both* paths through the interactions using

$$\phi_0 \mathsf{b}_1(X) + \phi_1 \mathsf{b}_1(X) X_1 + \phi_2 \mathsf{b}_2(X) + \phi_3 \mathsf{b}_2(X) X_1$$

that includes *both* $\mathsf{b}_1(X)$ and $\mathsf{b}_2(X)$.

Inverse Probability Weighting

For IPW estimation, in principle the construction in equation (IPW0)

$$\widehat{\mu}_{\text{IPW}}(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_i)Y_i}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}$$

works in the continuous case, as again

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}\left[\frac{\mathbb{1}_{\{z\}}(Z)Y}{f^{\mathcal{O}}_{Z|X}(Z|X)}\right] = \mu(z).$$

We may also use the *standardized weight* estimator in equation (IPW).

Inverse Probability Weighting

We have as usual

$$\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\frac{\mathbb{1}_{\{z\}}(Z)}{f^{\mathcal{O}}_{Z|X}(Z|X)}\right] = 1.$$

In the continuous setting, for any fixed z, the estimator includes those data for which

$$\mathbb{1}_{\{z\}}(z_i) = 1$$

that is, when $z_i = z$. When Z is treated as continuous, we will get at most one data point meeting the criterion for each z.

Inverse Probability Weighting

Therefore in practice we typically need to use an estimator based on

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

where, for some small $d > 0$,

$$A_z = (z - d, z + d)$$

is a d-neighbourhood of z.

Inverse Probability Weighting

It is sometimes recommended to use the *stabilized* weight based on the (estimated) marginal distribution $f_Z(z)$, that is

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

That is, say

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}_{\{A_z\}}(Z_i)f_Z^{\mathcal{O}}(Z_i)}{f_{Z|X}^{\mathcal{O}}(Z_i|X_i)}Y_i$$

Inverse Probability Weighting

These estimators do not rely on a mean model, but as before can be augmented with a proposed mean model

$$m(x, z)$$

such as

$$\widehat{\mu}_{\text{AIPW}}(z) = \sum_{i=1}^{n} W_{iz}(Y_i - m(X_i, Z_i)) + \frac{1}{n}\sum_{i=1}^{n} m(X_i, z).$$

Inverse Probability Weighting

This can also be fitted using the augmented outcome regression (AOR) approach based on augmented model

$$m(X_i, Z_i) + \phi_0 \frac{\mathbb{1}_{\{A_z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}$$

fitted using least squares.

Inverse Probability Weighting

In weighted least squares (WLS), estimation is carried out according to a proposed mean model, with minimization of a weighted least square objective function.

For example, the mean model $m(x,z) \equiv m(z) = \beta_0 + \psi_0 z$ might be fitted according to the WLS objective function

$$\sum_{i=1}^{n} w(x_i, z_i)(y_i - \beta_0 - \psi_0 z_i)^2$$

This model assumes a linear dependence on Z.

Inverse Probability Weighting

Note: need to take care with *positivity* violations

- cannot have that X predicts Z too precisely
- can check by inspecting the weights: should have average near 1.

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.

Advanced modelling

Could also consider advanced modelling methods for

- the outcome mean model

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

- the propensity or balancing score

$$\mathsf{e}(x) = \Pr[Z=1|X=x] \qquad \mathsf{b}(x) = \mathbb{E}_{Z|X}[Z|X=z]$$

Advanced modelling

In either case we can utilize

- linear/generalized linear models
- flexible models (eg splines)
- prediction approaches (eg machine learning methods, regression trees, neural networks)
- tree-based methods (eg CART, BART, random forests)
- ensemble methods (eg model averaging, boosting)

to construct fitted versions of each model.

Econometrics Journal (2018), volume **21**, pp. C1–C68.
doi: 10.1111/ectj.12097

Double/debiased machine learning for treatment and structural parameters

VICTOR CHERNOZHUKOV[†], DENIS CHETVERIKOV[‡], MERT DEMIRER[†], ESTHER DUFLO[†], CHRISTIAN HANSEN[§], WHITNEY NEWEY[†] AND JAMES ROBINS[‖]

Advanced modelling

Potential pitfalls:

1. The quantification of *uncertainty*;
   - no ready analytic answers,
   - typically relies on bootstrap;
   - large computational burden.
2. *Positivity violations*;
   - prediction (of treatment) is not the fundamental goal;
   - flexible modelling of the treatment assignment mechanism can predict treatment with too much precision, rendering causal comparisons impossible;
   - can be overcome using methods that target balance or overlap explicitly.

3. *Lack of theoretical guarantees*;
    - estimation of the nuisance models (eg propensity score) needs exhibit fast enough convergence
    - hard to study consistency, establish asymptotic normality
    - needs more advanced theoretical techniques (eg sample splitting)

RESEARCH ARTICLE

WILEY Statistics in Medicine

Should a propensity score model be super? The utility of ensemble procedures for causal adjustment

Shomoita Alam[1] | Erica E. M. Moodie[1] | David A. Stephens[2]

Statistics in Medicine, 2019, **38**: 1690–1702.

Variance estimation

Variance estimation can be carried out using the so-called 'robust' '*sandwich*' estimation procedure.

- based on large-sample (semiparametric) theory;
- `sandwich` package in `R` for standard `R` classes;
- some causal packages have built-in robust variance estimation; see for example `drgee`.
- need to account for estimation of *nuisance* parameters;
- for propensity score regression or G-estimation, estimation of nuisance parameters has a minimal effect on the variance estimate.

Bootstrap methods can also be used for regression or IPW approaches.

Variance 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;
- standard errors may be estimated in a robust fashion using the sandwich estimator of the asymptotic variance.

Variance estimation

For estimating function $\mathbf{U}(\cdot)$ for parameter θ, we define the estimator, $\widehat{\theta}$, as the solution to

$$\mathbf{U}(\theta) \equiv \mathbf{U}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}; \theta) = \sum_{i=1}^{n} \mathbf{U}(X_i, Y_i, Z_i; \theta) = \mathbf{0}.$$

The sandwich form of the asymptotic variance of $\widehat{\theta}$ takes the form

$$\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\top}$$

where

$$\mathbf{A} = -\mathbb{E}[\dot{\mathbf{U}}(\theta_0)] \qquad \mathbf{B} = \mathbb{E}[\mathbf{U}(\theta_0)\{\mathbf{U}(\theta)\}^{\top}]$$

Variance estimation

Here

- θ_0 is the true data generating parameter;
- $\dot{\mathbf{U}}$ is the matrix of first partial derivatives of $\mathbf{U}$ with respect to θ

$$\left[\dot{\mathbf{U}}(\theta)\right]_{j,k} = \frac{\partial U_j(\theta)}{\partial \theta_k}$$

- $\mathbf{A}$ is not always a symmetric matrix;
- in practice, $\mathbf{A}$ and $\mathbf{B}$ are estimated using the relevant sample averages, and θ_0 is replaced by $\widehat{\theta}$.

Example: Linear regression

For linear regression, we have

$$\mathbf{U}(\theta) = \mathbf{X}^\top (Y - \mathbf{X}\beta)$$
$$-\dot{\mathbf{U}}(\theta) = \mathbf{X}^\top \mathbf{X}$$

so

$$\mathbf{A} = \{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]\}^{-1} \qquad \mathbf{B} = \mathbb{E}[\mathbf{X}^\top (Y - \mathbf{X}\beta_0)^2 \mathbf{X}]$$

and the sandwich variance is

$$\mathbf{V} = \{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]\}^{-1} \{\mathbb{E}[\mathbf{X}^\top (Y - \mathbf{X}\beta_0)^2 \mathbf{X}]\} \{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]\}^{-1}$$

Key considerations

In light of the previous discussions, in order to facilitate causal comparisons, there are several considerations:

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.

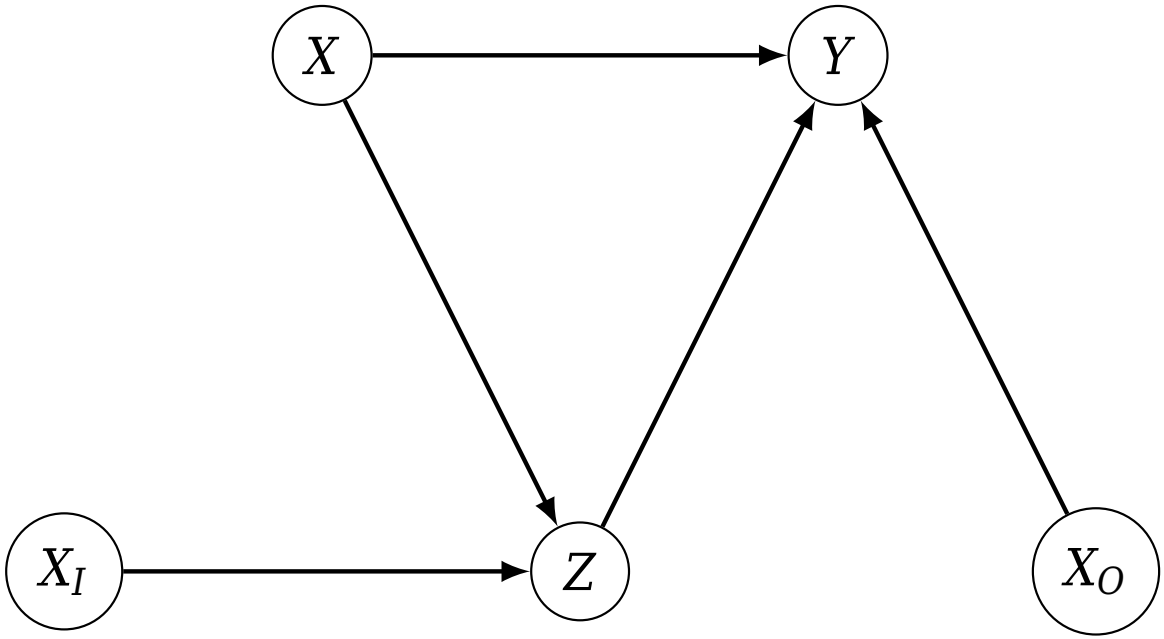
Key considerations

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 variable selection techniques 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.**
   The model for the balancing score must 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 propensity model to induce *balance*;
   - *positivity* (*overlap*): strata must contain sufficient data 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 11.

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 12.

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 ?

Numerical examples

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.

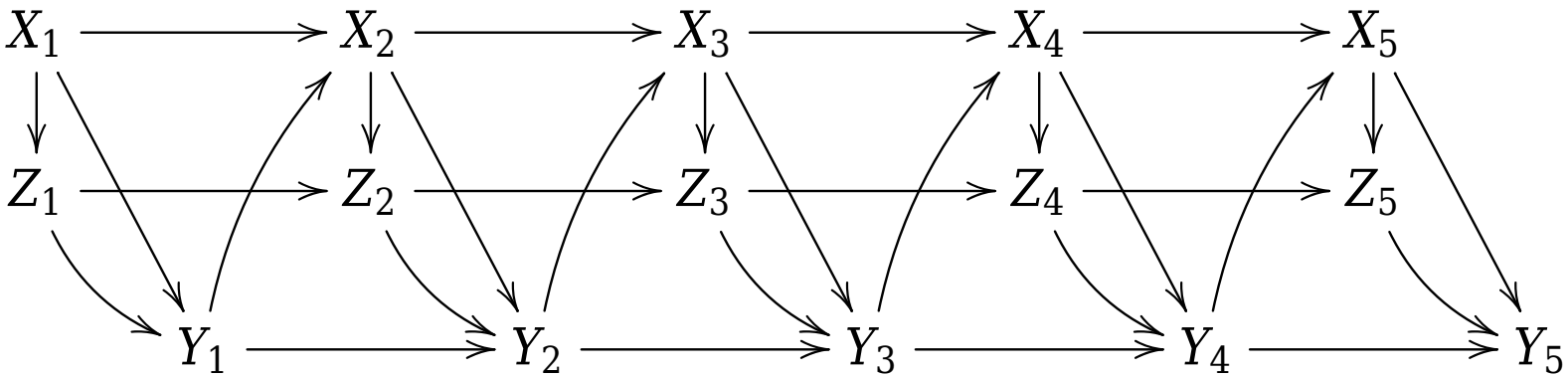
In this case, the possible effect of confounding of the exposure effect by the confounders is more complicated.

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

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

and then the corresponding sequence of potential outcomes

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

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.

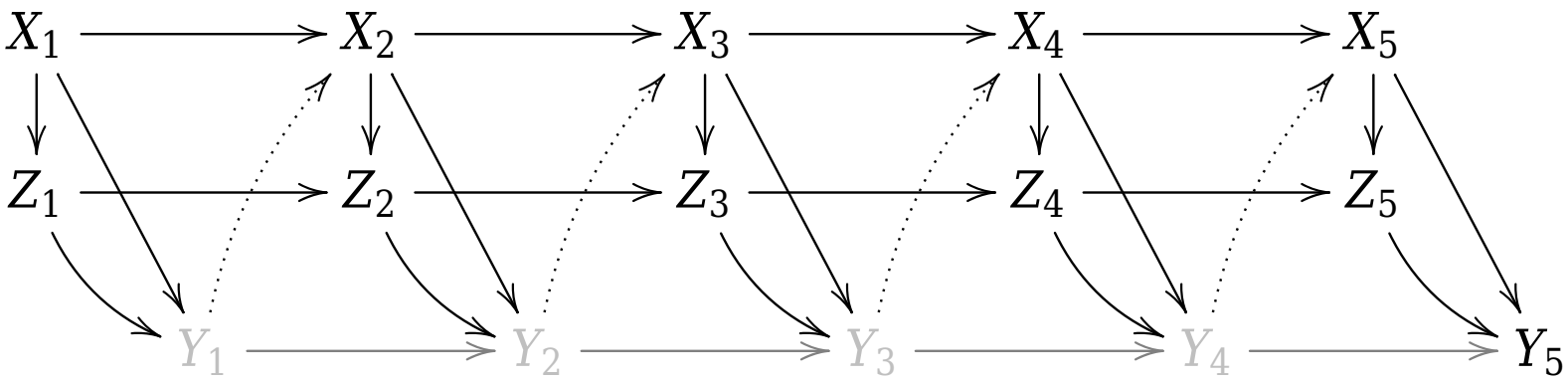
Estimation of Total Effects

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

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[8]

$$\log\left(\frac{\Pr(Y_{im} = 1|\widetilde{z}; \beta_0, \psi)}{\Pr(Y_{im} = 0|\widetilde{z}; \beta_0, \psi)}\right) = \beta_0 + \psi \sum_{j=1}^{m} z_j$$

as the (structural) *marginal* model.

[8] 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: The logic

- Inference is required under *hypothetical* population $\mathcal{E}$;
  - in population $\mathcal{E}$, the *conditional independence*

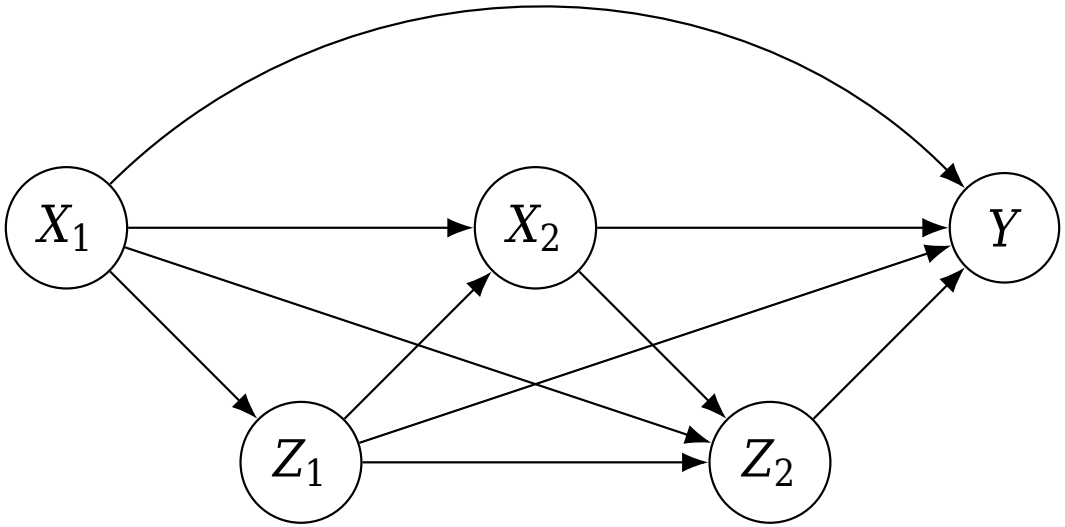
$$z_{ij} \perp\!\!\!\perp \tilde{x}_{ij} \mid \tilde{z}_{i(j-1)}$$

    holds.

- 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 $\mathcal{O}$.

Two time points

For a two time point setting:



In this formulation, the time ordering

$$X_1 \longrightarrow Z_1 \longrightarrow X_2 \longrightarrow Z_2 \longrightarrow Y$$

delimits the possible causal pathways.

Two time points

We can consider the expected counterfactual outcomes associated with treatment *patterns*

$$\mathbb{E}[Y(z_1, z_2)]$$

or equivalently

$$\mathbb{E}^{\mathcal{E}}_{Y|Z_1,Z_2}[Y|Z_1 = z_1, Z_2 = z_2]$$

where the experimental distribution $\mathcal{E}$ assumes randomized treatments.

Two time points

To learn about the APOs for different treatment patterns from observational data is not straightforward.

- Z_1 has a *direct* effect on Y, and a *mediated* effect via X_2 and Z_2;
- Z_2 has a *direct* effect on Y, but it is *confounded* by X_2; to remove this confounding we need to condition on X_2;
- However, conditioning on X_2 *blocks the directed path* from Z_1 to Y and hence affects the causal effect.

We cannot break the confounding by blocking paths by conditioning to get at the aggregate effect.

Marginal structural models

We may use *inverse weighting* to break the confounding as in the single interval case. For example, for APO

$$\mu(z_1, z_2) = \mathbb{E}^{\mathcal{E}}_{Y|Z_1,Z_2}[Y|Z_1 = z_1, Z_2 = z_2]$$

we may use the estimator

$$\widetilde{\mu}(z_1, z_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{1}_{\{z_1\}}(Z_{1i})\mathbb{1}_{\{z_2\}}(Z_{2i})}{f^{\mathcal{O}}_{Z_1,Z_2|X_1,X_2}(Z_{1i}, Z_{2i}|X_{1i}, X_{2i})} Y_i$$

Each outcome data point is re-weighted by the IPW weight across the whole treatment sequence.

Marginal structural models

In the re-weighted data, *model-based* analysis can also be used: for example, we could propose a marginal model

$$\mathbb{E}^{\mathcal{E}}_{Y|Z_1,Z_2}[Y|Z_1 = z_1, Z_2 = z_2] = \beta_0 + \psi_1 z_1 + \psi_2 z_2$$

or, using the *total* treatment

$$\mathbb{E}^{\mathcal{E}}_{Y|Z_1,Z_2}[Y|Z_1 = z_1, Z_2 = z_2] = \beta_0 + \psi_0(z_1 + z_2)$$

and then perform a *weighted least squares analysis* (WLS) to estimate (ψ_1, ψ_2) or ψ_0.

Such a model is termed a *marginal structural model* (MSM).

Marginal structural models

That is for example

$$(\widehat{\beta}_0, \widehat{\psi}_1, \widehat{\psi}_2) = \arg \min_{(\beta_0, \psi_1, \psi_2)} \sum_{i=1}^{n} w_i (y_i - \beta_0 - \psi_1 z_{1i} - \psi_2 z_{2i})^2$$

where

$$w_i = \frac{1}{f^{\mathcal{O}}_{Z_1, Z_2 | X_1, X_2}(z_{1i}, z_{2i} | x_{1i}, x_{2i})}$$

where

$$f^{\mathcal{O}}_{Z_1, Z_2 | X_1, X_2}(z_1, z_2 | x_1, x_2) = f^{\mathcal{O}}_{Z_1 | X_1}(z_1 | x_1) f^{\mathcal{O}}_{Z_2 | X_1, X_2, Z_1}(z_2 | x_1, x_2, z_1).$$

Marginal structural models

An alternative weight

$$w_i = \frac{f^{\mathcal{O}}_{Z_1,Z_2}(z_{1i}, z_{2i})}{f^{\mathcal{O}}_{Z_1,Z_2|X_1,X_2}(z_{1i}, z_{2i}|x_{1i}, x_{2i})}$$

could be used, where

$$f^{\mathcal{O}}_{Z_1,Z_2}(z_{1i}, z_{2i}) = f^{\mathcal{O}}_{Z_1}(z_1) f^{\mathcal{O}}_{Z_2|Z_1}(z_2|z_1).$$

This generalizes the earlier form of IPW estimator.

If a *non-parametric* model for $f^{\mathcal{O}}_{Z_1,Z_2}(z_{1i}, z_{2i})$ is adopted, then the new weight essentially reduces to the original weight.

Marginal structural models

Using a *parametric* model,

$$w_i = \frac{f^{\mathcal{O}}_{Z_1,Z_2}(z_{1i}, z_{2i}; \widehat{\alpha})}{f^{\mathcal{O}}_{Z_1,Z_2|X_1,X_2}(z_{1i}, z_{2i}|x_{1i}, x_{2i}; \widehat{\gamma})}$$

where

- $\alpha = (\alpha_1, \alpha_2)$ is estimated from the model

$$f^{\mathcal{O}}_{Z_1,Z_2}(z_1, z_2; \alpha) = f^{\mathcal{O}}_{Z_1}(z_1; \alpha_1) f^{\mathcal{O}}_{Z_2|Z_1}(z_2|z_1; \alpha_2)$$

- $\gamma = (\gamma_1, \gamma_2)$ is estimated from the model

$$f^{\mathcal{O}}_{Z_1,Z_2|X_1,X_2}(z_1, z_2|x_1, x_2; \gamma) = f^{\mathcal{O}}_{Z_1|X_1}(z_1|x_1; \gamma_1) \times f^{\mathcal{O}}_{Z_2|X_1,X_2,Z_1}(z_2|x_1, x_2, z_1; \gamma_2)$$

Marginal structural models

It is also possible to carry out a *conditional* analysis given baseline predictors X_1 which are not subject to the influence of any treatment: for example

$$\mathbb{E}^{\mathcal{E}}_{Y|Z_1,Z_2,X_1}[Y|Z_1 = z_1, Z_2 = z_2, X_1 = x_1] = \beta_0 + x_1\beta_1 + \psi_1 z_1 + \psi_2 z_2$$

for which the so-called *stabilized* weights

$$w_i = \frac{f^{\mathcal{O}}_{Z_1,Z_2|X_1}(z_{1i}, z_{2i}|x_{1i})}{f^{\mathcal{O}}_{Z_1,Z_2|X_1,X_2}(z_{1i}, z_{2i}|x_{1i}, x_{2i})}$$

should be used.

Marginal structural models

Separate models are needed for numerator and denominator

- Denominator:

$$f^{\mathcal{O}}_{Z_1,Z_2|X_1,X_2}(z_1,z_2|x_1,x_2) = f^{\mathcal{O}}_{Z_1|X_1}(z_1|x_1) f^{\mathcal{O}}_{Z_2|X_1,X_2,Z_1}(z_2|x_1,x_2,z_1)$$

- Numerator:

$$f^{\mathcal{O}}_{Z_1,Z_2|X_1,X_2}(z_1,z_2|x_1) = f^{\mathcal{O}}_{Z_1|X_1}(z_1|x_1) f^{\mathcal{O}}_{Z_2|X_1,Z_1}(z_2|x_1,z_1)$$

parameterized by α and γ respectively, say.

That is, after cancelling terms,

$$w_i = \frac{f^{\mathcal{O}}_{Z_2|X_1,Z_1}\left(z_2|x_1,z_1;\widehat{\alpha}\right)}{f^{\mathcal{O}}_{Z_2|X_1,X_2,Z_1}\left(z_2|x_1,x_2,z_1;\widehat{\gamma}\right)}$$

This weight may be less extreme than the unstabilized counterpart.

Note that conditioning on X_1 in the outcome model is necessary to account for possible confounding.

Note

The term ‘stabilized’ is slightly misleading; the introduction of the numerator term *changes the estimand*, so the original and stabilized versions of the MSM estimate *different* quantities.

In many cases the stabilized weights will be more uniform, and this has the effect of reducing estimator variance, but the estimation target is changed.

MSM Example

Part 5

Practical Challenges

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.

Data complexity

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.

Optimal Treatment Strategies

Causal methods can be used to develop *optimal treatment strategies*

- observational data and causal adjustment methods lead to inference about treatment effects
- this information can be used to develop optimal approaches to treatment decision-making for future studies
- choose the treatment that leads to the best possible outcome

Typically, the optimal treatment rule is

- *tailored* to the individual subject
- *adaptive* to treatment and intermediate outcome history

Part 6

Conclusions

Conclusions

- Causal inference methods provide answers to important questions concerning the impact of hypothetical exposures;
- Causal graphs are useful in formulating inference;
- 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.

Key remaining challenges

- Flexible representations of model components;
- Model selection;
- Scale and complexity of observational data;

Collaborators & Acknowledgments

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





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)

References

- The marginal structural model
  - Hernán et al. (2000)
  - Hernán et al. (2001)
- Model selection
  - Brookhart et al. (2006)
- Longitudinal studies
  - Moodie and Stephens (2012)
  - Graham et al. (2014)

References

- High-dimensional settings
  - Schneeweiss et al. (2009)
- Advanced modelling methods
  - Chernozhukov et al. (2018)
  - Alam et al. (2019)
- Bayesian methods
  - Rubin (1978)
  - McCandless et al. (2009)
  - An (2010)
  - Saarela et al. (2015)

References

Alam, S., Moodie, E. E. M., and Stephens, D. A. (2019). Should a propensity score model be super? the utility of ensemble procedures for causal adjustment. Statistics in Medicine **38,** 1690–1702.

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.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal **21,** C1–C68.

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.

References

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.

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.

References

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.

Appendix: Monte Carlo Methods

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

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

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

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

Back

Monte Carlo & Importance sampling

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 $\mathbb{E}[g(Y)]$.

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

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

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

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

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

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

Importance sampling is based on forming *weighted* averages

- we re-weight samples from f_Y^* so that we can estimate quantities relating to f_Y
- this is like '*standardization*' (eg standardized mortality rate) in epidemiology.

Appendix: Confounding Bias Example

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

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

and

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

Back

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

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

say.

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.

Simple confounding example

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*.

Appendix: Probability & Causal Graphs

Probabilistic models

A *joint* probability distribution

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

describes how the data are generated. This model specifies

- the *marginal* distributions

  $$f_X(x) \qquad f_Y(y) \qquad f_Z(z)$$

  that describe how the variables behave individually,
- the *conditional* distributions such as

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

  etc. that describe how the variables behave when one or more variable is *fixed*

Back

We have the possible decompositions

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

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

and so on, for any ordering of the variables.

We can *always* consider this kind of sequential decomposition, which is termed a *chain rule factorization*.

Independence

Two random variables X, Z are *independent*

$$X \perp\!\!\!\perp Z$$

if and only if, for all values (x, z),

$$\begin{aligned} f_{X,Z}(x, z) &= f_X(x) f_Z(z) \\ f_{Z|X}(z|x) &= f_Z(z) \\ f_{X|Z}(x|z) &= f_X(x) \end{aligned}$$

i.e. knowledge of X does not influence our assessment of Z.

Independence

We can consider *conditional independence*: say

$$Y \perp\!\!\!\perp Z \mid X$$

if and only if, *for all* (x, z, y)

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

i.e. knowledge of Y does not influence our assessment of Z if we *fix* $X = x$.

Example

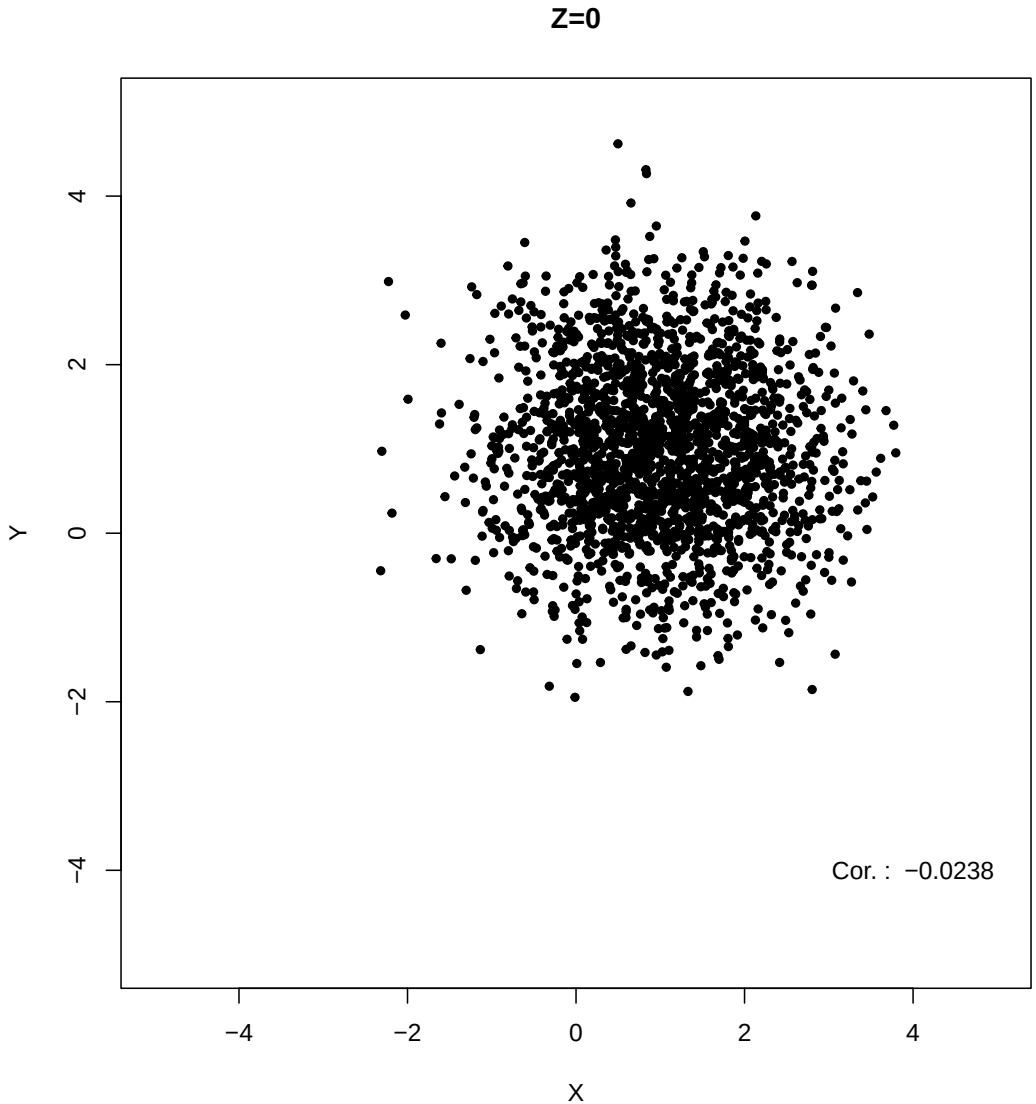
Three variables

- X and Y vary continuously,
- Z is binary.

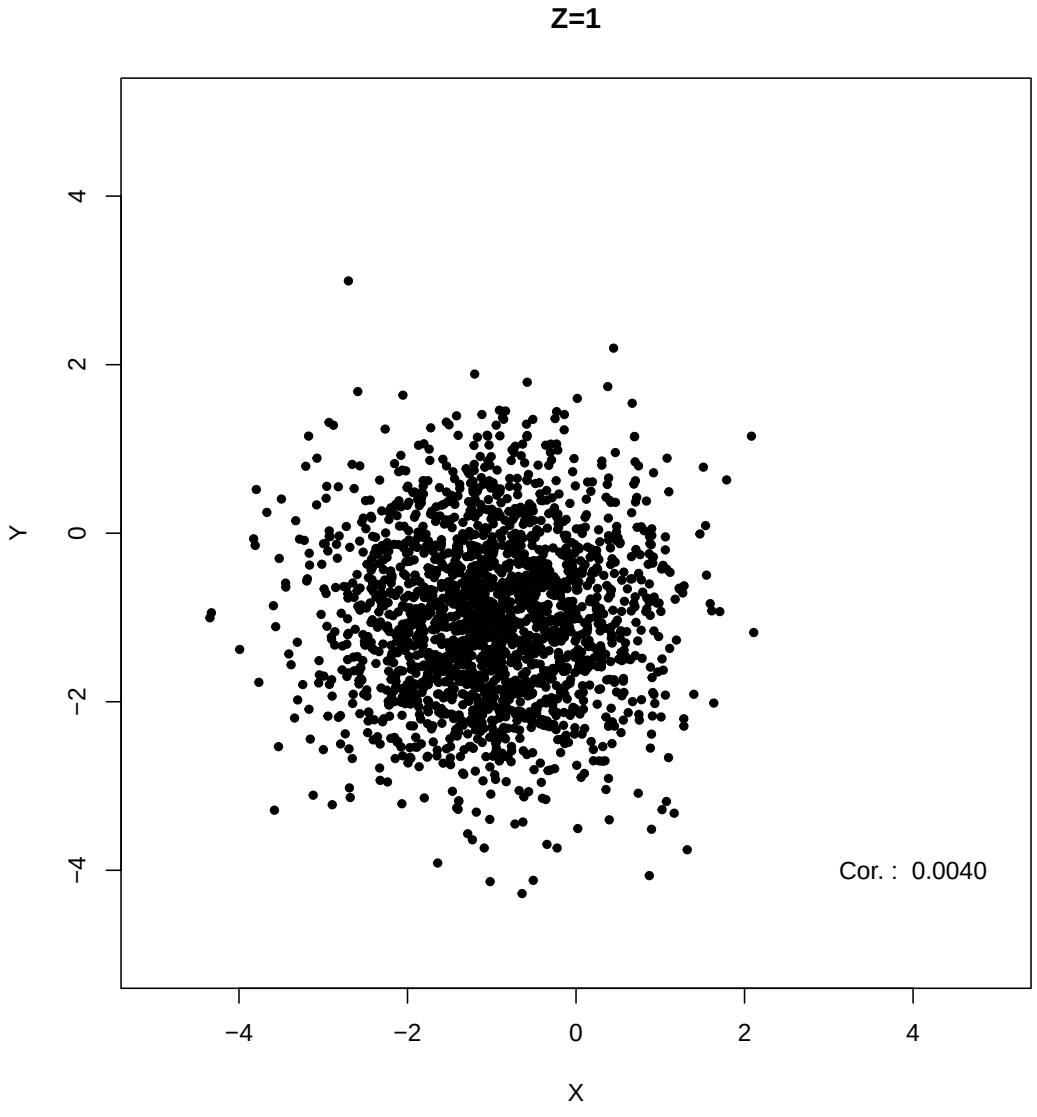
We can study the distribution of the data for X and Y

- for each level of Z separately,
- pooled over Z levels.

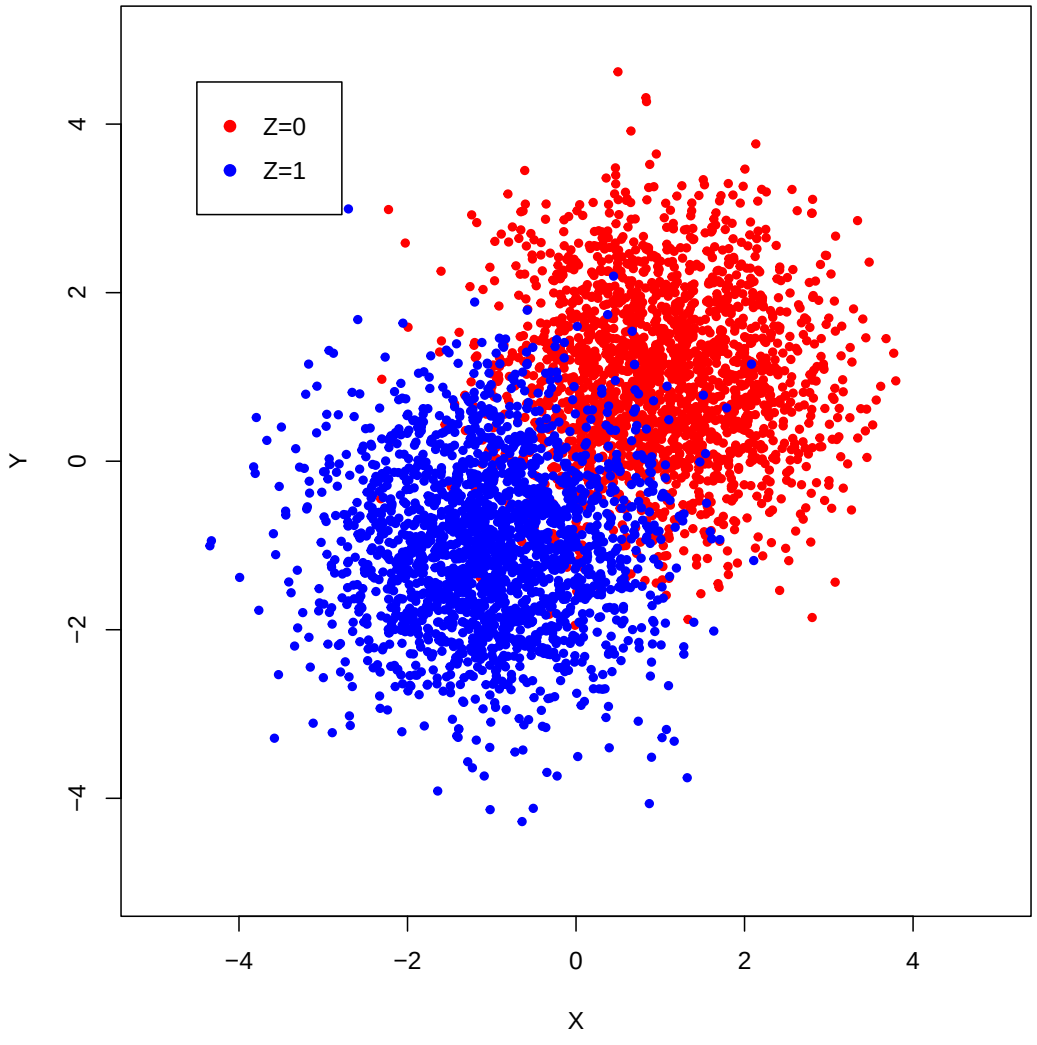
Example



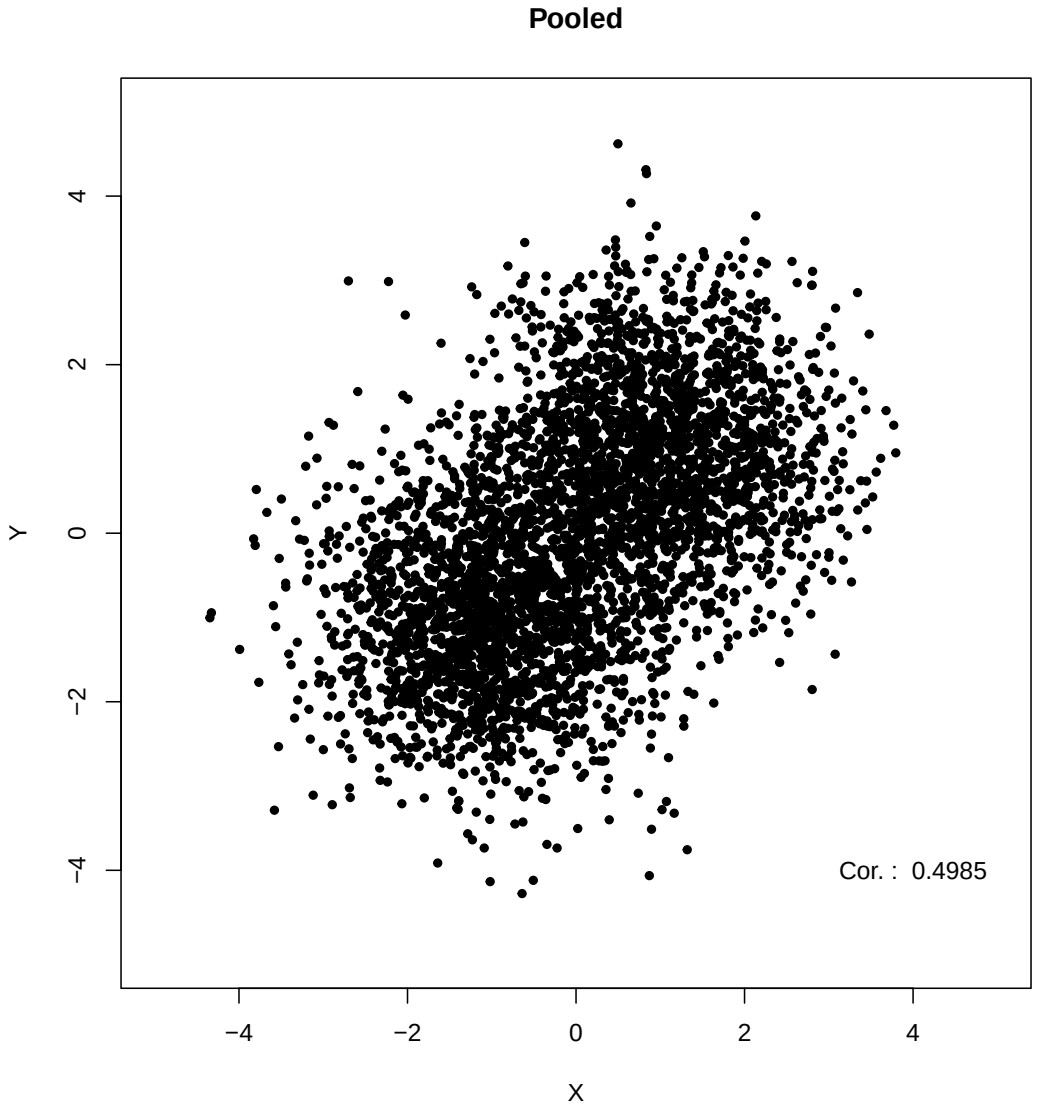
Example



Example



Example



Example

We see that

- for $Z = 0$ and $Z = 1$ separately, X and Y are *uncorrelated*;
- overall X and Y are *positively correlated*.

Thus X and Y are

conditionally unrelated given Z

but are

unconditionally related.

This illustrates that conditioning can remove (or *block*) dependence.

We have a chain rule factorization

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

We might then *assume* the *conditional independence*

$$Z \perp\!\!\!\perp Y|X$$

so that

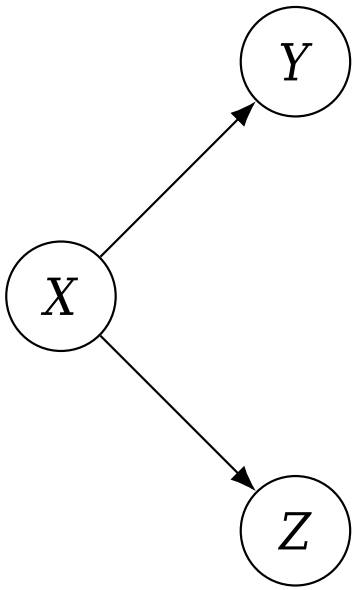
$$f_{Z|X,Y}(z|x,y) = f_{Z|X}(z|x)$$

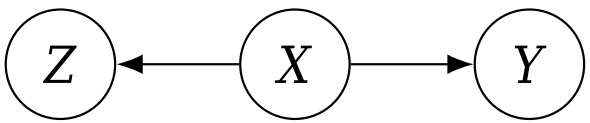
and so

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

Causal Graphs

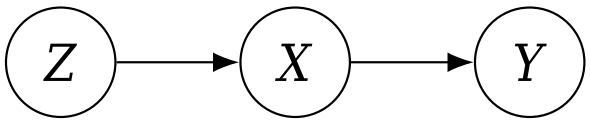
We can depict the conditional independence using a graph:





This type of graph is sometimes called a *fork*.

The other common type of graph component is a *chain*



which implies the factorization

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

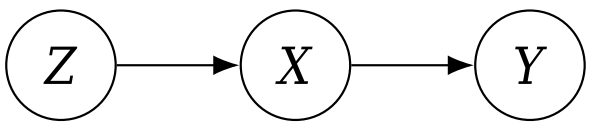
and the conditional independence

$$Y \perp\!\!\!\perp Z|X$$

That is, there are two ways the conditional independence

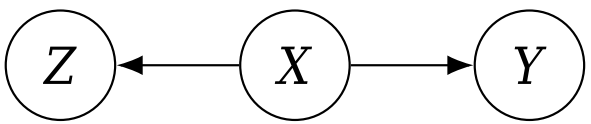
$$Y \perp\!\!\!\perp Z|X$$

could be represented



Chain

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



Fork

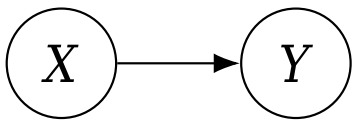
$$f_X(x)f_{Z|X}(z|x)f_{Y|X}(y|x)$$

Causal Graphs

- Nodes Ⓧ, Ⓨ, Ⓩ denote the variables;
- Edges with *arrows* indicate the nature of dependence in the chain rule factorization;
- *Directed* arrows specify the conditional independence assumptions;

Causal Graphs

- Nodes without *incoming* edges are *founders*;



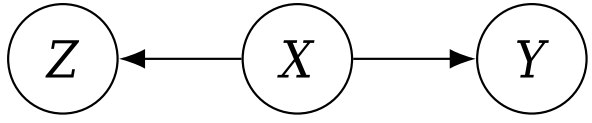
corresponds to

$$f_X(x) f_{Y|X}(y|x)$$

Causal Graphs

- Nodes with only *outgoing* edges act to *block* dependence.

  For example, in



  so that

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

  it follows that

$$Z \perp\!\!\!\perp Y | X.$$

  However, it also follows that, in general

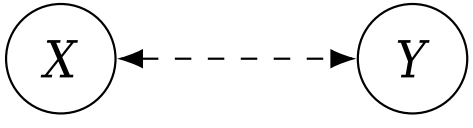
$$Y \not\perp\!\!\!\perp Z$$

  (recall the earlier scatterplots)

Causal Graphs

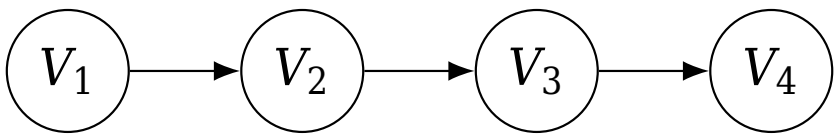
- *Nodes* or *vertices*, $V_1, V_2, \ldots$, represent variables.
- *Edges*, $E_1, E_2, \ldots$, represent dependencies.
- Two nodes are *adjacent* if there is an edge between them.
  - edges can be *directed*, denoted using arrows, or *undirected*;
  - if all edges are directed, the graph is directed.

Note: we can use 'bidirected' (edges with an arrow at each end) to indicate general dependence between two variables



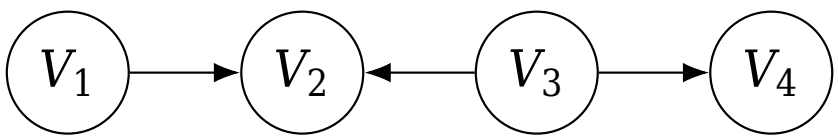
although these will be less important in causal settings.

- A *path* is a sequence of edges that connects two nodes;
  - a *directed* path is a path where the directions of arrows on edges are obeyed



Directed path from V_1 to V_4

whereas an *undirected* path is a path that is not directed.

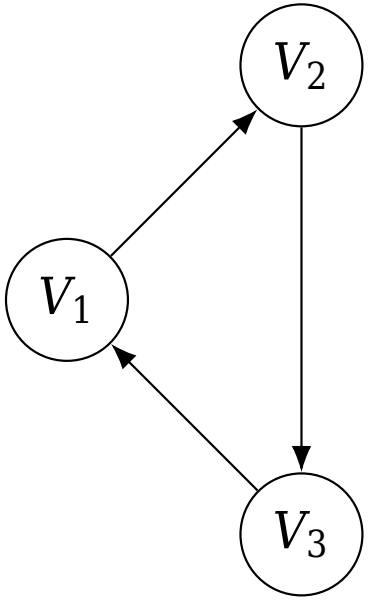


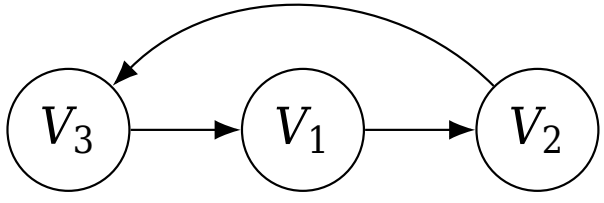
Undirected path from V_1 to V_4

  - two nodes are *connected* if a path exists between them, and *disconnected* otherwise.

Causal Graphs

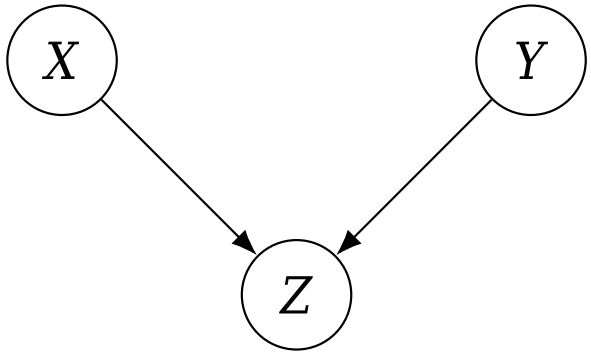
In general, a graph may contain *cycles*, that is, directed paths that *start* and *end* at the *same* node.





Directed acyclic graph (DAG): a directed graph with no cycles.

Causal Graphs



PARENTS

CHILD

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

In this DAG, we have $X \perp\!\!\!\perp Y$:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

e.g. X and Y represent the scores on two dice rolled independently, Z is the total score

$$Z = X + Y.$$

However, conditioning on $Z = z$

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

in general. That is,

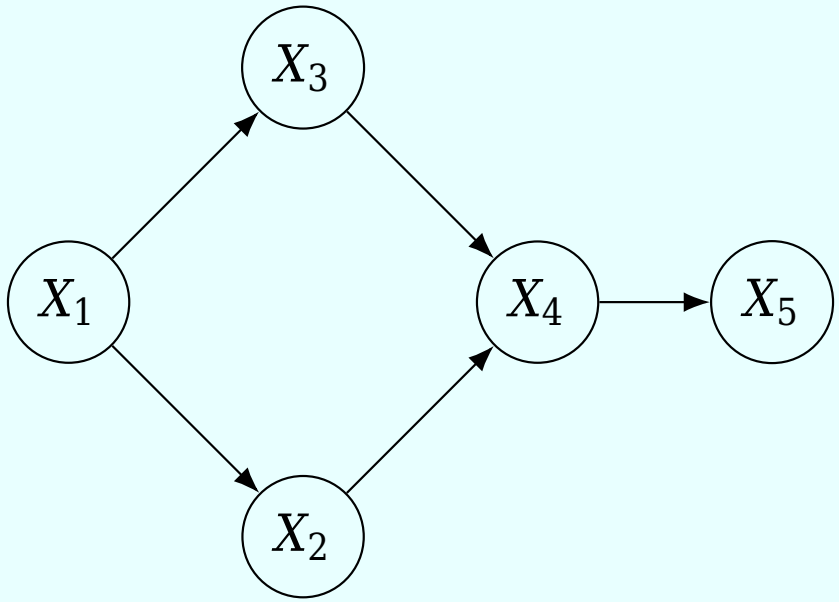
$$X \perp\!\!\!\perp Y$$

but

$$X \not\perp\!\!\!\perp Y \mid Z$$

Conditioning on Z *induces dependence*; the node is termed a *collider*.

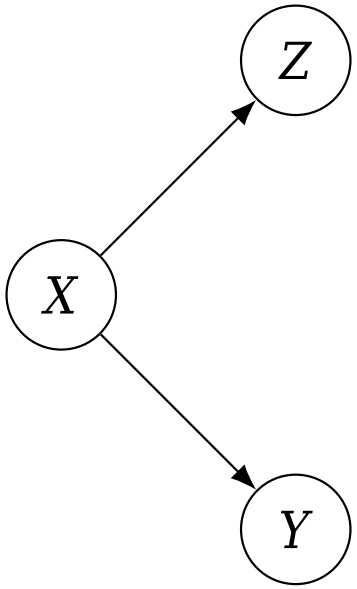
Example: Factorization



$$f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1}(x_3|x_1) f_{X_4|X_2,X_3}(x_4|x_2, x_3) f_{X_5|X_4}(x_5|x_4)$$

Structural models and equations

When we write



what precisely does the symbol $\longrightarrow$ mean ?

Structural models and equations

A *structural* interpretation states that we

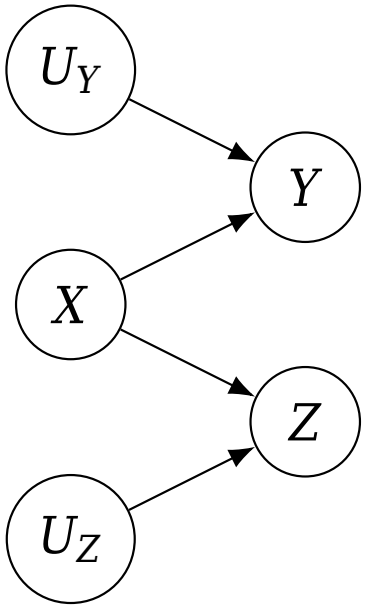
- generate X independently,
- generate Y and Z independently as functions of the realized X, for example

$$Y = 3X$$

$$Z = 4X + 9$$

Structural models and equations

X, U_Z, U_Y independent



$$Y = g_1(X, U_Y)$$

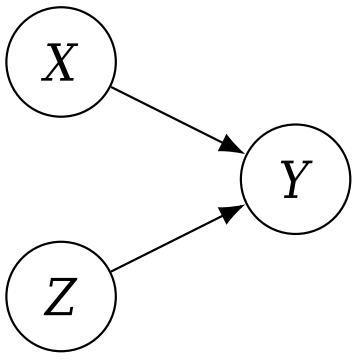
$$Z = g_2(X, U_Z)$$

For example

$$Y = X + U_Y$$

$$Z = X + U_Z$$

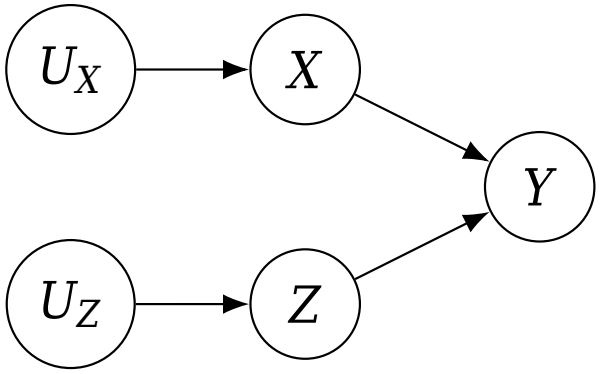
Structural models and equations



$$Y = g(X, Z)$$

Fixing $X = x$ and $Z = z$ fixes $Y = g(x, z)$.

Structural models and equations



$$X = g_1(U_X)$$

$$Z = g_2(U_Z)$$

$$Y = g(X, Z)$$

If we know $X = x$ and $Z = z$, then we do not need to know the values of U_X and U_Z to determine Y. That is

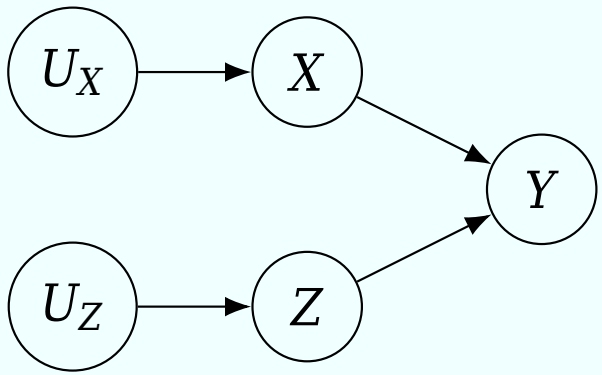
$$Y \perp\!\!\!\perp (U_X, U_Z) \mid (X, Z).$$

We can interpret *causation* in terms of these functions.

Structural models and equations

- X *causes* Y if it appears in the function, g, that assigns Ys value;
- X *causes* Y if, in the graph representing the joint distribution, there is a *directed path* from X to Y;
- X is a *direct cause* of Y if there is an arrow from X to Y.

Note



$$X = g_1(U_X)$$

$$Z = g_2(U_Z)$$

$$Y = g(X, Z)$$

so that

$$Y = g(X, Z) = g(g_1(U_X), g_2(U_Z))$$

so both (X, Z) and (U_X, U_Y) can be interpreted as causes of Y.

- X and Z are direct causes,
- U_X and U_Y are indirect causes.

Note

We will proceed by assuming that in a practical setting, the structural relationship and the corresponding causal graph is *known* before any analysis can be carried out.

- Usually in practice this requires expert knowledge;
- Learning the causal graph from data is a hard problem;
- If we obtain all variables simultaneously, it is not possible to learn which of the possible factorizations is the data generating one; for example, we cannot distinguish

$$f_X(x)f_{Y|X}(y|x) \qquad \text{from} \qquad f_Y(y)f_{X|Y}(x|y)$$

i.e. does X cause Y or does Y cause X ?

Note

In order for X to cause Y, we must have that X *precedes* Y temporally.

The structural equations form the variables on the left hand side from the variables on the right hand side

$$Y = g(X, Z)$$

that is, we *first* generate X and Z, and *then* generate Y.

That is, there must be a *temporal* ordering.

d-separation

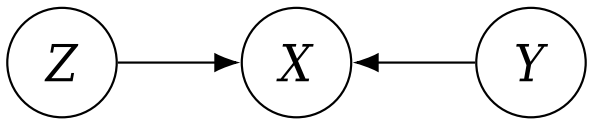
To assess whether

$$Y \perp\!\!\!\perp Z \qquad \text{or} \qquad Y \perp\!\!\!\perp Z \,|\, X$$

for any distribution compatible with the DAG, we must assess whether there is any way for 'information' to 'flow' between Z and Y, maybe once X has been accounted for.

d-separation

First, recall the *collider* graph



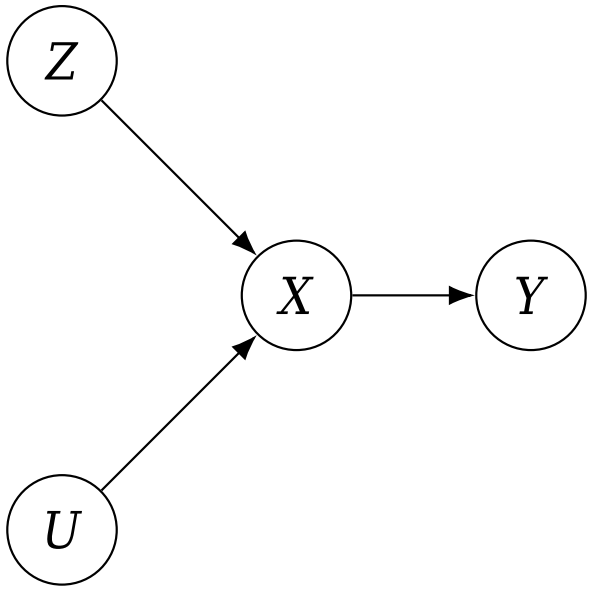
X is a collider on the path between Z and Y. Therefore

$$Y \perp\!\!\!\perp Z \qquad \text{but} \qquad Y \not\perp\!\!\!\perp Z \mid X$$

Note that a *directed path* from one node to another *cannot* contain a collider.

d-separation

The notion of being a collider is *path-specific*: for example



- X is a *collider* on path $Z \to X \to U$
- X is *not a collider* on path $Z \to X \to Y$.

Unconditional d-separation

Consider a general path (directed or undirected) between Z and Y.

The path is *open* (or *unblocked*) if there is *no collider* on the path;

- if there is a collider, the path is *closed* (*blocked*).

Z and Y are *d-separated* if there is *no open path between them*;

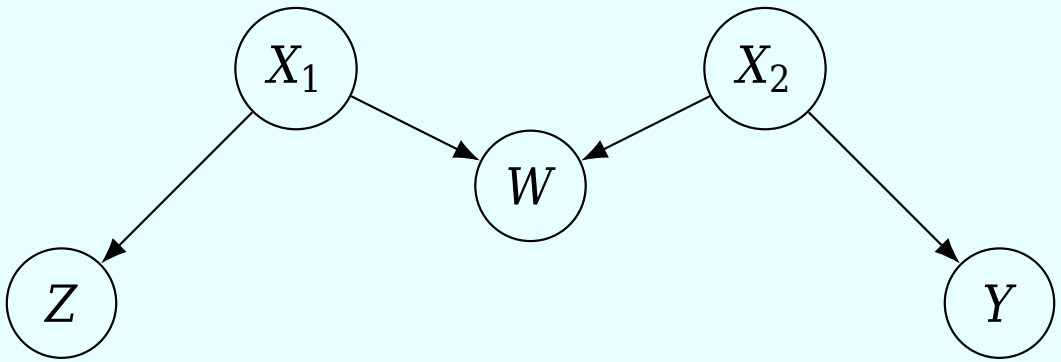
If there is an open path, Z and Y are *d-connected*.

- this path must comprise *chains* or *forks* only

Unconditional d-separation

Example: Diabetes example (Rothman et al. p 188)

- X_1 family income
- X_2 genetic risk
- W parental diabetes
- Z low educational attainment
- Y diabetes of subject



Unconditional d-separation

Example: Diabetes example (Rothman et al. p 188)

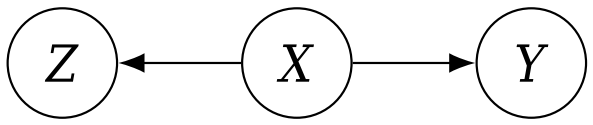
Z and Y are d-separated; there is one path between Z and Y, but it is blocked by the collider W.

$$f_{X_1}(x_1)f_{X_2}(x_2)f_{W|X_1,X_2}(w|x_1,x_2)f_{Z|X_1}(z|x_1)f_{Y|X_2}(y|x_2)$$

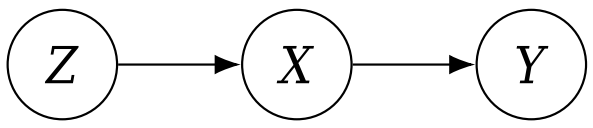
and Z and Y are *independent*.

Conditional d-separation

For a *non-collider* X: *conditioning* on X:



$$Z \perp\!\!\!\perp Y \mid X$$

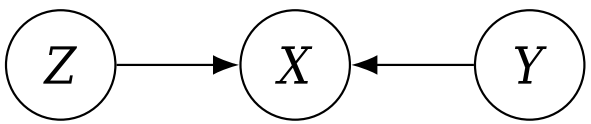


$$Z \perp\!\!\!\perp Y \mid X$$

Conditioning *blocks* the path.

Conditional d-separation

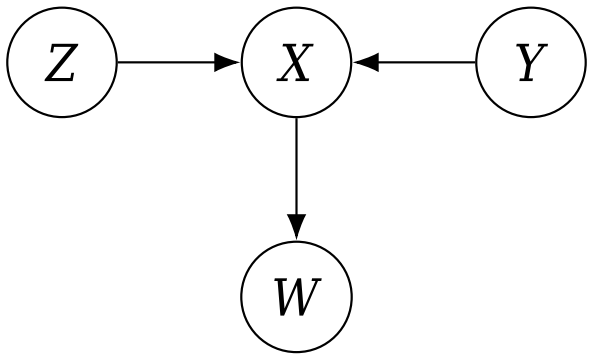
For a *collider* X: conditioning on X *opens* the path



$$Z \not\perp\!\!\!\perp Y \mid X$$

Conditional d-separation

Consider the following DAG:



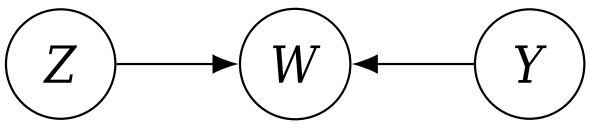
and conditioning on *a descendant*, W, of X:

$$f_{Z,Y,W}(x,y,w) = f_Z(z)f_Y(y)\int f_{X|Z,Y}(x|z,y)f_{W|X}(w|x)\,dx$$

$$= f_Z(z)f_Y(y)f_{W|Z,Y}(w|z,y)$$

Conditional d-separation

Therefore we have that



$$Z \not\perp\!\!\!\perp Y \mid W$$

and so W *is a collider* in the reduced graph.

Conditional d-separation

Therefore

(i) conditioning on a *non-collider* X *blocks* the path at X;

(ii) conditioning on a *collider* X or a *descendant* W of X *opens* the path at X;

Conditional d-separation

Consider two nodes X and Y with possibly several open paths connecting them. Suppose S is a set of variables.

- S *blocks* a path if, after conditioning on S, the path is *closed*;
- S *unblocks* a path if after conditioning the path is *open*;
- If S blocks *every path*, then X and Y are *d-separated* by S.

Conditional d-separation

- If S d-separates X and Y, then

$$X \perp\!\!\!\perp Y \mid S,$$

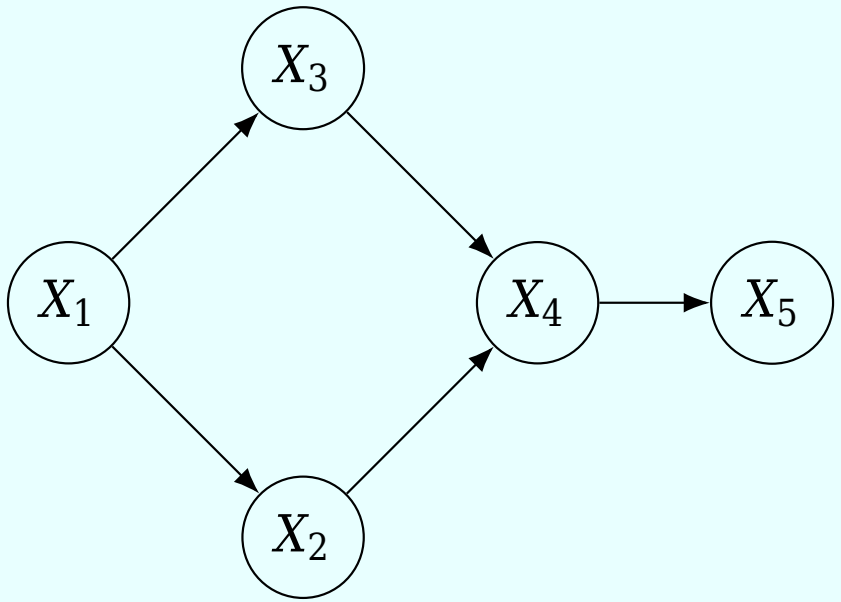
so that

$$f_{X|Y,S}(x|y,s) \equiv f_{X|S}(x|s) \quad \forall (x,y,s).$$

X and Y are *conditionally independent* given S.

Conditional d-separation

Example:

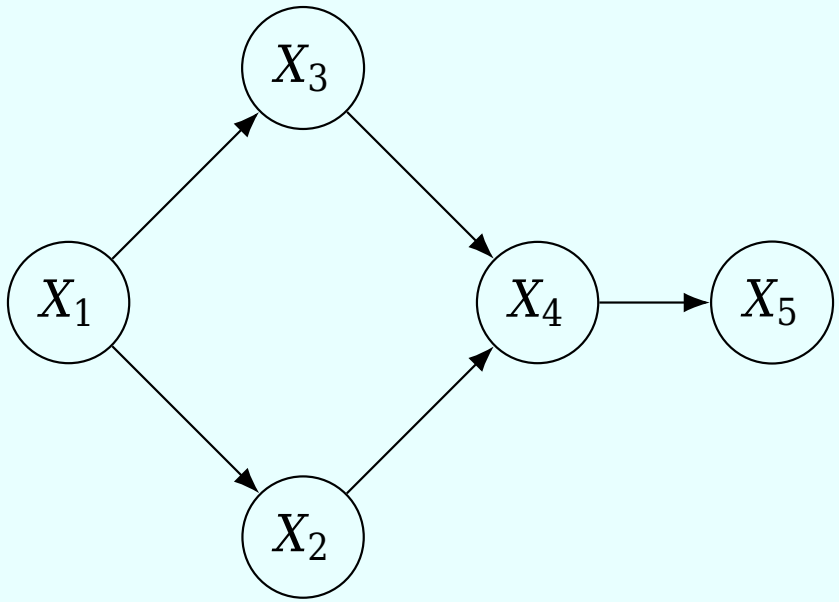


$\{X_2\}$ and $\{X_3\}$ are d-separated by $\{X_1\}$, and $X_2 \perp\!\!\!\perp X_3 \mid X_1$.

- there are two paths between X_2 and X_3;
  - $X_2 \to X_1 \to X_3$: blocked by conditioning on X_1.
  - $X_2 \to X_4 \to X_3$: blocked by the collider at X_4, and $X_4 \notin \{X_1\}$.

Conditional d-separation

Example:



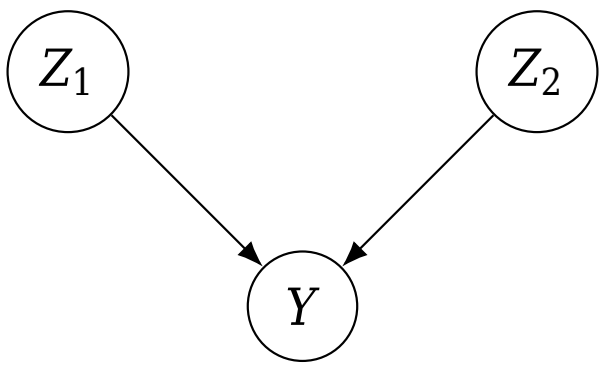
$\{X_2\}$ and $\{X_3\}$ are *not* d-separated by $\{X_1, X_5\}$:

- $X_2 \not\perp\!\!\!\perp X_3 \mid (X_1, X_5)$.
- X_5 is a descendant of collider X_4;

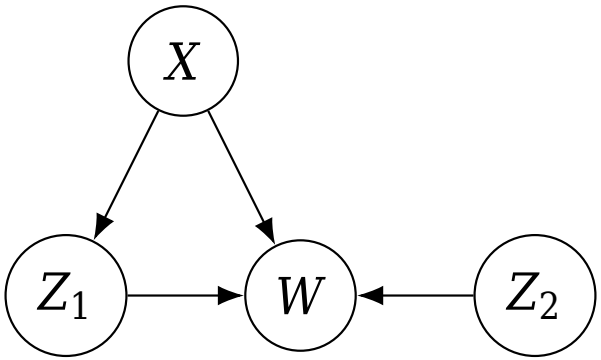
Selection bias

Conditioning on the common effect of two causes renders the two causes dependent;

- this is known as *selection bias* or *Berkson bias*
- it is the effect we observe in the collider graph



Selection bias

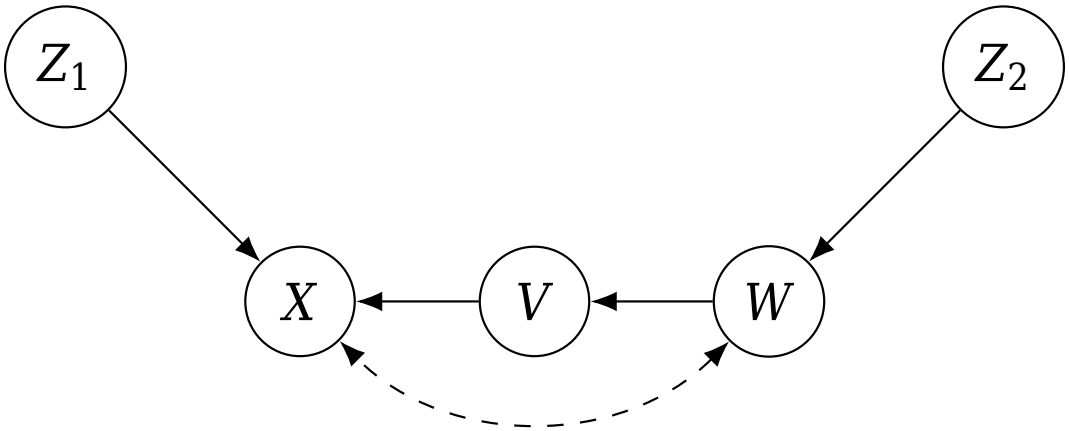


Here $Z_1 \perp\!\!\!\perp Z_2$: there are two paths to consider

- $Z_1 \to X \to W \to Z_2$
- $Z_1 \to W \to Z_2$

both blocked by collider W. Therefore $Z_1 \not\perp\!\!\!\perp Z_2 \mid \{W\}$.

Selection bias



Here $Z_1 \perp\!\!\!\perp Z_2$: there are two paths to consider

- $Z_1 \to X \to V \to W \to Z_2$ is blocked by the collider X.
- $Z_1 \to X \to W \to Z_2$ is blocked by the colliders X and W.

Therefore $Z_1 \not\perp\!\!\!\perp Z_2 \mid \{X, W\}$.

d-separation implication

If X and Y are d-separated by S then

$$X \perp\!\!\!\perp Y \mid S$$

for *all* distributions compatible with the graph; conversely, if they are not d-separated, then X and Y are *dependent* given S for at least one distribution compatible with the graph.

Interventions

Intervening set the level of Z to z has the effect of

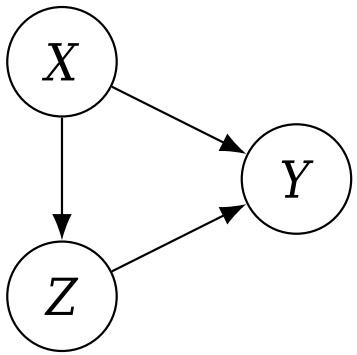
- removing all *incoming* arrows to Z
- switching the marginal for Z to the *degenerate distribution* $f_Z^*(.)$

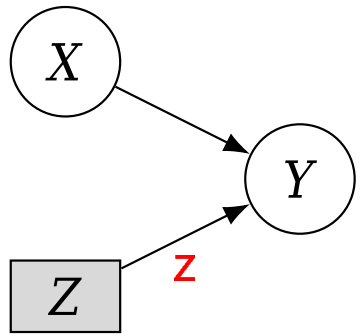
$$f_Z^*(z) = \mathbb{1}_{\{z\}}(z) \quad z \in \mathbb{R}.$$

That is, Z takes the value z with probability 1.

Interventions

In the earlier example

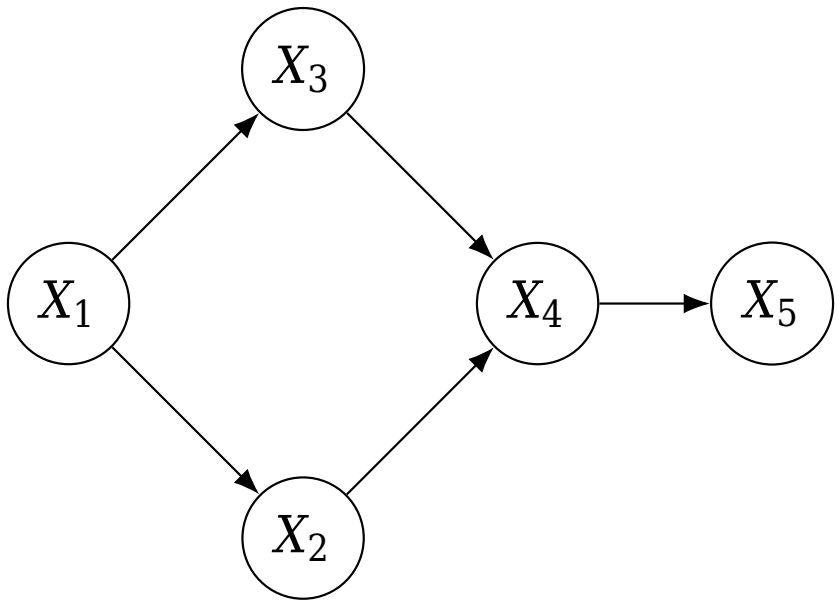




$$f_X(x) f_{Z|X}(z|x) f_{Y|X,Z}(y|x,z)$$

$$f_X(x) f^*_Z(z) f_{Y|X,Z}(y|x,z)$$

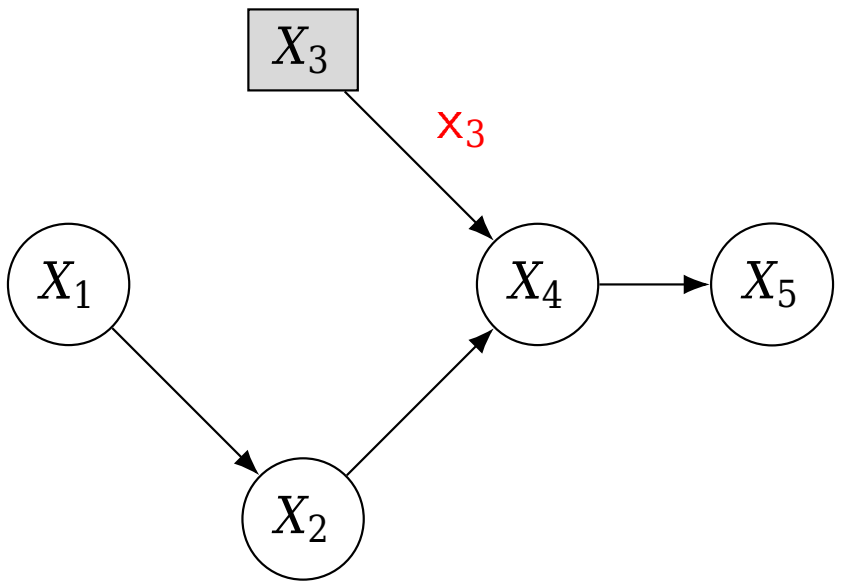
Consider the DAG



$$f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1}(x_3|x_1) f_{X_4|X_2,X_3}(x_4|x_2, x_3) f_{X_5|X_4}(x_5|x_4)$$

Interventions

Suppose we *intervene* to set $X_3 = x_3$. The relevant DAG is



$$f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f^*_{X_3}(x_3) f_{X_4|X_2,X_3}(x_4|x_2, x_3) f_{X_5|X_4}(x_5|x_4)$$

and X_1 is *no longer a cause* of X_3.

Graphical representation of bias

We aim to understand the effect of Z on Y.

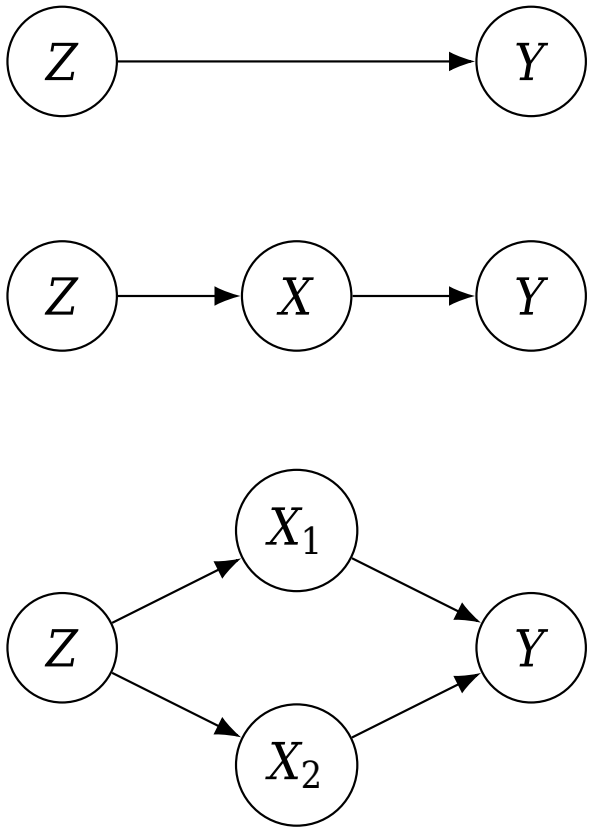
- An *open* undirected path between Z and Y allows for the *association* between Z and Y to be *modified* by the presence of other variables.

  This is known as a *biasing* path.

- By *'association'*, we mean some form of *correlation*.

Graphical representation of bias

- The *association* between Z and Y is *unbiased* for the effect of Z on Y if the only open paths between them are *directed paths*.



Graphical representation of bias

A set of nodes S is *sufficient* to control bias in the association between Z and Y if

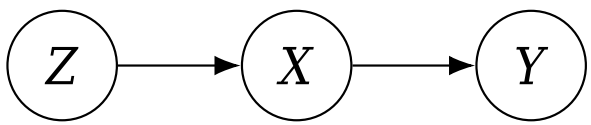
- conditional on S, the *open* paths between Z and Y are precisely the *directed* paths between Z and Y.

S is *minimally sufficient* if it is the smallest sufficient set.

Graphical representation of bias

Note: Conditioning on descendants of Z

(i) *blocks* directed paths

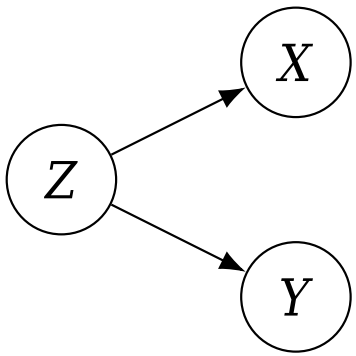


$Z \perp\!\!\!\perp Y \mid X$ but $Z \not\perp\!\!\!\perp Y$

(ii) may *unblock* or *create* paths that lead to *biasing* of the effect of Z on Y.

Graphical representation of bias

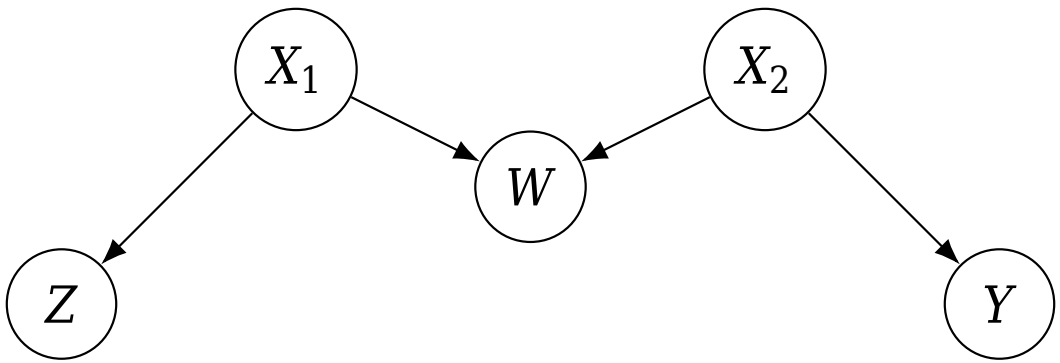
(iii) may be unnecessary in statistical terms: for example



In this graph, conditioning on X will not affect bias.

Graphical representation of bias

Undirected paths from Z to Y are termed *backdoor* paths (relative to Z) if they *start* with an arrow pointing *into* Z.



The only path from Z to Y is a backdoor path; however, it is not open because of the collider W.

Graphical representation of bias

Before conditioning

- *all biasing* paths in a DAG are backdoor paths, and
- all *open* backdoor paths are biasing paths.

To obtain an unbiased estimate of the effect of Z on Y, all backdoor paths between Z and Y must be *blocked*.

Graphical representation of bias

Set S satisfies the *backdoor criterion* with respect to Z and Y if

(i) S *contains no descendant* of Z, and

(ii) there is *no open backdoor path* from Z to Y after *conditioning* on S.

Confounding

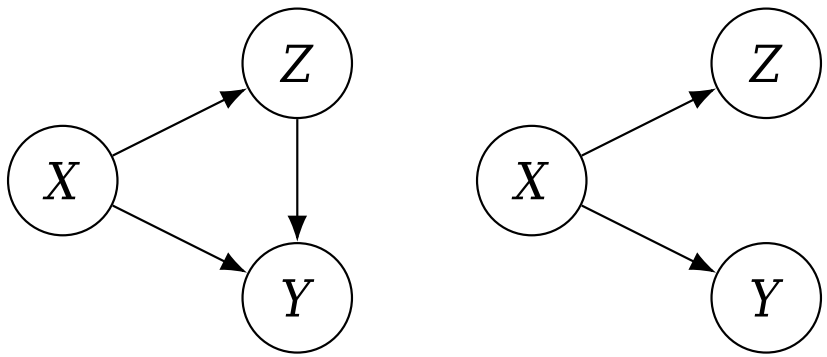
A *confounding path* between Z and Y is

(i) a *biasing* path (that is, an *undirected open path*) that

(ii) *ends* with an arrow into Y.

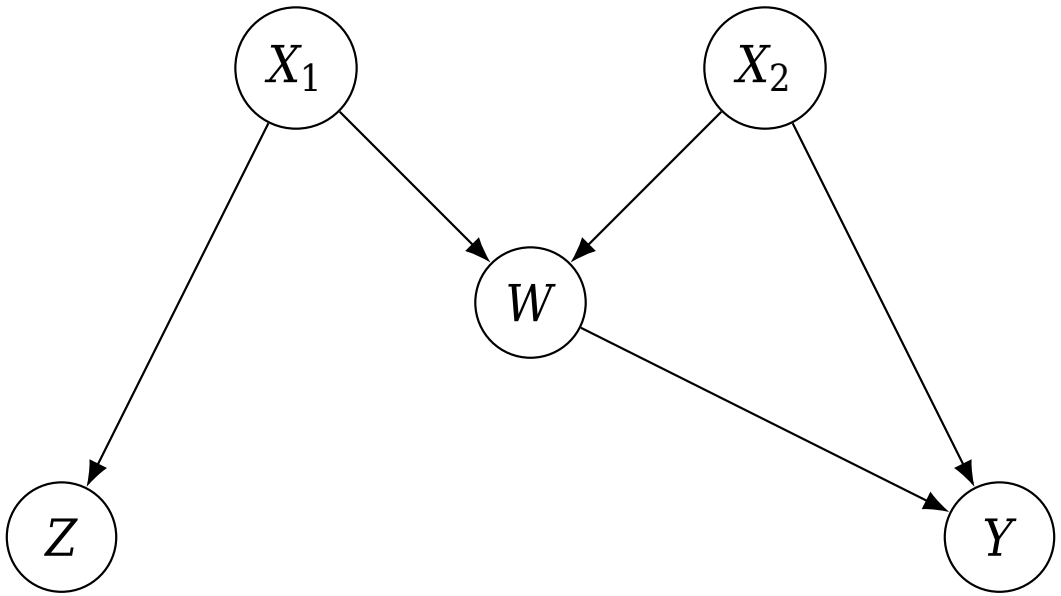
Confounding

Variables on a confounding path are termed *confounders*.



X is a confounder in both cases.

Confounding



W is a collider on the undirected path from Z to Y

$$\text{Path 1: } Z \to X_1 \to W \to X_2 \to Y$$

and hence this path is blocked.

Confounding

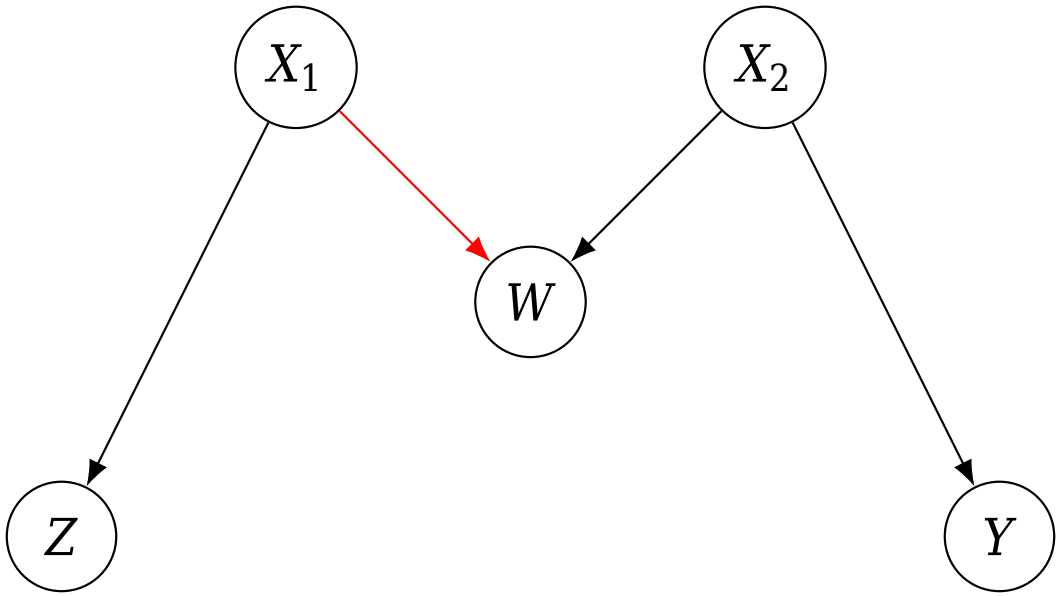
However unconditional on W, the effect of Z on Y is confounded by the backdoor path

$$\text{Path 2: } Z \to X_1 \to W \to Y.$$

Conditioning on W alone opens Path 1, therefore to block both paths, we need to condition on

$$S \equiv \{W, X_2\}.$$

Confounding

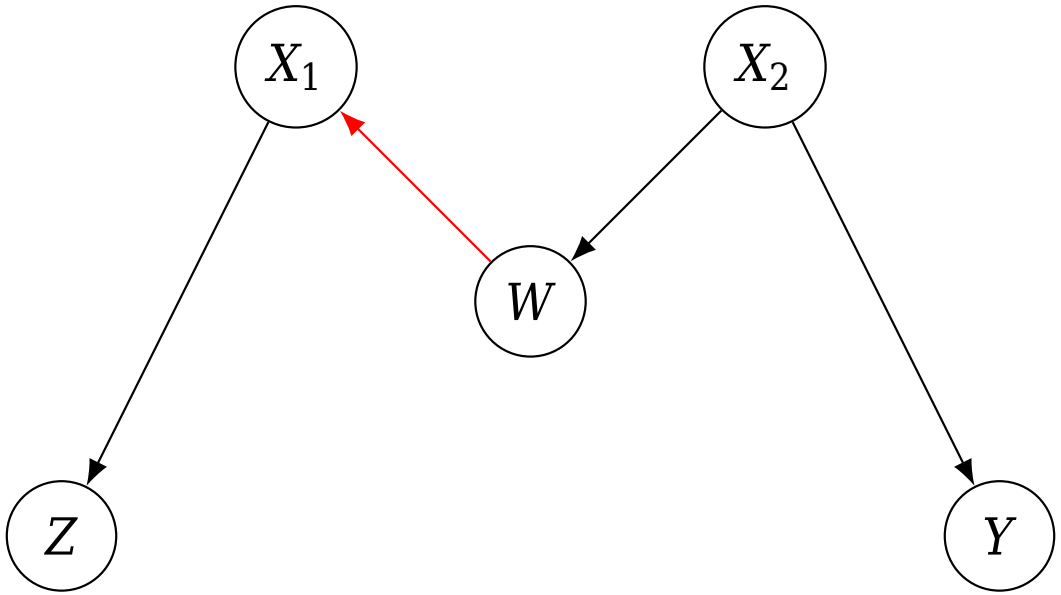


Conditioning on W *opens* the confounding path. Therefore $Z \perp\!\!\!\perp Y$ (as there is no open path between them), but

$$Z \not\perp\!\!\!\perp Y \mid W$$

Further conditioning on either $\{X_1\}$ or $\{X_2\}$ blocks the path.

Confounding



Conditioning on W *blocks* the confounding path. Therefore conditioning on any one of

$$\{X_1\}, \{W\}, \{X_2\}$$

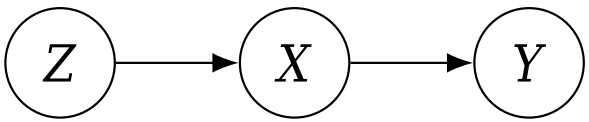
will block the path.

Direct and indirect effects

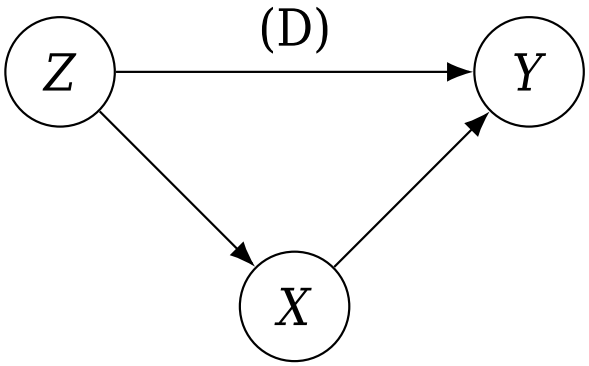
For the effect of Z on Y relative to X:

- *Direct effect:* A direct effect of Z on Y is the effect captured by a *directed* path from Z to Y that does not pass through X.
- *Indirect effect:* An indirect effect of X on Y that is captured by directed paths that pass through X.
    - X is termed an *intermediate* or *mediator* variable.

Direct and indirect effects



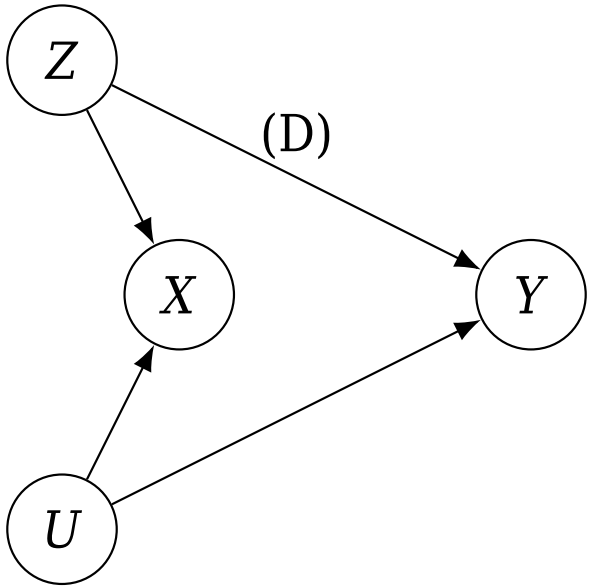
Indirect effect



Direct (D) & Indirect effect

X is a mediator of the indirect effect

Direct and indirect effects



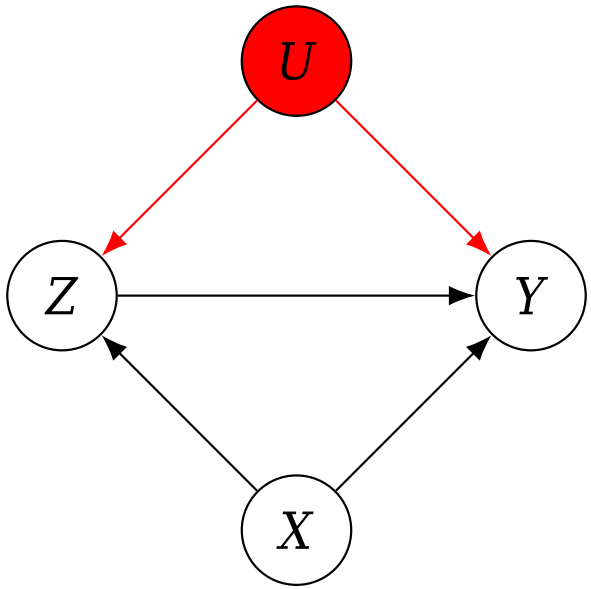
No indirect effect

Direct effect is not confounded

X is a collider, so there is no other open path from Z to Y.

Unmeasured confounding

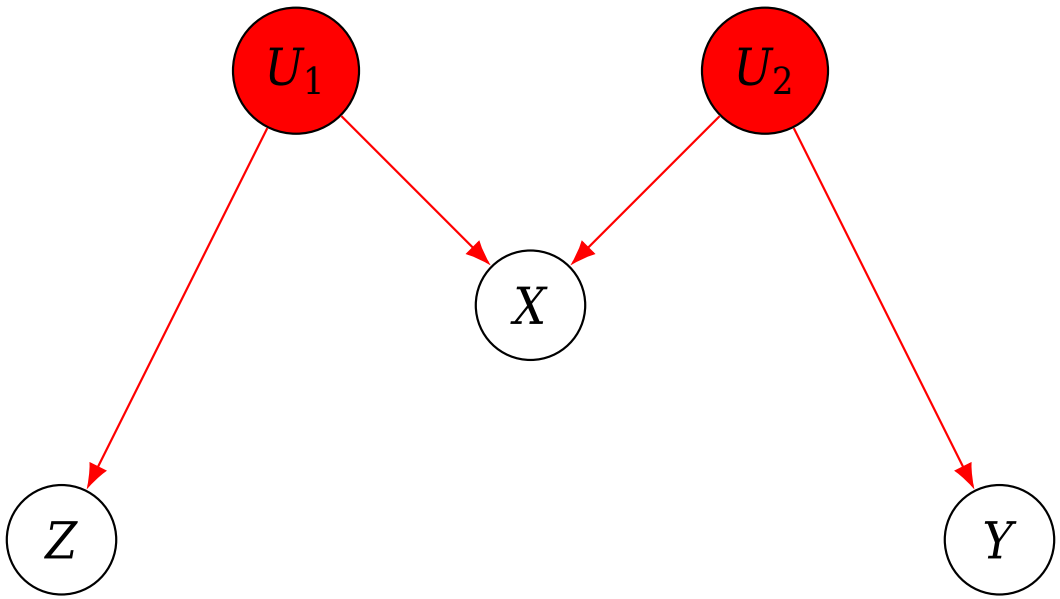
Suppose that in reality there is a further variable U that is a confounder, but is unmeasured in the observed data.



There is a hidden confounding path $Z \to U \to Y$. Conditioning on U is not possible, as we are unaware of its existence.

Unmeasured confounding

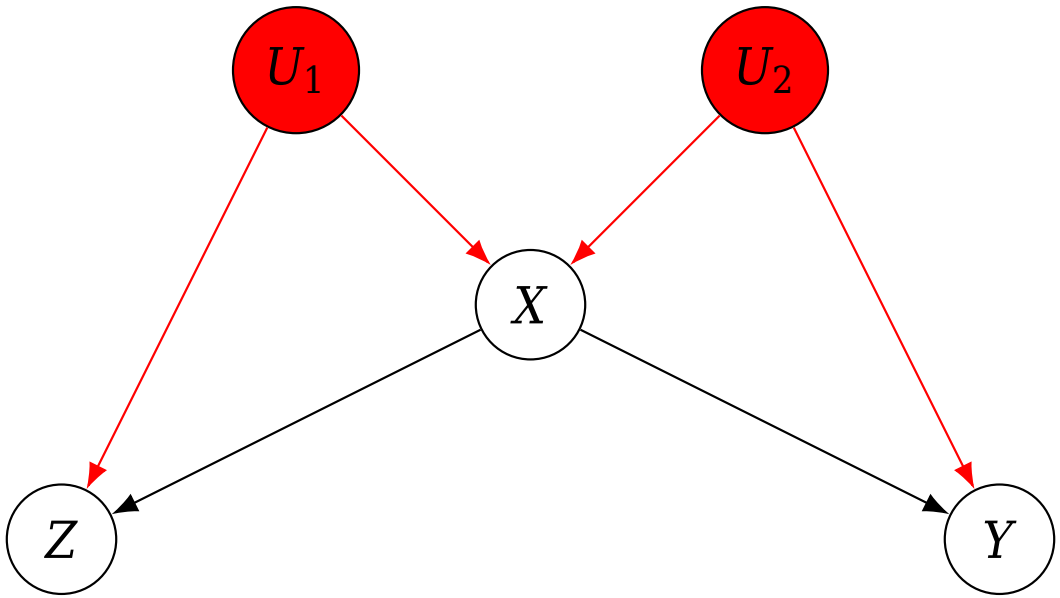
With two unmeasured confounders:



We have that X, Y and Z are *independent*; the (true but hidden) path between Z and Y is blocked at collider X.

Unmeasured confounding

Suppose we condition on X:

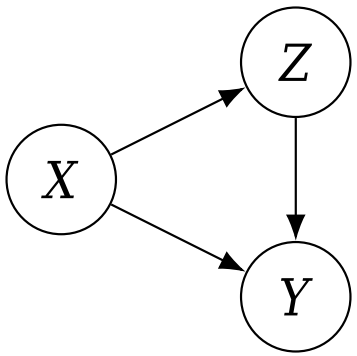


In the modelled DAG, $Y \perp\!\!\!\perp Z \mid X$; however, conditioning on X *opens* the *hidden* path through U_1 and U_2, so there is now an open biasing path.

This is sometimes referred to as the *M-bias* phenomenon.

Implications for Statistical Modelling

We have seen that conditioning on variables can close biasing paths, allowing an unbiased assessment of the causal effect of Z on Y.



The open, undirected path

$$Z \to X \to Y$$

can be blocked by conditioning on X.

Implications for Statistical Modelling

If all the variables are jointly Normally distributed, then this conditioning can be achieved by including X as a predictor in a linear regression model of Y on Z.

That is, we can fit the linear model where

$$\mathbb{E}[Y|X = x, Z = z] = \beta_0 + \beta_1 x + \psi z$$

and estimate the direct effect of Z on Y by estimating ψ.

Implications for Statistical Modelling

Note

Blocking confounding paths (e.g. by conditioning) is not quite the end of the story.

Typically we need to utilize *parametric* inference, and there is usually a requirement that certain parametric models are *correctly specified*.

Summary

In a statistical formulation of a causal inference problem

1. We identify *treatment* Z and *outcome* Y
2. We form the *DAG* representing the relationships between Z and Y which contains other measured variables X.
3. The causal effect of Z on Y flows down *open* and *directed* paths from Z to Y;
    - there may be a *direct* effect if there is an arrow from Z into Y;
    - there may also be *indirect* effects if the directed path passes through *mediating* variables.

Summary

4. If there are *undirected* paths from Z to Y that are open, then these paths may induce *bias* in estimation of the effect of Z on Y.
5. In order to obtain unbiased estimation, the open undirected (biasing) paths must be *blocked*; typically this is done by *conditioning* on variables on those paths.
6. A *collider* node blocks a path; however, conditioning on the collider *opens* the path at that node.

Appendix: Propensity Score Regression Example

Example

Example:

In this example we have

- two confounders X_1 and X_2
- binary treatment Z

where the treatment effect model depends on Z only, or on both Z and X_1.

Back

Example

Example:

Suppose that we have the following data generating model:

- Confounders: $(X_1, X_2)^\top \sim Normal_2((1,1)^\top, \Sigma)$ with

$$\Sigma = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.5 \end{bmatrix} \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix} \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}$$

- Treatment: $Z|X_1, X_2 \sim Bernoulli(e(X_1, X_2))$, where

$$e(x_1, x_2) = \frac{\exp\{1 + x_1 - 2x_2\}}{1 + \exp\{1 + x_1 - 2x_2\}}$$

- Outcome: $Y|X, Z \sim Normal(\mu(X, Z), 1)$, where

$$\mu(x, z) = (2 + 3x_1 + x_2 + x_1 x_2) + z$$

Example

Example:

We consider fitting the parametric model

$$m(x, z; \beta, \psi) = (\beta_0 + \beta_1 x_1) + z\psi_0$$

which is mis-specified due to the 'treatment-free' model specification. The true values is $\psi_0 = 1$.

Example

```
#n=1000
#Correct specification
> round(coef(summary(lm(Y~X1+X2+X1:X2+Z))),4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7134     0.6222  4.3608   0.0000
X1            2.2156     0.6869  3.2254   0.0013
X2            0.2882     0.4807  0.5996   0.5489
Z             1.0150     0.0674 15.0572   0.0000
X1:X2         1.7421     0.4721  3.6905   0.0002
```

```
#Incorrect specification
> round(coef(summary(lm(Y~X1+Z))),4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  -4.2613     0.4034 -10.5631        0
X1           11.4990     0.3888  29.5760        0
Z             0.6366     0.0762   8.3523        0
```

Example

Example:

In the correctly specified model, we have

$$\widehat{\psi}_0 \ : \ 1.0150 \ (0.0674)$$

however in the incorrectly specified model we have

$$\widehat{\psi}_0 \ : \ 0.6366 \ (0.0762)$$

This effect persists at even larger sample sizes.

Example

Example:

Now consider fitting the parametric model

$$m(x, z; \beta, \psi, \phi) = (\beta_0 + \beta_1 x_1) + z\psi_0 + e(x_1, x_2)\phi_0$$

which considers the additional propensity score term.

Initially, we will set

$$e(x_1, x_2) = \frac{\exp\{1 + x_1 - 2x_2\}}{1 + \exp\{1 + x_1 - 2x_2\}}$$

that is, using the true value.

Example

```
#Propensity score regression
> round(coef(summary(lm(Y~X1+Z+eX))),4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   4.1718     0.5609   7.4377        0
X1            5.1662     0.4701  10.9907        0
Z             1.0172     0.0682  14.9069        0
eX           -4.6374     0.2430 -19.0815        0
```

Example

Example:

We now have

$$\widehat{\psi}_0 \ : \ 1.0172 \ (0.0682)$$

and so correct estimation of ψ_0 has been recovered.

Example

Example:

Now suppose

$$\mu(x, z) = (2 + 3x_1 + x_2 + x_1x_2) + z(1 + x_1 + x_2)$$

and using the propensity score regression model

$$m(x, z; \beta, \psi, \phi) = (\beta_0 + \beta_1 x_1) + z(\psi_0 + \psi_1 x_1 + \psi_2 x_2) + e(x_1, x_2)\phi_0$$

Example

```
#n=1000
#Correct specification
> round(coef(summary(lm(Y~X1+X2+X1:X2+Z+Z:X1+Z:X2))),4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5674     0.9486  3.7609   0.0002
X1            1.2812     1.0109  1.2675   0.2053
X2            0.1155     0.6004  0.1923   0.8475
Z            -0.2023     0.9313 -0.2173   0.8281
X1:X2         1.9903     0.5672  3.5090   0.0005
X1:Z          2.3744     1.0558  2.2488   0.0247
X2:Z          0.8420     0.2091  4.0272   0.0001
```

Example

```
#Incorrect specification
> round(coef(summary(lm(Y~X1+Z+Z:X1+Z:X2))),4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.4874     0.5541 -8.0981        0
X1           11.7187     0.5363 21.8503        0
Z             6.4906     0.8778  7.3941        0
X1:Z         -6.5766     0.9644 -6.8196        0
Z:X2          2.8785     0.1642 17.5344        0
```

Example

```
#Propensity score regression
> round(coef(summary(lm(Y~X1+Z+Z:X1+Z:X2+eX))),4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   3.9848     0.7752   5.1403   0.0000
X1            5.4002     0.6565   8.2252   0.0000
Z             1.4774     0.8716   1.6951   0.0904
eX           -4.7679     0.3313 -14.3913   0.0000
X1:Z          0.6533     1.0113   0.6460   0.5184
Z:X2          0.8889     0.2036   4.3664   0.0000
```

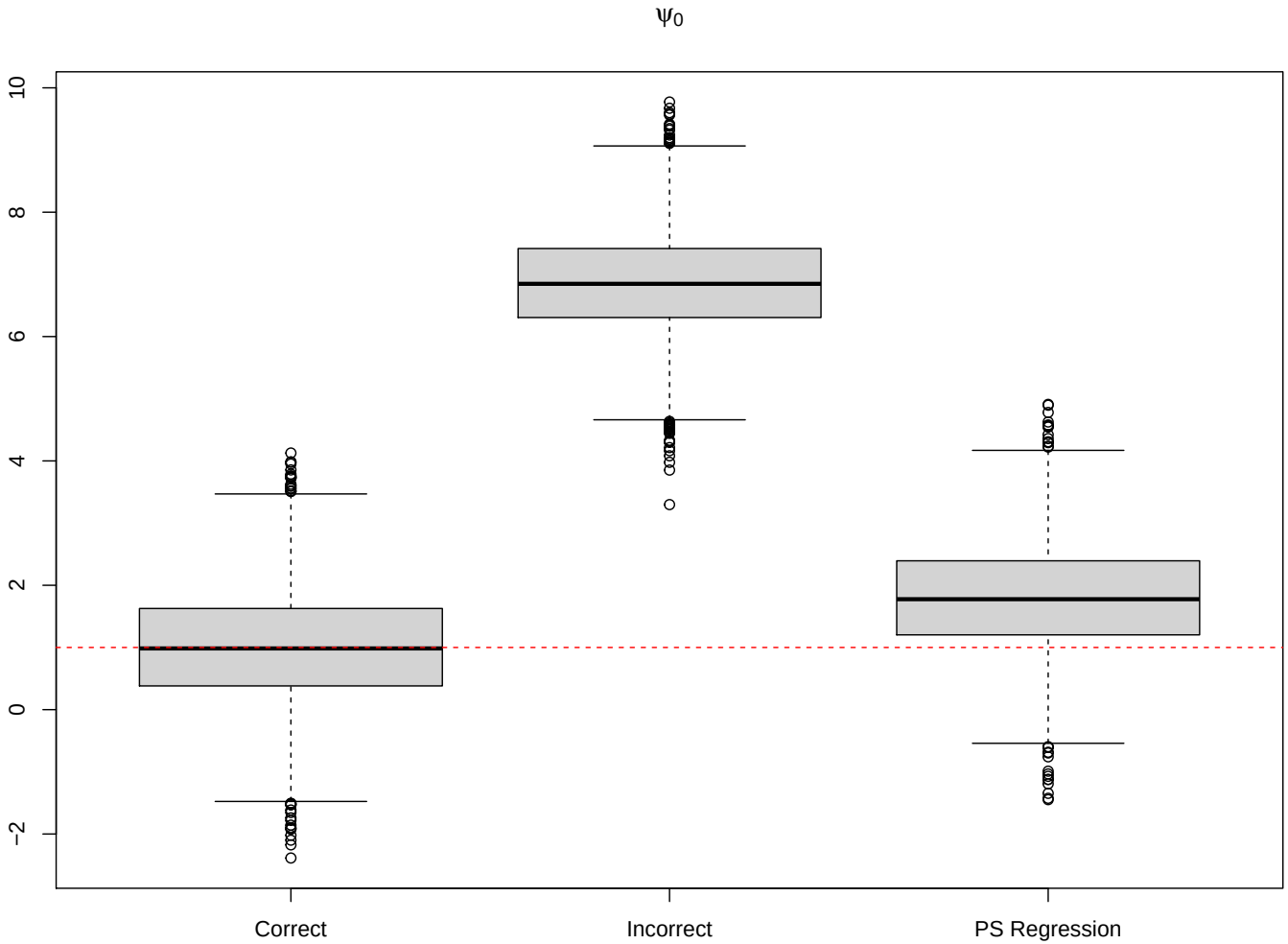
Example

Example:

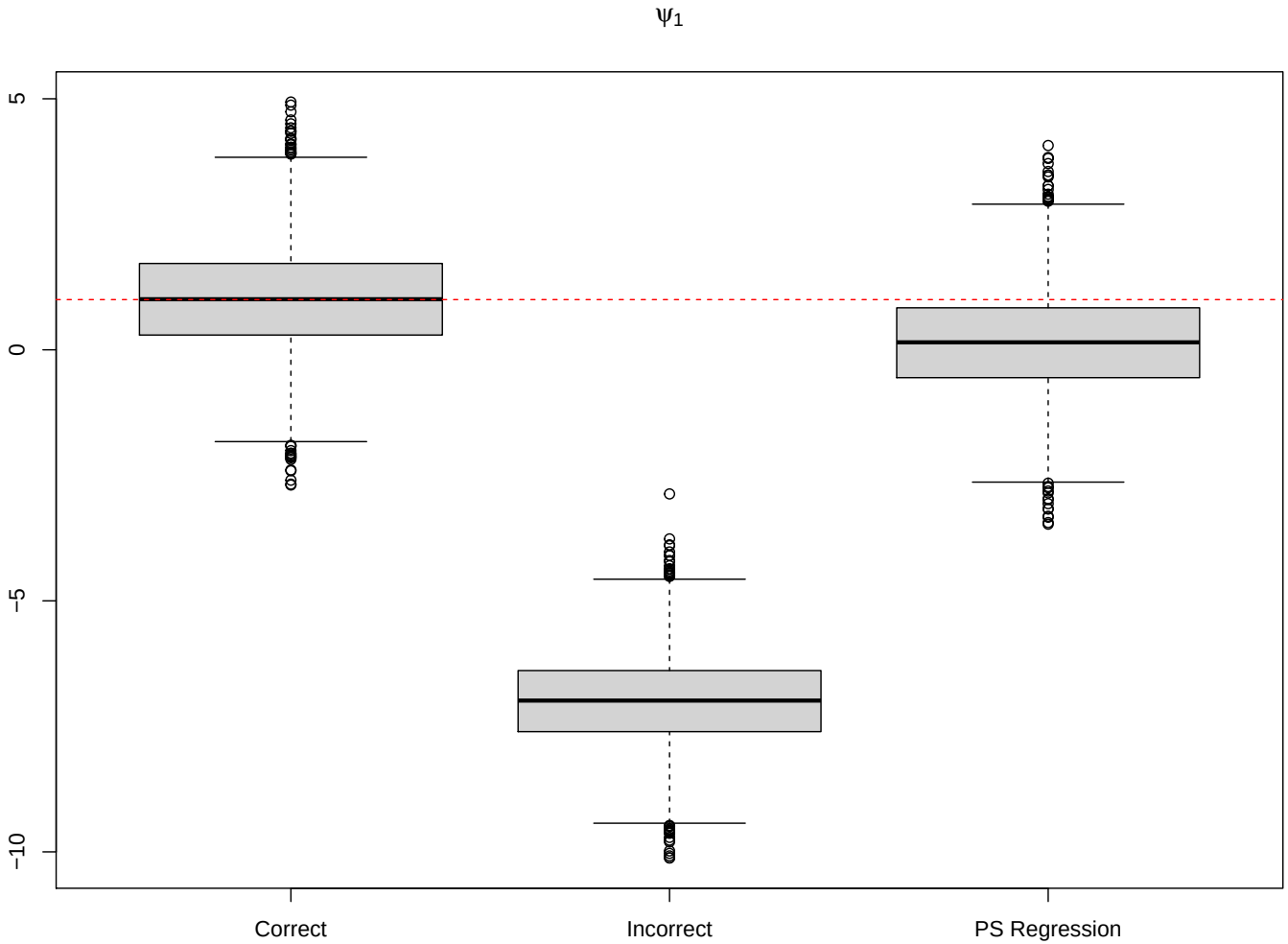
Hard to conclude anything due to the inherent variability, but it seems that including the propensity score does improve the estimation of (ψ_0, ψ_1, ψ_2).

Need to do a larger simulation study: we perform 5000 replications, and inspect the boxplots of the estimates for the three parameters.

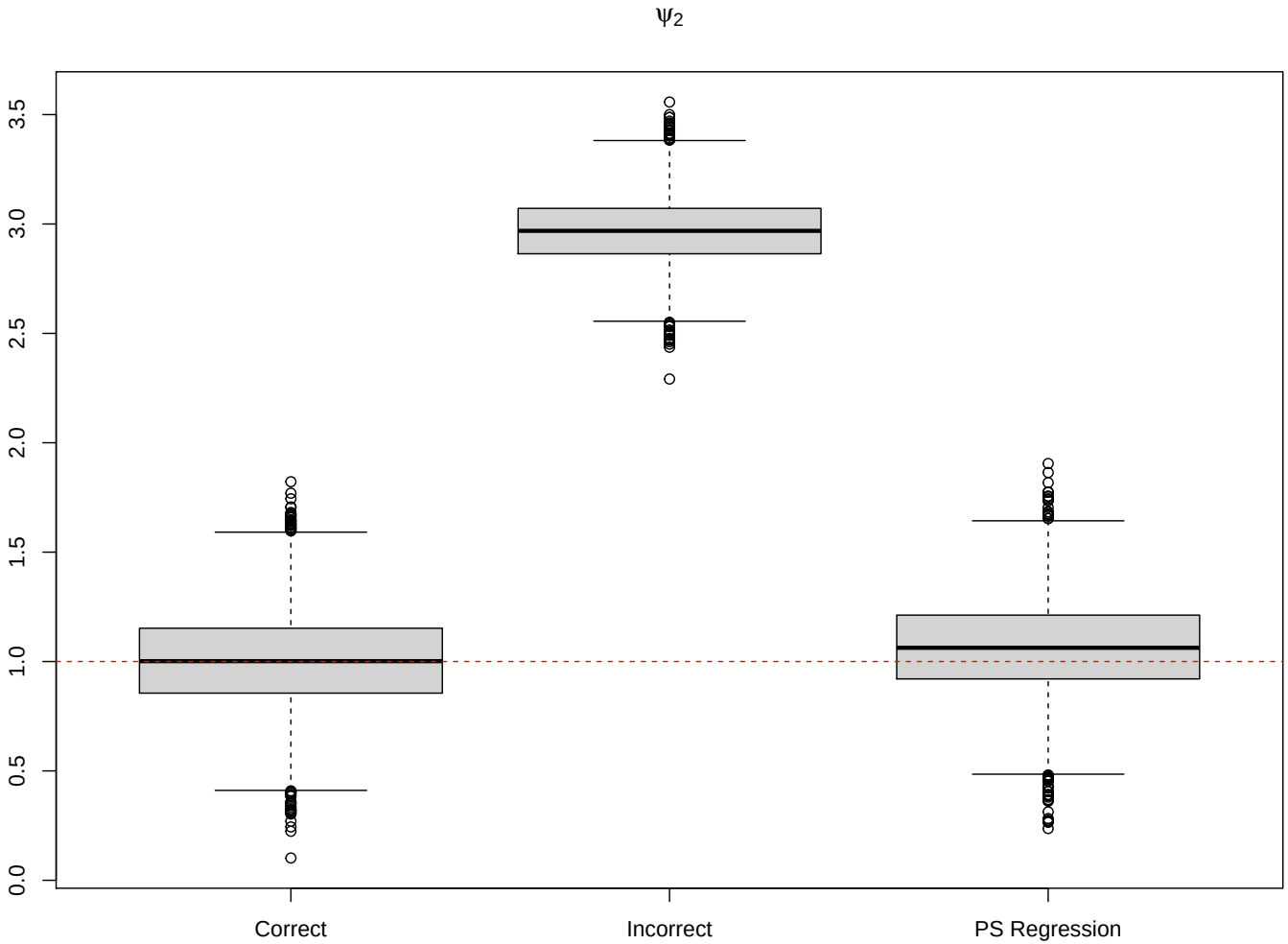
Example



Example



Example



Example

Example:

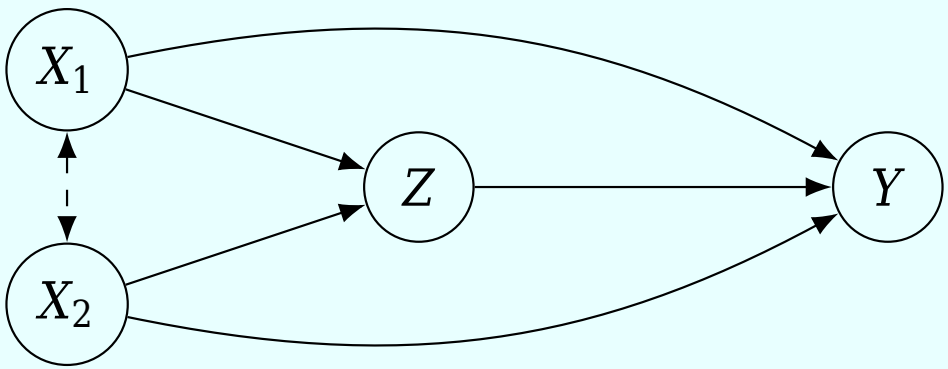
This confirms that including the propensity score *does* improve the estimation of (ψ_0, ψ_1, ψ_2), even if the treatment-free model component is incorrectly specified.

However, it seems that there is still a small amount of bias.

Example

Example:

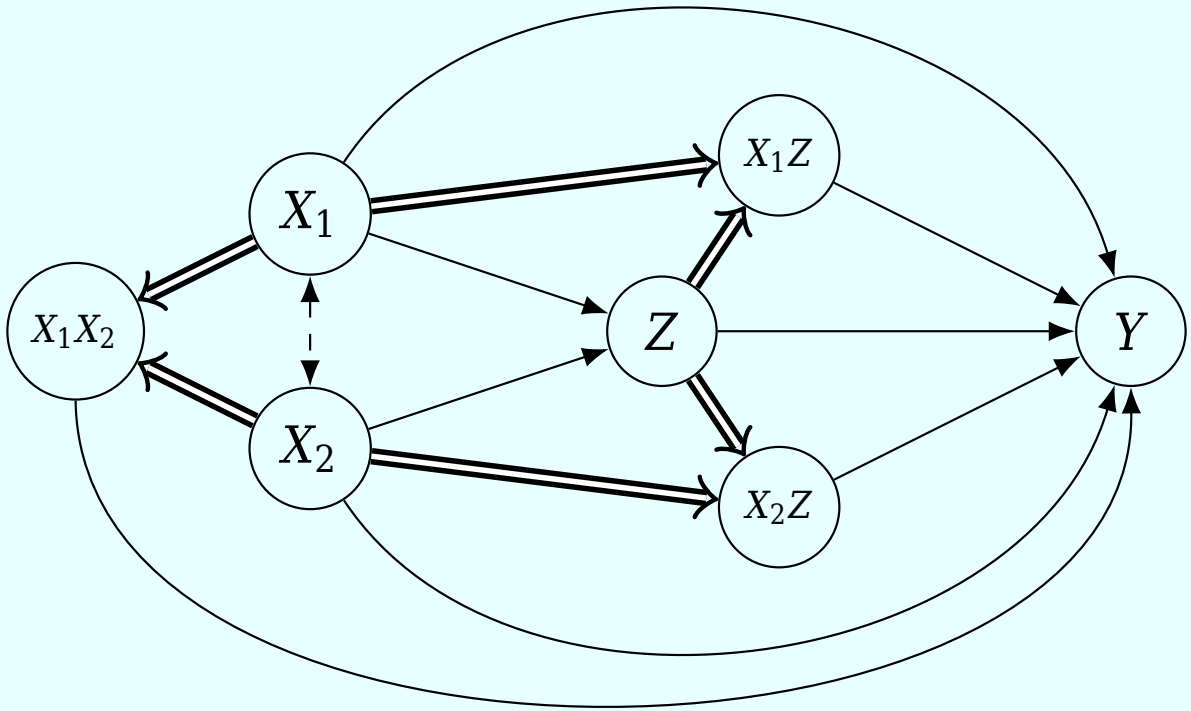
Here is a version of the DAG for the data generating model



Example

Example:

However, a more accurate DAG includes the *interactions*.



Example

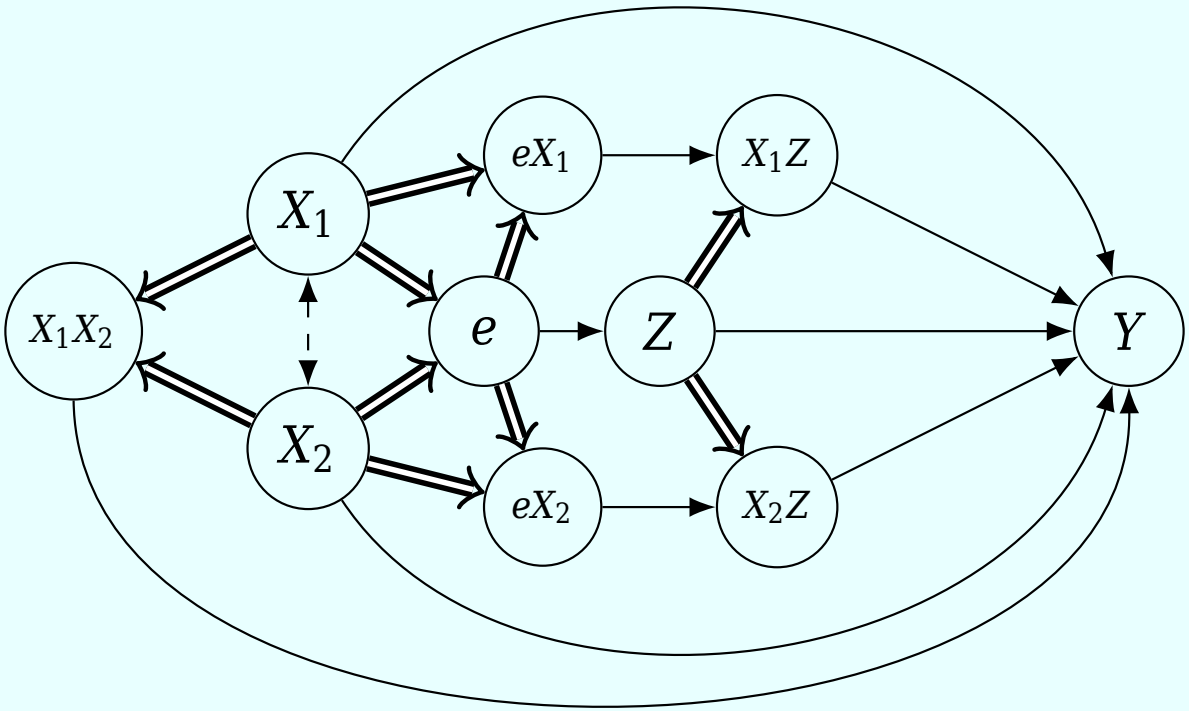
Example:

We need to block the open paths via the interactions. This can be achieved by using the model

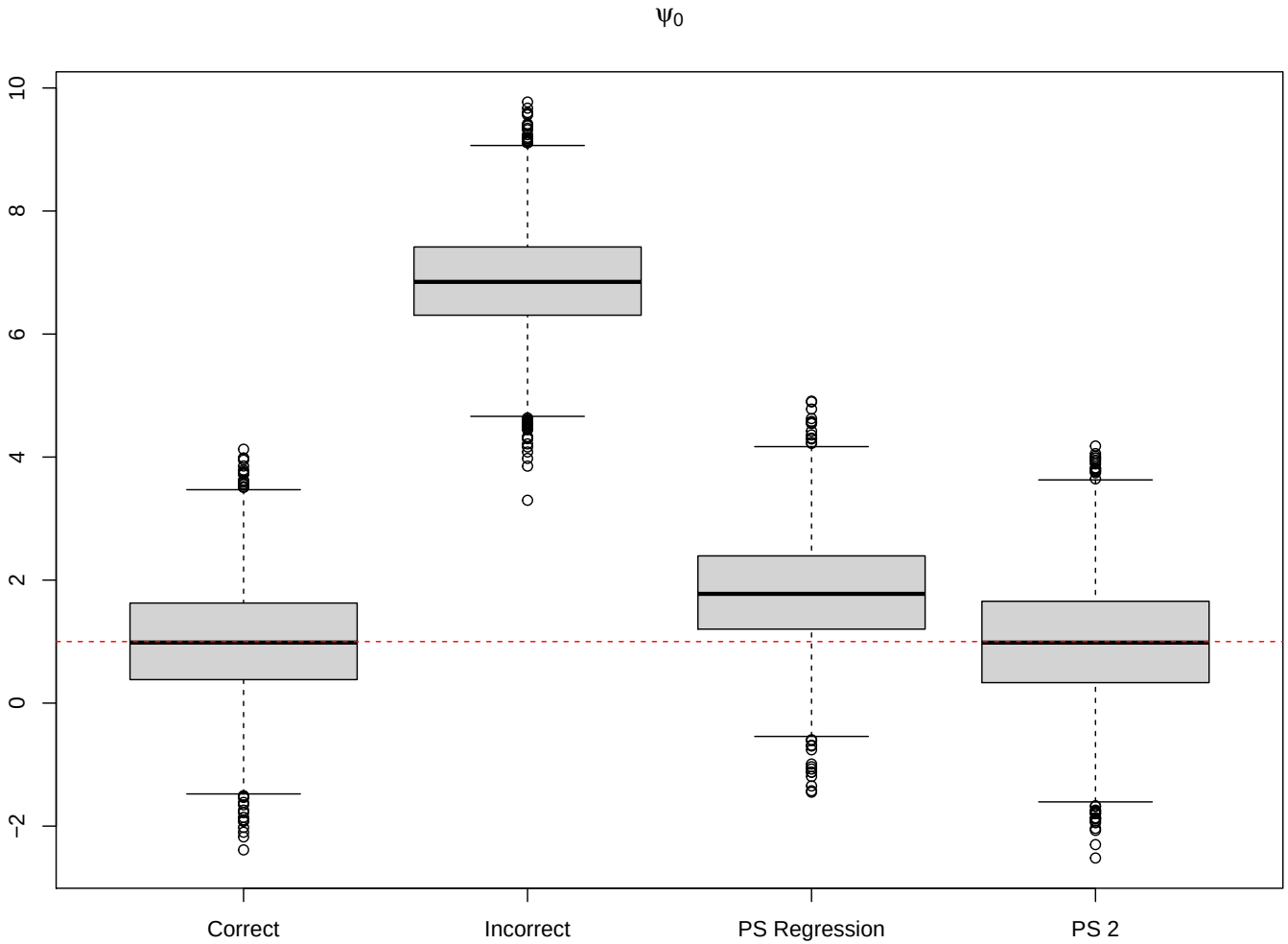
$$m(x, z; \beta, \psi, \phi) = (\beta_0 + \beta_1 x_1) + z(\psi_0 + \psi_1 x_1 + \psi_2 x_2) + e(x_1, x_2)(\phi_0 + \phi_1 x_1 + \phi_2 x_2)$$

Conditioning on $e(X), e(X)X_1$ and $e(X)X_2$ blocks the paths.

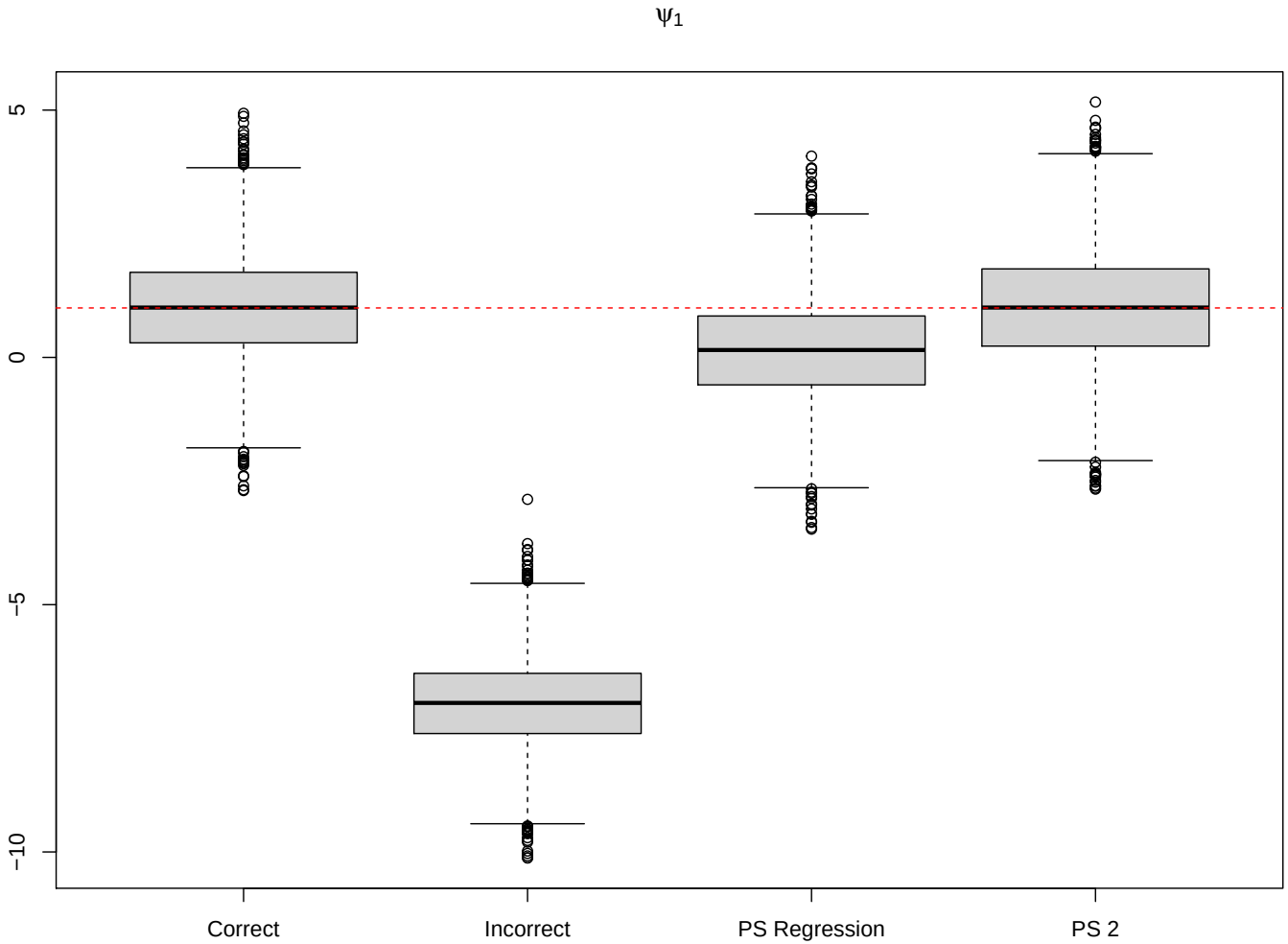
Example:



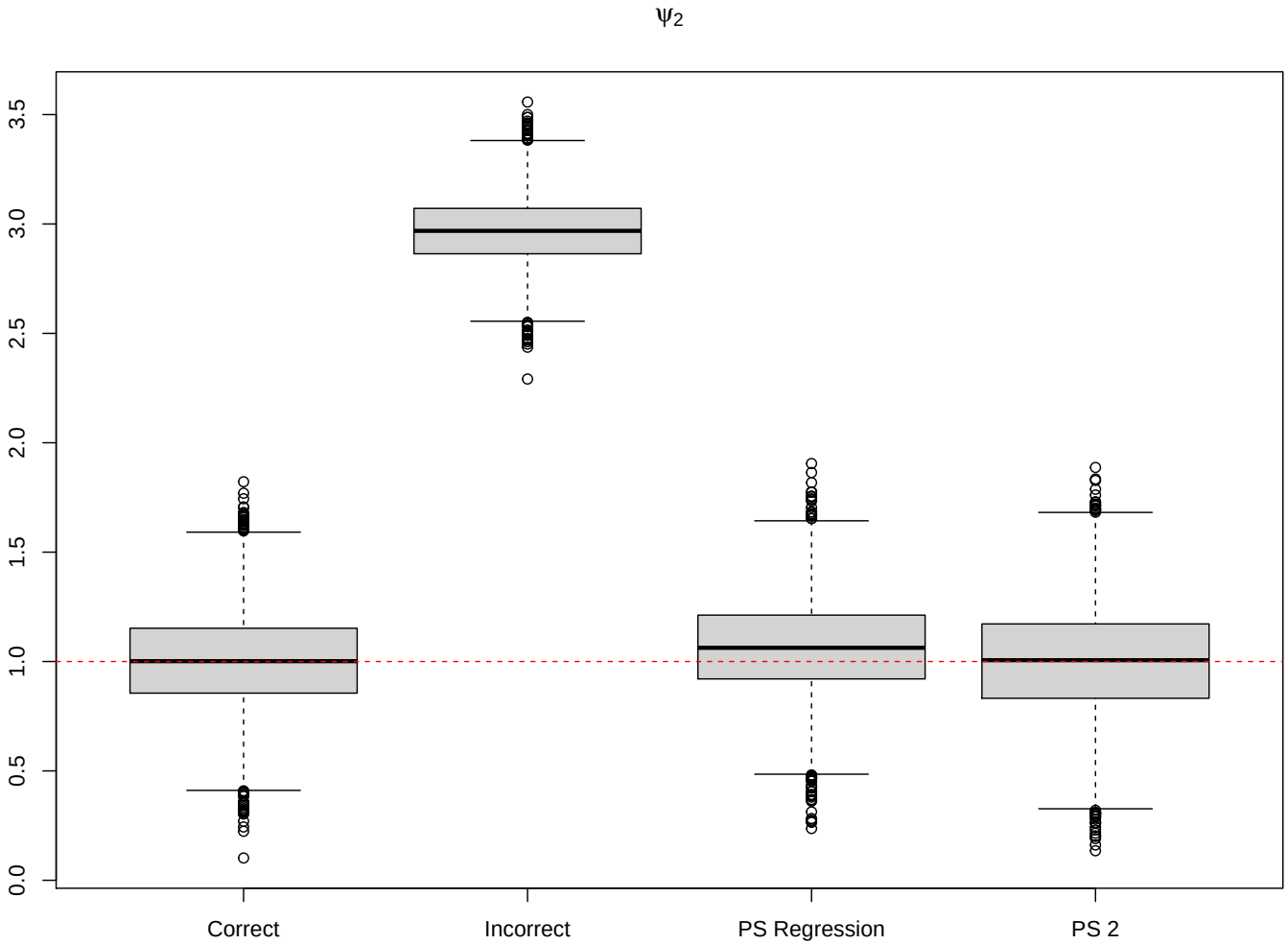
Example



Example



Example


ψ_2
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Correct
Incorrect
PS Regression
PS 2

Example

Example:

The augmented propensity score regression model (PS 2) improves the performance.

Note, however, that the variances of the estimators from propensity score regression model are slightly *larger* than those arising from the correctly specified model.

- 10% to 20% larger in this simulation.

Example

Example:

In this analysis, we may estimate the ATE by taking the average difference of the two fitted values under the proposed model, that is

$$\widehat{\delta} = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\psi}_0 + \widehat{\psi}_1 x_{i1} + \widehat{\psi}_2 x_{i2}).$$

Example

Example:

We need to take care in estimating the APO. In the data generating model, with

$$\mu(x, z) = (2 + 3x_1 + x_2 + x_1 x_2) + z(1 + x_1 + x_2)$$

we have that

$$\mu(\mathsf{z}) = 2 + 3\mathbb{E}[X_1] + \mathbb{E}[X_2] + \mathbb{E}[X_1 X_2] + \mathsf{z}(1 + \mathbb{E}[X_1] + \mathbb{E}[X_2]).$$

This cannot in general be estimated correctly using

$$\widehat{\mu}(\mathsf{z}) = \frac{1}{n} \sum_{i=1}^{n} m(x_i, \mathsf{z}; \widehat{\beta}, \widehat{\psi}, \widehat{\phi}).$$

Example

Note

This type of adjustment works for a *linear* outcome model; however, for other types of model such as

- log-linear
- logistic

more care needs to be taken.

Appendix: Marginal Structural Model Example

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

See Saarela et al. 2015, *Biometrics*.

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.

Back

MSM 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.

MSM 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).

MSM 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.

MSM 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.

MSM 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.

MSM Real Data Example

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

We considered the structural model

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

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

MSM 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*.