MATH 598/MATH 782

Introduction to Causal Inference Methods

Dr David A. Stephens

Department of Mathematics & Statistics
Room 1225, Burnside Hall

david.stephens@mcgill.ca

Part 1
Introduction

Objective

The objective of *causal inference* is to quantify the effect of an *intervention*:

- in a multi-variable system, suppose we are able to manipulate (i.e. alter the value of) one of the variables separately from all other variables;
- we wish to report the impact of that manipulation on one or more of the other variables.

In many scientific enterprises, this is a primary objective.

Some basic probability calculus

Consider three random variables: X, Y and Z. Ultimately we will collect data

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

which are observed values of the variables.

A *probabilistic model* for the data comprises a *joint density*

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

or for discrete variables a *joint mass function*, which represents how the data are generated.

Some basic probability calculus

This joint model automatically specifies

- the *marginal* distributions, $f_X(x)$, $f_Y(y)$ and $f_Z(z)$;
- the *conditional* distributions

$$f_{X|Y}(x|y) \qquad f_{X|Z}(x|z) \qquad f_{Y|X}(y|x) \qquad \cdots$$

and

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

etc.

Some basic probability calculus

We have the *chain rule factorization*

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

but also

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

and so on, for any ordering of the variables.

Some basic probability calculus

Marginalization:

$$f_Y(y) = \iint f_{X,Y,Z}(x,y,z)\ dx\ dz$$

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

$$\equiv \iint f_{Y|X,Z}(y|x,z) f_{X|Z}(x|z) f_Z(z)\ dx\ dz$$

Some basic probability calculus

Conditioning: provided $f_{X,Z}(x,z) > 0$,

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

$$\equiv \frac{f_{X,Y,Z}(x,y,z)}{\int f_{X,Y,Z}(x,t,z)\ dt}$$

Some basic probability calculus

Note

- for *discrete* variables, integrals replaced by sums,

$$f_{Z|X}(z|x) = \frac{f_{X,Z}(x,z)}{f_X(x)} \equiv \frac{\Pr[X = x, Z = z]}{\Pr[X = x]}$$

$$= \frac{\Pr[X = x, Z = z]}{\sum_t \Pr[X = x, Z = t]}.$$

- Can have *mixed* cases: Z discrete, X continuous.

Some basic probability calculus

Expectations: we can compute the summary

$$\begin{aligned}
\mathbb{E}_Y[Y] &= \int y \; f_Y(y) \; dy \\
&\equiv \int y \left\{ \iint f_{X,Y,Z}(x,y,z) \; dx \; dz \right\} \; dy \\
&\equiv \int y \left\{ \iint f_{Y|X,Z}(y|x,z) f_{Z|X}(z|x) f_X(x) \; dx \; dz \right\} \; dy \\
&\equiv \iint \left\{ \int y \; f_{Y|X,Z}(y|x,z) \; dy \right\} f_{Z|X}(z|x) f_X(x) \; dx \; dz
\end{aligned}$$

Some basic probability calculus

We may denote

$$\int y \; f_{Y|X,Z}(y|x,z) \; dy \equiv \mathbb{E}_{Y|X,Z}[Y|X=x,Z=z]$$

that is, as a *conditional expectation*. Thus

$$\mathbb{E}_Y[Y] = \iint \mathbb{E}_{Y|X,Z}[Y|X=x,Z=z] f_{Z|X}(z|x) f_X(x) \; dx \; dz$$

which we may also re-write

$$\mathbb{E}_Y[Y] = \mathbb{E}_{X,Z}\left[\mathbb{E}_{Y|X,Z}[Y|X,Z]\right]$$

which is known as *iterated expectation*.

Some basic probability calculus

Note

The quantity

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

is a function of the two values (x, z) and therefore is *non-random*, whereas

$$\mathbb{E}_{Y|X,Z}[Y|X, Z]$$

is a function of (X, Z) and is therefore a *random variable*.

Some basic probability calculus

Consider the conditional expectation $\mathbb{E}_{Y|Z}[Y|Z = z]$ for some fixed value z. We have

$$\begin{aligned}
\mathbb{E}_{Y|Z}[Y|Z = z] &= \int y \, f_{Y|Z}(y|z) \, dy \\
&= \iint y \, f_{Y|X,Z}(y|x,z) f_{X|Z}(x|z) \, dy \, dx \\
&= \iiint \mathbb{1}_{\{z\}}(v) y \, f_{Y|X,Z}(y|x,v) f_{X|Z}(x|v) \, dy \, dx \, dv
\end{aligned}$$

where

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

is the *indicator function*.

Some basic probability calculus

That is,

$$\mathbb{E}_{Y|Z}[Y|Z = z] = \iiint y \; f_{Y|X,Z}(y|x,v) f_{X|Z}(x|v) f_V(v) \; dy \; dx \; dv$$

$$\equiv \mathbb{E}_{X,V}[\mathbb{E}_{Y|X,V}[Y|X,V]]$$

where V is a *degenerate* random variable with

$$f_V(v) = \Pr[V = v] = \mathbb{1}_{\{z\}}(v) = \begin{cases} 1 & v = z \\ 0 & v \neq z \end{cases} .$$

Some basic probability calculus

Independence: Two random variables X, Z are *independent*

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

if and only if

$$f_{X,Z}(x, z) = f_X(x) f_Z(z) \qquad \forall (x, z) \in \mathbb{R}^2$$

that is, *for all* $(x, z) \in \mathbb{R}^2$, or equivalently

$$f_{X|Z}(x|z) = f_X(x) \qquad \forall (x, z) \text{ s.t. } f_Z(z) > 0$$

or

$$f_{Z|X}(z|x) = f_Z(z) \qquad \forall (x, z) \text{ s.t. } f_X(x) > 0.$$

Some basic probability calculus

Note

- For three variables, we require for independence

$$f_{X,Y,Z}(x,y,z) = f_X(x)f_Y(y)f_Z(z) \qquad \forall (x,y,z) \in \mathbb{R}^3$$

- We can consider *conditional independence*: say

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

if and only if

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

for all $(x,z,y) \in \mathbb{R}^3$ where the conditional densities are well-defined.

Some basic probability calculus

Note

Suppose that X and V are two random variables, but suppose that V is *degenerate* at some fixed value $v_0 \in \mathbb{R}$, that is,

$$\Pr[V = v_0] = 1.$$

Consider the joint distribution of X and V: we have that for arbitrary x

$$f_{X,V}(x, v) = \begin{cases} g(x, v_0) & x \in \mathbb{R}, v = v_0 \\ 0 & x \in \mathbb{R}, v \neq v_0 \end{cases}.$$

for some function $g(x, v)$.

Some basic probability calculus

Note

Therefore, marginally

$$f_X(x) = g(x, v_0)$$

which must be a density in x. Hence for all $(x, v) \in \mathbb{R}^2$

$$f_{X,V}(x, v) = f_X(x) f_V(v)$$

and hence *X and V are independent*

$$X \perp\!\!\!\perp V.$$

Some basic probability calculus

In the previous calculation, suppose X and Z are independent:

$$\begin{aligned}
\mathbb{E}_{Y|Z}[Y|Z = z] &= \iint y \; f_{Y|X,Z}(y|x,z) f_{X|Z}(x|z) \; dy \; dx \\
&\equiv \iint y \; f_{Y|X,Z}(y|x,z) f_X(x) \; dy \; dx \qquad \text{as } X \perp\!\!\!\perp Z \\
&\equiv \mathbb{E}_X[\mathbb{E}_{Y|X,Z}[Y|X,z]]
\end{aligned}$$

Some basic probability calculus

That is, we can compute $\mathbb{E}_{Y|Z}[Y|Z = z]$ by

- fixing $Z = z$ independently of X,
- computing for each fixed x

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

  say,

- averaging the result over the distribution $f_X(x)$

$$\mathbb{E}_X[\mu(X, z)]$$

Some basic probability calculus

Regression: we might propose

$$\mathbb{E}_{Y|X,Z}[Y|X = \mathrm{x}, Z = \mathsf{z}] = \beta_0 + \beta_1 \mathrm{x} + \psi_0 \mathsf{z}$$

or

$$\mathbb{E}_{Y|X,Z}[Y|X = \mathrm{x}, Z = \mathsf{z}] = \beta_0 + \beta_1 \mathrm{x} + \psi_0 \mathsf{z} + \psi_1 \mathrm{x}\mathsf{z}$$

etc. for some *parameters* β, ψ to specify the *mean model*

$$\mu(\mathrm{x}, \mathsf{z}; \beta, \psi).$$

Some basic probability calculus

Binary variables: suppose X, Y, Z are binary and consider

$$\Pr[X = x, Y = y, Z = z] \qquad (x, y, z) \in \{0, 1\}^3.$$

Suppose $\Pr[Z = 0]\Pr[Z = 1] \neq 0$, that is

$$\Pr[Z = 0] > 0 \qquad \text{and} \qquad \Pr[Z = 1] > 0.$$

Some basic probability calculus

We have for $(y, z) \in \{0, 1\}^2$

$$\Pr[Y = y|Z = z] = \sum_{x=0}^{1} \Pr[Y = y|X = x, Z = z]\Pr[X = x|Z = z]$$

$$= \Pr[Y = y|X = 0, Z = z]\Pr[X = 0|Z = z]$$

$$+ \Pr[Y = y|X = 1, Z = z]\Pr[X = 1|Z = z]$$

$$\neq \Pr[Y = y|X = 0, Z = z]\Pr[X = 0]$$

$$+ \Pr[Y = y|X = 1, Z = z]\Pr[X = 1]$$

in general.

Some basic probability calculus

Simpson's Paradox:

Consider

$$X = \begin{cases} 0 & \text{Group 0} \\ 1 & \text{Group 1} \end{cases}$$

$$Z = \begin{cases} 0 & \text{Treatment A} \\ 1 & \text{Treatment B} \end{cases}$$

$$Y = \begin{cases} 0 & \text{Not cured} \\ 1 & \text{Cured} \end{cases}$$

Some basic probability calculus

Data:

$X = 0$

Z \ Y	0	1
0	36	234
1	6	81

$X = 1$

Z \ Y	0	1
0	25	55
1	71	192

Collapsing over X:

Z \ Y	0	1
0	61	289
1	77	273

Some basic probability calculus

Estimated cure rates for the two treatment groups:

- In Group 0 ($X = 0$):

$$Z = 0: \ \frac{234}{270} \simeq 0.87 \qquad Z = 1: \ \frac{81}{87} \simeq 0.93$$

- In Group 1 ($X = 1$):

$$Z = 0: \ \frac{55}{80} \simeq 0.69 \qquad Z = 1: \ \frac{192}{263} \simeq 0.73$$

In the pooled data:

$$Z = 0: \ \frac{289}{350} \simeq 0.83 \qquad Z = 1: \ \frac{273}{350} \simeq 0.78$$

Some basic probability calculus

Therefore in each of the two Groups *separately*,

Treatment B beats Treatment A

but in the *pooled* data, it seems

Treatment A beats Treatment B.

Some basic probability calculus

Not surprising:

$$\Pr[Y=1|Z=1] = \Pr[Y=1|X=0,Z=1]\Pr[X=0|Z=1] + \Pr[Y=1|X=1,Z=1]\Pr[X=1|Z=1]$$

and we have from the data

$$\widehat{\Pr}[Y=1|X=0,Z=1] = 0.93 \quad \text{(a)}$$
$$\widehat{\Pr}[Y=1|X=0,Z=0] = 0.87 \quad \text{(b)}$$
$$\widehat{\Pr}[Y=1|X=1,Z=1] = 0.73 \quad \text{(c)}$$
$$\widehat{\Pr}[Y=1|X=1,Z=0] = 0.69 \quad \text{(d)}$$

Some basic probability calculus

$$\widehat{\Pr}[Y = 1|Z = 1] = 0.78 \qquad \text{i.e. } (1 - w_1)ⓐ + w_1ⓒ$$
$$\widehat{\Pr}[Y = 1|Z = 0] = 0.83 \qquad \text{i.e. } (1 - w_0)ⓑ + w_0ⓓ$$

where

$$w_1 = \widehat{\Pr}[X = 1|Z = 1] = \frac{263}{263 + 87} \simeq 0.75$$

and

$$w_0 = \widehat{\Pr}[X = 1|Z = 0] = \frac{80}{80 + 270} \simeq 0.22.$$

Some basic probability calculus

The weights

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

are substantially different, representing (in the joint distribution rather than the data) *dependence* between X and Z.

- There is an *imbalance* between the two treatments when considering the representation of the two Groups of individuals;
- as the probability of cure is different for the two groups, this imbalance affects the conclusions from the pooled data.

Some basic probability calculus

Note

It is important to consider whether we wish to report

- a *conditional on x* comparison

$$\Pr[Y = 1 | X = x, Z = 1] \quad \text{vs} \quad \Pr[Y = 1 | X = x, Z = 0]$$

- a *marginal* comparison

$$\Pr[Y = 1 | Z = 1] \quad \text{vs} \quad \Pr[Y = 1 | Z = 0].$$

Some basic probability calculus

This kind of result is not limited to discrete variables: suppose

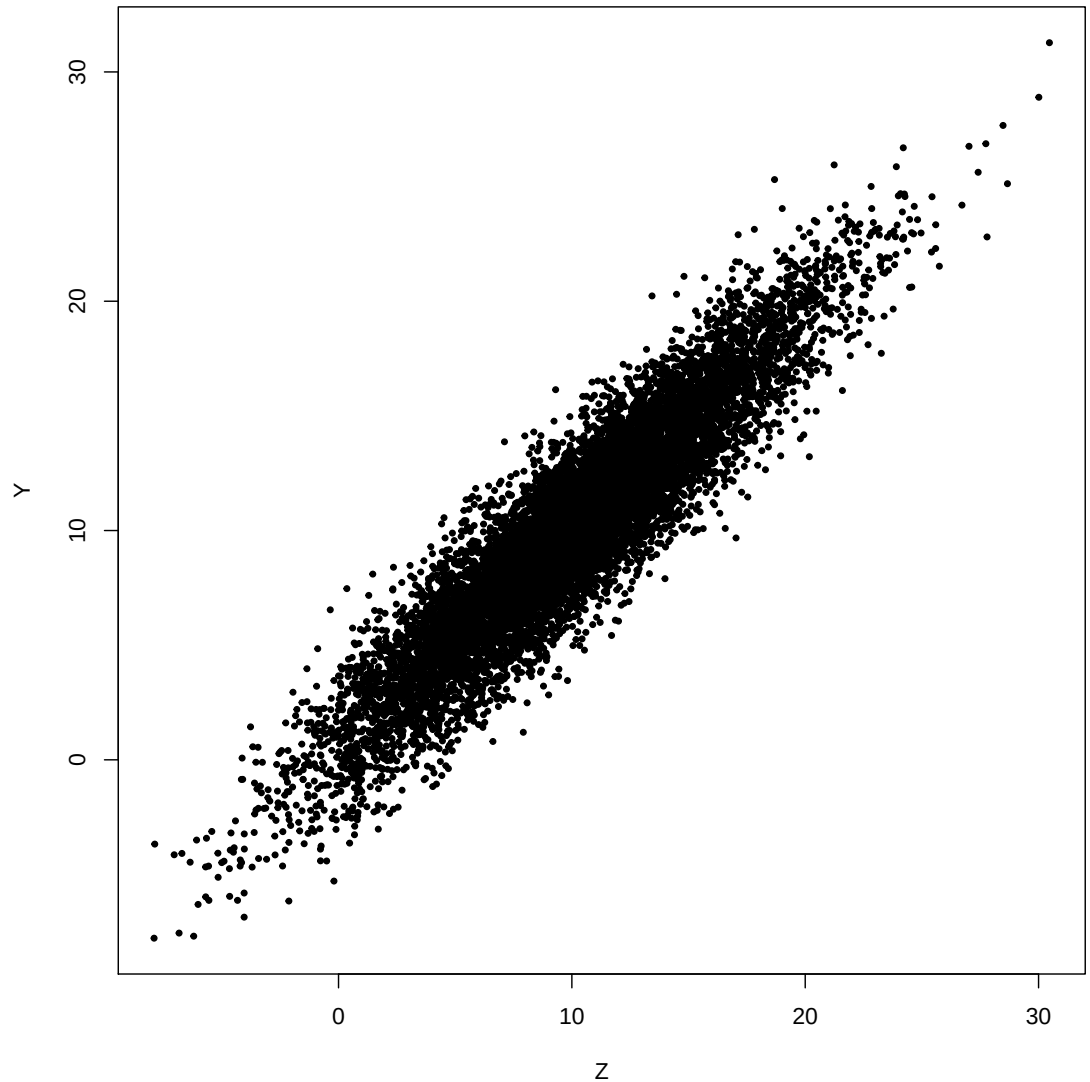
$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim Normal_3(\mu, \Sigma)$$

constructed as follows:

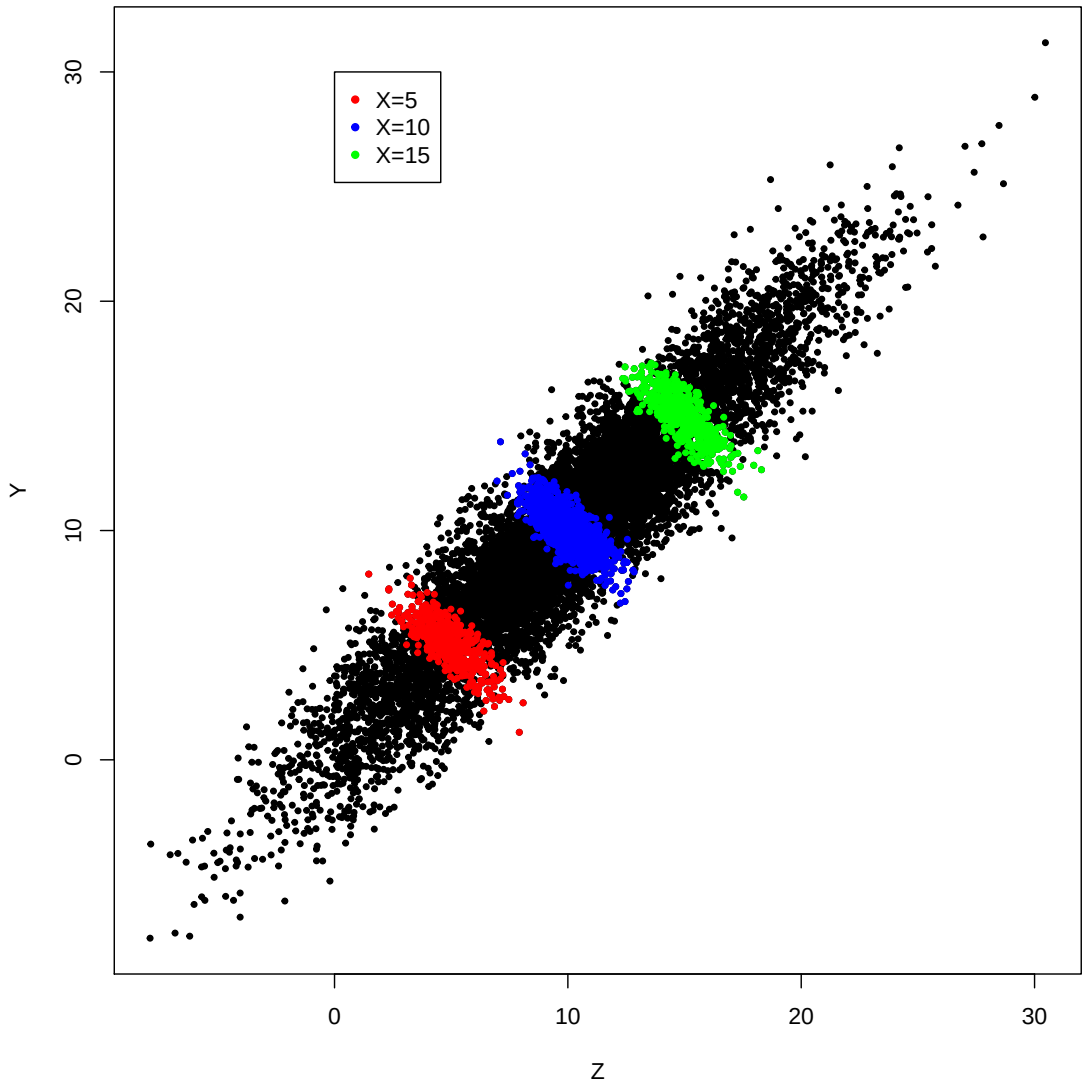
- *Marginal* for X: $X \sim Normal(\mu_X, \sigma_X^2)$
- *Conditional* for (Y, Z) given $X = x$:

$$(Y, Z)|X = x \sim Normal_2\left(\begin{pmatrix} x \\ x \end{pmatrix}, \begin{pmatrix} 1.0 & -0.9 \\ -0.9 & 1.0 \end{pmatrix}\right)$$

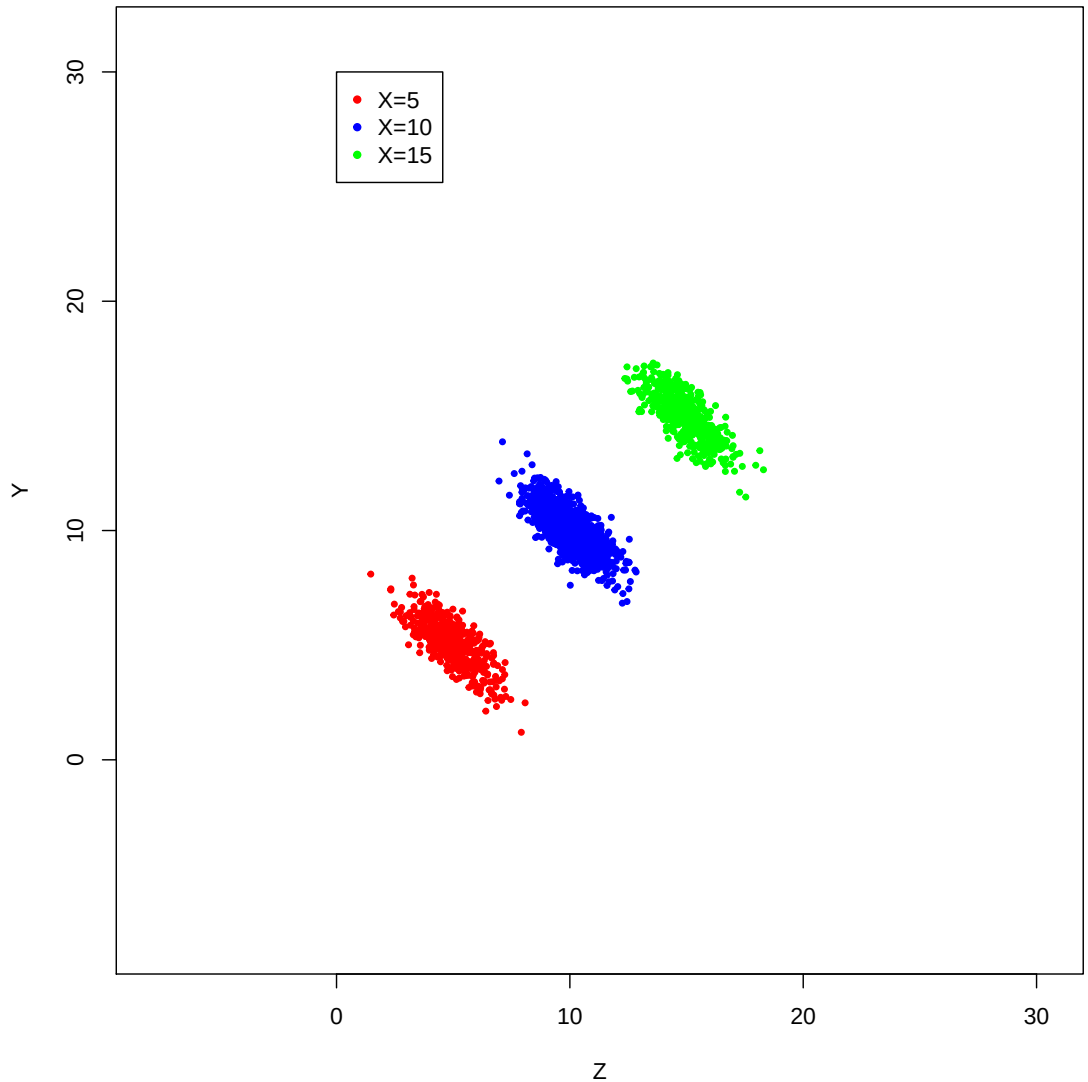
Some basic probability calculus



Some basic probability calculus



Some basic probability calculus



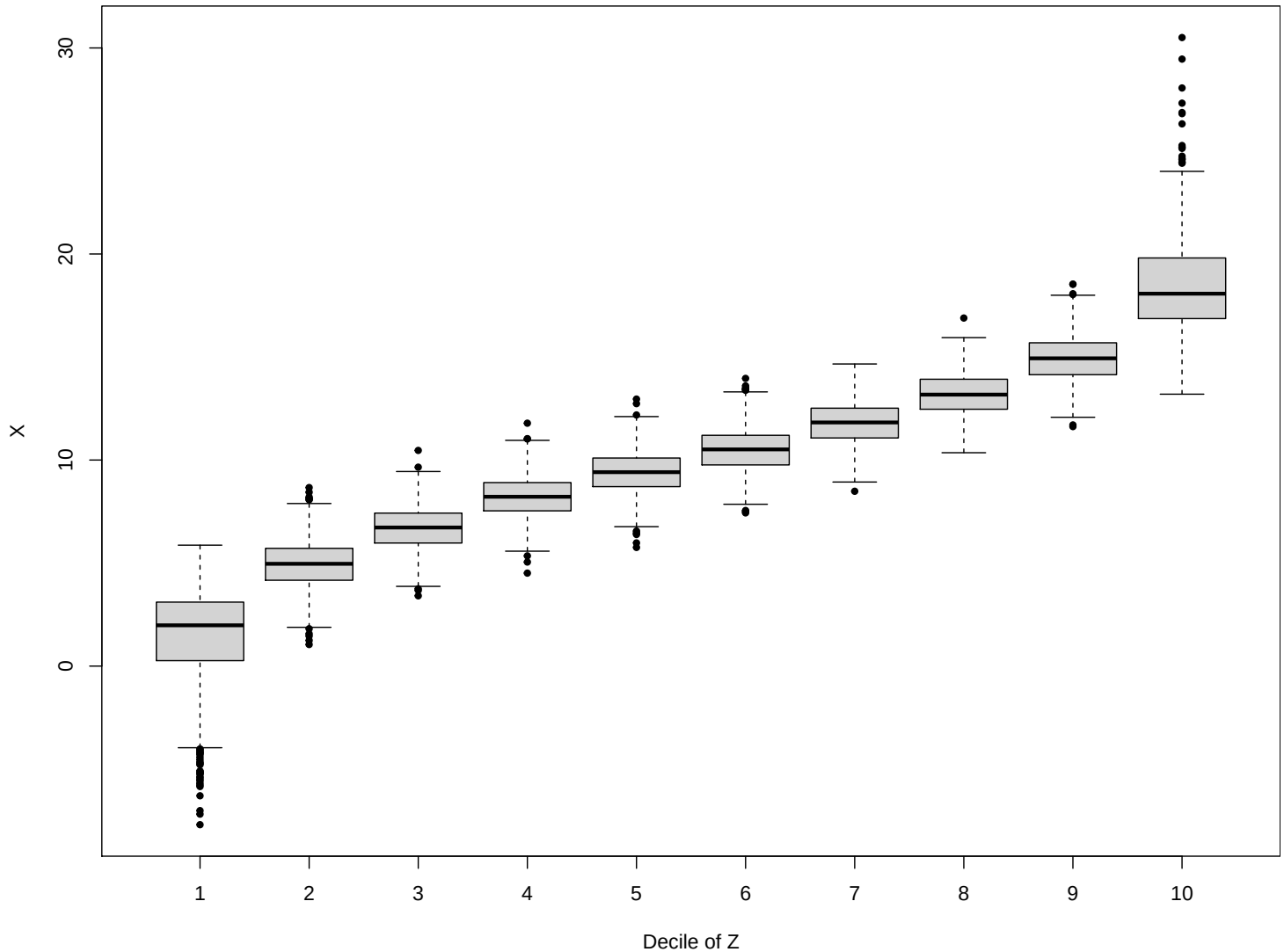
Some basic probability calculus

In this model

- Marginally, Y and Z are *positively* correlated;
- Conditionally on any $X = x$, Y and Z have *negative* correlation (by construction the correlation is -0.9).

We can examine the distribution of X for each Z: the following boxplot splits the data by deciles of Z.

Some basic probability calculus



Some basic probability calculus

By standard theory for the multivariate Normal distribution

$$Y|X = x, Z = z \sim Normal(x + \rho(z - x), (1 - \rho^2))$$

that is

$$\mathbb{E}[Y|X = x, Z = z] = x + \rho(z - x) = \rho z + (1 - \rho)x$$

Some basic probability calculus

Also

$$Z|X = x \sim Normal(x, 1)$$

and so

$$f_{X|Z}(x|z) \propto f_{Z|X}(z|x) f_X(x)$$

$$\equiv Normal\left(\frac{z + \mu_X/\sigma_X^2}{1 + 1/\sigma_X^2}, \frac{1}{1 + 1/\sigma_X^2}\right)$$

Some basic probability calculus

Therefore

$$\begin{aligned}
\mathbb{E}_{Y|Z}[Y|Z=z] &= \mathbb{E}_{X|Z}\left[\mathbb{E}_{Y|X,Z}[Y|X,Z=z]\middle|Z=z\right] \\
&= \mathbb{E}_{X|Z}\left[\rho z + (1-\rho)X|Z=z\right] \\
&= \rho z + (1-\rho)\mathbb{E}_{X|Z}\left[X|Z=z\right] \\
&= \rho z + (1-\rho)\frac{z+\mu_X/\sigma_X^2}{1+1/\sigma_X^2} \\
&= \frac{(1-\rho)\mu_X}{\sigma_X^2+1} + \left(\rho + (1-\rho)\frac{\sigma_X^2}{\sigma_X^2+1}\right) z
\end{aligned}$$

Some basic probability calculus

In this system, X and Z are *not independent*, and so the marginal effect on Y of changing Z is measured by the coefficient of z in $\mathbb{E}_{Y|Z}[Y|Z = z]$, that is,

$$\rho + (1 - \rho)\frac{\sigma_X^2}{\sigma_X^2 + 1}$$

whereas if we imagine manipulating Z *independently* of X, the effect is measured by the coefficient in

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

that is, ρ.

Some basic probability calculus

Note that in this system, the conditional model for Y, given $X = x$ and $Z = z$, in particular the mean model

$$\mathbb{E}_{Y|X,Z}[Y|X = x, Z = z] = \rho z + (1 - \rho)x$$

is *unchanged* irrespective of any assumption about the (X, Z) distribution.

Thus the critical distinction concerns whether we imagine Z being manipulated independently of X.

Some basic probability calculus

Here we have

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \sim Normal_3 \left(\begin{bmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{bmatrix} \right)$$

where by iterated expectation we can conclude

$$\mu_Y = \mu_Z = \mu_X.$$

Some basic probability calculus

By the general result for the multivariate normal distribution

$$\begin{bmatrix} Y \\ Z \end{bmatrix} \Big| X = x \sim Normal_2 \left(\begin{bmatrix} \mu_X \\ \mu_X \end{bmatrix} + \frac{1}{\sigma_X^2} \begin{bmatrix} \sigma_{XY} \\ \sigma_{XZ} \end{bmatrix} (\mathbf{x} - \mu_X), \Sigma_{YZ.X} \right)$$

where

$$\Sigma_{YZ.X} = \begin{bmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z^2 \end{bmatrix} - \frac{1}{\sigma_X^2} \begin{bmatrix} \sigma_{XY} \\ \sigma_{XZ} \end{bmatrix} \begin{bmatrix} \sigma_{XY} & \sigma_{XZ} \end{bmatrix}$$

Some basic probability calculus

We must have by construction

$$\Sigma_{YZ.X} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

so that $\sigma_{XY} = \sigma_{XZ} = \sigma_X^2$ and

$$\sigma_Y^2 = 1 + \frac{\sigma_{XY}^2}{\sigma_X^2} = 1 + \sigma_X^2$$

$$\sigma_Z^2 = 1 + \sigma_X^2$$

$$\sigma_{YZ} = \rho + \frac{\sigma_{XY}\sigma_{XZ}}{\sigma_X^2} = \rho + \sigma_X^2.$$

Some basic probability calculus

Here

$$\rho = \mathrm{Corr}[Y, Z|X = x] = \rho_{YZ.X}$$

is the *partial correlation* between Y and Z given $X = x$, which is different from

$$\rho_{YZ} = \frac{\sigma_{YZ}}{\sqrt{\sigma_Y^2 \sigma_Z^2}} = \frac{\rho + \sigma_X^2}{1 + \sigma_X^2}$$

which is the ordinary *correlation*.

Some basic probability calculus

Example:

See `knitr 01`

Regression models

- Y scalar
- $\mathbf{x}$ is $1 \times p$
- β is $p \times 1$

Often we model using a *linear combination*

$$\mathbb{E}[Y|\mathbf{x}] = g(\mathbf{x}\beta)$$

for some mapping function $g(.)$, and assume

$$\text{Var}[Y|\mathbf{x}] = V(\mathbf{x})$$

for some non-negative function $V(.)$.

Regression models

Most commonly for continuous-valued Y

$$\mathbb{E}[Y|\mathbf{x}] = \mathbf{x}\beta$$

and

$$\text{Var}[Y|\mathbf{x}] = \sigma^2$$

Regression models

For a data set of size n comprising

- outcome data $\mathbf{y} = (y_1, \ldots, y_n)^\top$
- predictor data $\mathbf{X}$ – an $n \times p$ matrix

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix}$$

we assume

$$\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta \qquad \mathrm{Var}[\mathbf{Y}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$$

Regression models

This is equivalent to the model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

where ε is an $(n \times 1)$ vector of random variables with

$$\mathbb{E}[\varepsilon|\mathbf{X}] = \mathbf{0}_n \qquad \text{Var}[\varepsilon|\mathbf{X}] \equiv \mathbb{E}[\varepsilon\varepsilon^\top|\mathbf{X}] = \sigma^2\mathbf{I}_n$$

Regression models

We may choose to treat $\mathbf{X}$ as *fixed* or *random* quantities.

- with $\mathbf{X}$ fixed, estimate parameters β and σ^2 using *ordinary least squares* (OLS)

$$\widehat{\beta} = \arg\min_{\beta} (\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{y} - \mathbf{X}\beta)$$

that is, $\widehat{\beta}$ solves

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}_p$$

so that

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

and

$$\widehat{\sigma}^2 = \frac{1}{n-p} (\mathbf{y} - \mathbf{X}\widehat{\beta})^\top (\mathbf{y} - \mathbf{X}\widehat{\beta})$$

Note that

$$\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\beta) = \sum_{i=1}^{n} \mathbf{x}_i^\top(y_i - \mathbf{x}_i\beta)$$

showing the form of the *estimating function*.

Regression models

- with $\mathbf{X}$ random, using the model equation

$$\mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{X}\beta + \mathbf{X}^\top \varepsilon$$

and taking expectations with respect to the *joint* distribution

$$\mathbb{E}[\mathbf{X}^\top \mathbf{Y}] = \mathbb{E}[\mathbf{X}^\top \mathbf{X}]\beta + \mathbb{E}[\mathbf{X}^\top \varepsilon].$$

By assumption

$$\mathbb{E}[\mathbf{X}^\top \varepsilon] = \mathbb{E}_{\mathbf{X}}[\mathbf{X}^\top \mathbb{E}_{\varepsilon|\mathbf{X}}[\varepsilon|\mathbf{X}]] = \mathbf{0}_p.$$

Thus

$$\mathbb{E}[\mathbf{X}^\top \mathbf{Y}] = \mathbb{E}[\mathbf{X}^\top \mathbf{X}]\beta$$

and provided $\mathbb{E}[\mathbf{X}^\top \mathbf{X}]$ is non-singular, we have

$$\beta = \{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]\}^{-1}\mathbb{E}[\mathbf{X}^\top \mathbf{Y}].$$

Note also that

$$\mathbb{E}[\mathbf{X}^\top (\mathbf{Y} - \mathbf{X}\beta)] = \mathbf{0}_p$$

Using the method of moments, we have that

$$\{\mathbb{E}[\mathbf{X}^\top \mathbf{X}]\}^{-1} \text{ is estimated by } \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i \right\}^{-1}$$

and

$$\mathbb{E}[\mathbf{X}^\top \mathbf{Y}] \text{ is estimated by } \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top y_i$$

yielding an identical result to OLS.

Regression models

By standard theory, we have for

$$\widehat{\beta}_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

that as the sample size grows the corresponding estimator

$$\widehat{\beta}_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

has good behaviour

Regression models

- *consistency*: as $n \longrightarrow \infty$,

$$\widehat{\beta}_n \xrightarrow{p} \beta_{\text{TRUE}}$$

with β_{TRUE} the true (data generating) value.

- *asymptotic normality*

$$\sqrt{n}(\widehat{\beta}_n - \beta_{\text{TRUE}}) \xrightarrow{d} Normal_p(\mathbf{0}_p, \sigma^2 \mathbf{V})$$

where

$$\mathbf{V}^{-1} = \operatorname*{plim}_{n \longrightarrow \infty} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i \right\}$$

provided the limit exists and is non-singular.

Note

This theory holds assuming *correct specification* of $\mathbb{E}_{\mathbf{Y}|\mathbf{X}}[\mathbf{Y}|\mathbf{X}]$.

- $\mathbf{Y} - \mathbf{X}\beta$ is *uncorrelated* with the columns of $\mathbf{X}$.

A parallel theory holds under *mis-specification*; however, most critically we *do not* obtain consistent estimators if the mean model is mis-specified.

Regression models

Example:

Suppose we specify

$$\mathbb{E}_{Y|\mathbf{x}}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + z(\psi_0 + \psi_1 x_1).$$

so that

$$\mathbf{x}_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & z_i & z_i x_{i1} \end{bmatrix}.$$

Using the above formulae, this leads us to estimates

$$\widehat{\beta}_n = \begin{bmatrix} \widehat{\beta}_0 & \widehat{\beta}_1 & \widehat{\beta}_2 & \widehat{\psi}_0 & \widehat{\psi}_1 \end{bmatrix}^\top.$$

Example:

Then we may estimate the expected response for Z set to take the value z as

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

Then, if we compare $z = 1$ with $z = 0$ we get the estimated difference

$$\widehat{\mathbb{E}}[Y|x_1, x_2, 1] - \widehat{\mathbb{E}}[Y|x_1, x_2, 0] = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\psi}_0 + \widehat{\psi}_1 x_{i1})$$

Moment-based estimation & sample averages

The idea of *moment-based estimation* is to estimate expectations using *sample averages*.

- Sample mean:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  is an estimator of

$$\mu = \mathbb{E}_X[X] = \int x f_X(x)\, dx$$

Moment-based estimation & sample averages

- Generalized version:

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

  is an estimator of

$$\mathbb{E}_X[g(X)] = \int g(x) f_X(x)\, dx$$

Moment-based estimation & sample averages

We are approximating the integral

$$\int g(x) f_X(x)\ dx$$

by the *empirical* version

$$\int g(x) \widehat{f}_n(x)\ dx$$

where

$$\widehat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i\}}(x)$$

We can think of this as a type of *Monte Carlo* calculation.

Moment-based estimation & sample averages

Monte Carlo estimation is reliant on the fact that as $n \longrightarrow \infty$, we have certain types of *convergence*:

- *Laws of large numbers:* Suppose $X_1, \ldots, X_n, \ldots$ are iid random variables. Then, usually,

$$\frac{1}{n} \sum_{i=1}^{n} g(X_i) \xrightarrow[\text{p}]{\text{a.s.}} \mathbb{E}_X[g(X)]$$

Moment-based estimation & sample averages

- *Central Limit Theorems:* Under mild conditions on the joint distribution of random variables $X_1, \ldots, X_n, \ldots$,

$$a_n \left(\frac{1}{n} \sum_{i=1}^{n} g(X_i) - b_n \right) \xrightarrow{d} Normal(\mu, \sigma^2)$$

for suitable choices of the sequences $\{a_n\}$ and $\{b_n\}$.

Moment-based estimation & sample averages

Essentially, standardized sums of random variables have stable long-run behaviour. For example,

$$\overline{X}_n \xrightarrow{p} \mathbb{E}_X[X] \qquad \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}_X[X^2]$$

as $n \longrightarrow \infty$, and so on.

Moment-based estimation & sample averages

Importance sampling: The identity

$$\int g(x)f(x)\,dx = \int g(x)f(x)\frac{f_0(x)}{f_0(x)}\,dx = \int \frac{g(x)f(x)}{f_0(x)}f_0(x)\,dx$$

where f_0 is a probability density with support including the support of f: that is we must choose f_0 such that

$$f_0(x) > 0 \quad \text{whenever} \quad f(x) > 0$$

Moment-based estimation & sample averages

That is,

$$\mathbb{E}_f[g(X)] = \mathbb{E}_{f_0}\left[\frac{g(X)f(X)}{f_0(X)}\right]$$

so that an estimator of the LHS is

$$\widehat{I}_N^{(f_0)}(g) = \frac{1}{N}\sum_{i=1}^{N}\frac{g(X_i)f(X_i)}{f_0(X_i)}$$

where $X_1, \ldots, X_N \sim f_0(.)$.

- $\widehat{I}_N^{(f_0)}$ is termed the *importance sampling* estimator.
- f_0 is termed the *importance sampling density*.

Moment-based estimation & sample averages

Note

The importance sampling method tells us that even if we have an expectation that we need to estimate for distribution f, we can instead use variates (or 'data') sampled from a *different* distribution f_0.

Moment-based estimation & sample averages

By careful choice of f_0, the estimator can have better performance than the Monte Carlo estimator in finite samples.

Note that

$$\widehat{I}_N^{(f_0)}(g) = \frac{1}{N}\sum_{i=1}^{N}\frac{f(X_i)}{f_0(X_i)}g(X_i) = \frac{1}{N}\sum_{i=1}^{N} w_0(X_i)g(X_i)$$

say, where

$$w_0(X_i) = \frac{f(X_i)}{f_0(X_i)}$$

is the *importance sampling weight*.

Moment-based estimation & sample averages

Note that

$$\mathbb{E}_{f_0}[w_0(X)] \equiv \mathbb{E}_{f_0}\left[\frac{f(X)}{f_0(X)}\right] = \int f(x)\,dx = 1$$

so

$$\mathbb{E}_{f_0}\left[\frac{1}{N}\sum_{i=1}^{N} w_0(X_i)\right] = 1$$

although for any realization

$$\frac{1}{N}\sum_{i=1}^{N} w_0(x_i) \neq 1$$

in general.

Moment-based estimation & sample averages

Example:

Consider the two distributions for variables X, Y, Z:

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

$$f^* : f^*_{X,Y,Z}(x,y,z) = f_X(x) f^*_Z(z) f_{Y|X,Z}(y|x,z) \qquad \text{i.e. } X \perp\!\!\!\perp Z$$

so that

$$\frac{f^*_{X,Y,Z}(x,y,z)}{f_{X,Y,Z}(x,y,z)} = \frac{f^*_Z(z)}{f_{Z|X}(z|x)}$$

Moment-based estimation & sample averages

Example:

Thus, for any function $g(x, y, z)$, using the importance sampling idea

$$\mathbb{E}_{f^*}[g(X, Y, Z)] = \mathbb{E}_f\left[\frac{f_Z^*(Z)}{f_{Z|X}(Z|X)} g(X, Y, Z)\right]$$

provided, for all z such that $f_Z^*(z) > 0$, we have $f_{Z|X}(z|x) > 0$ for all x.

Moment-based estimation & sample averages

Example:

We are *reweighting* contributions to the expectation to account for the fact that

- under f^*, each contribution $g(x, y, z)$ gets weight determined by $f^*_Z(z)$;
- under f, each contribution $g(x, y, z)$ gets weight determined by $f_{Z|X}(z|x)$

Part 2
Causal Graphs

Causal Graphs

The structure of a joint distribution is essentially specified the set of *conditional distributions* that appear in the chain rule factorization. In general we have

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

but perhaps we might assume that

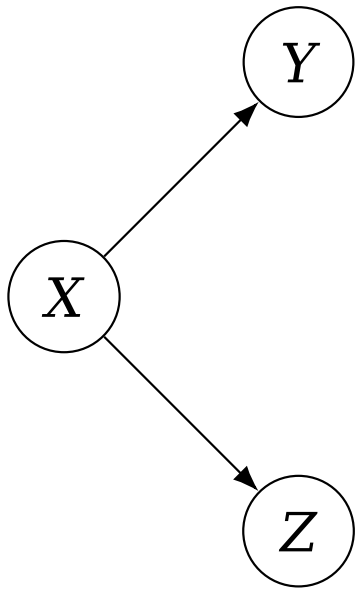
$$Z \perp\!\!\!\perp Y|X$$

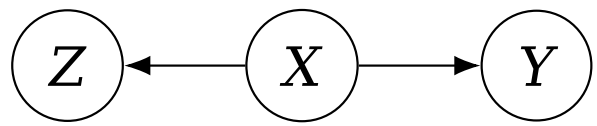
so that $f_{Z|X,Y}(z|x,y) = f_{Z|X}(z|x)$ and

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

Causal Graphs

We can depict the conditional independence using a graph:





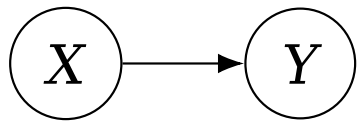
This type of graph is sometimes called a *fork*.

Causal Graphs

- Nodes (X), (Y), (Z) 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*;

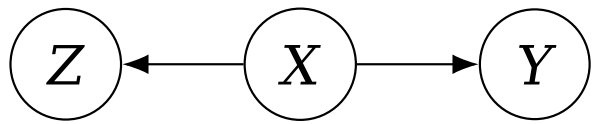


corresponds to

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

Causal Graphs

- Nodes with only *outgoing* edges act to *block* dependence: in



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

it is evident that $Y \perp\!\!\!\perp Z|X$.

Causal Graphs

However, note that

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

By standard probability calculus

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

$$f_Y(y) = \int f_{Y|X}(y|x) f_X(x)\ dx$$

$$f_Z(z) = \int f_{Z|X}(z|x) f_X(x)\ dx$$

so in general

$$f_{Y,Z}(y,z) \neq f_Y(y) f_Z(z).$$

Causal Graphs

A (causal) graph G is described using the following elements:

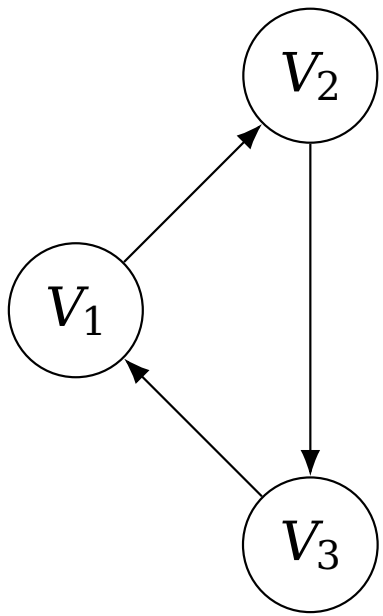
- A set of *nodes* or *vertices*, $V_1, V_2, \ldots$, representing variables.
- A set of *edges*, $E_1, E_2, \ldots$, representing dependencies.
- Two nodes are *adjacent* if there is an edge between them.
- Edges can be *directed*, denoted using arrows, or *undirected*; if all edges are directed, the graph is directed.
- The graph with the arrow directions removed is termed the *skeleton*.

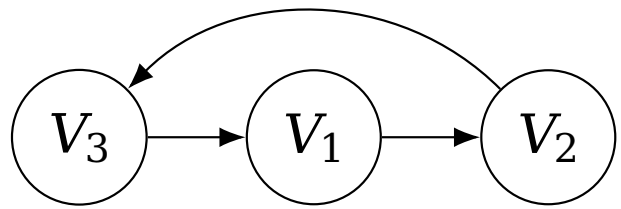
Causal Graphs

- A *path* between to nodes V_1 and V_2 is a sequence of edges connecting those nodes;
  - a *directed* path is a path where the directions of arrows on edges are obeyed.
  - 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.

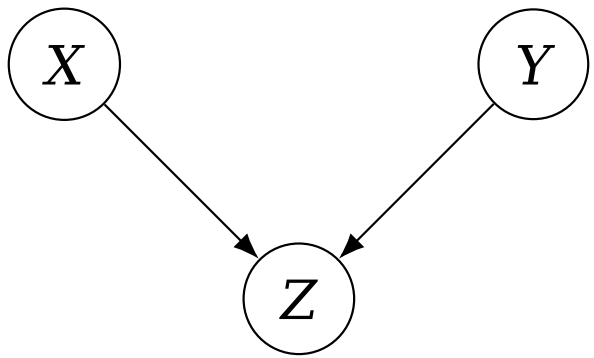




A directed graph that has no cycles is termed a *directed acyclic graph* (DAG).

Causal Graphs

The language of *'kinship'* may be used to describe graphical connections:



PARENTS

CHILD

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

Causal Graphs

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

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

as

$$\int f_{Z|X,Y}(z|x,y)\ dz = 1$$

Causal Graphs

However, conditioning on $Z = z$

$$
\begin{aligned}
f_{X,Y|Z}(x, y|z) &= \frac{f_{X,Y,Z}(x, y, z)}{f_Z(z)} && \text{definition} \\
&= \frac{f_X(x) f_Y(y) f_{Z|X,Y}(z|x, y)}{f_Z(z)} && \text{by assumption} \\
&= f_X(x) f_Y(y) \frac{f_{Z|X,Y}(z|x, y)}{f_Z(z)} \\
&\neq f_{X|Z}(x|z) f_{Y|Z}(y|z) && \text{also} \neq f_X(x) f_Y(y)
\end{aligned}
$$

in general, as the function in line 3 does not factorize into a function of (x, z) and a function of (y, z), which would be necessary for conditional independence.

Causal Graphs

That is,

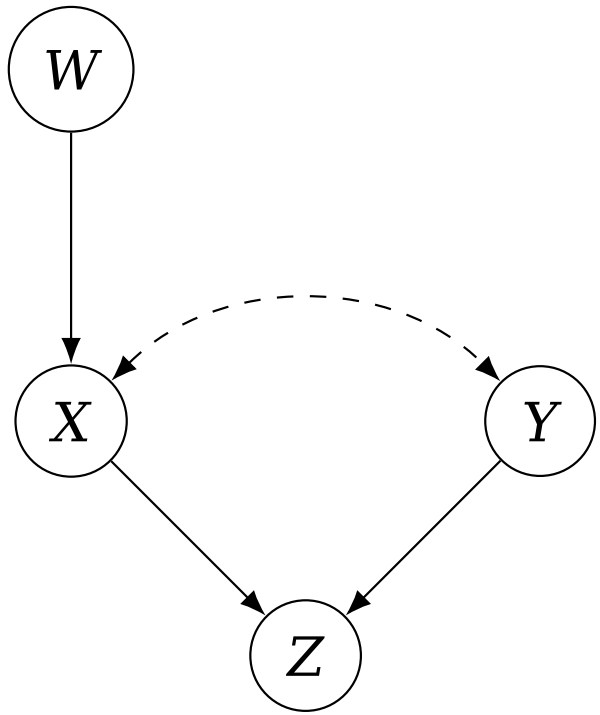
$$X \perp\!\!\!\perp Y$$

but

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

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

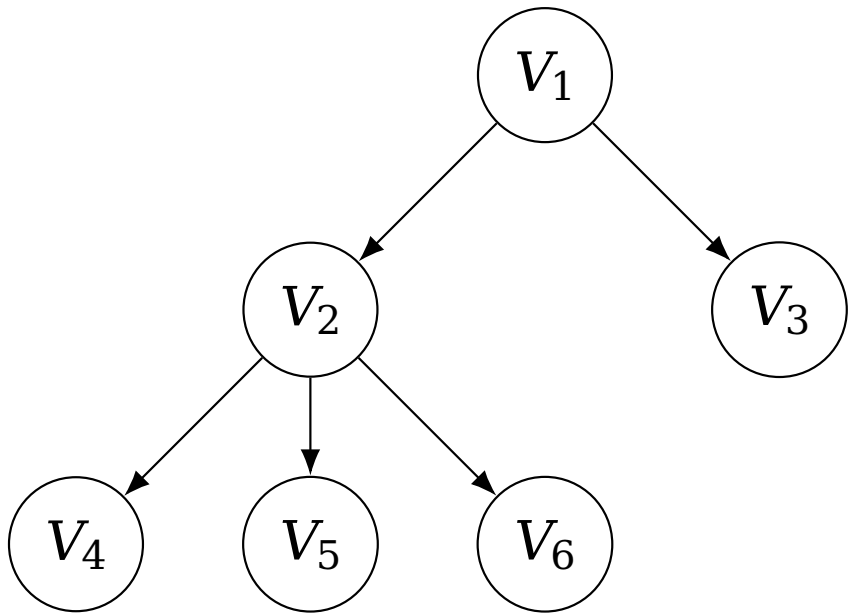
Causal Graphs



- X and Y are *spouses*, and *parents* of Z.
- X, Y and W are *ancestors* of Z.
- X is a *child* of W.
- Z is a *child* of X and Y.

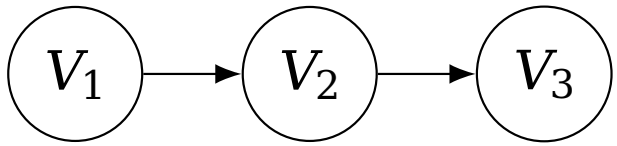
Causal Graphs

A *tree* is a graph where each node has at most one *parent*.



Causal Graphs

A *chain* is a graph where each node has at most one *child*.



Causal Graphs

For variables $X_1, X_2, \ldots, X_d$, define for $j = 1, \ldots, d$ the *set of parents* of X_j, denoted

$$\mathrm{PA}_j \equiv \{X_1^{(j)}, \ldots, X_{n_j}^{(j)}\}, \text{ say}$$

such that for $X_k \notin \mathrm{PA}_j$,

$$X_j \perp\!\!\!\perp X_k \mid X_1^{(j)}, \ldots, X_{n_j}^{(j)}$$

and that no proper subset of PA_j yields the conditional independence. That is, PA_j is the *smallest* set of variables for which the conditional independence statement holds.

Causal Graphs

We have for the chain rule factorization

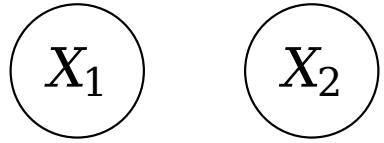
$$f_{X_1,\ldots,X_d}(x_1,\ldots,x_d) = f_{X_1}(x_1)\prod_{j=2}^{d} f_{X_j|X_1,\ldots,X_{j-1}}(x_j|x_1,\ldots,x_{j-1})$$

$$\equiv f_{X_1}(x_1)\prod_{j=2}^{d} f_{X_j|\mathrm{PA}_j}(x_j|x_1^{(j)},\ldots,x_{n_j}^{(j)})$$

To construct the factorization, we start with the *founders* for which the parent set is *empty*.

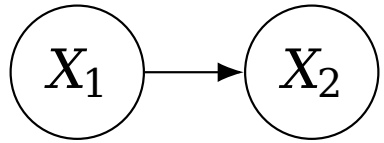
Causal Graphs

X_1, X_2



$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$$

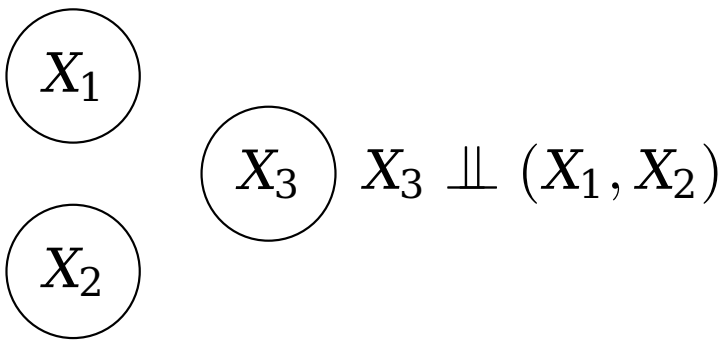
or

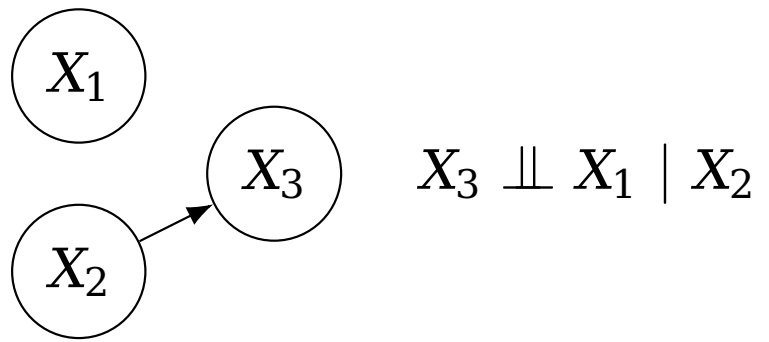


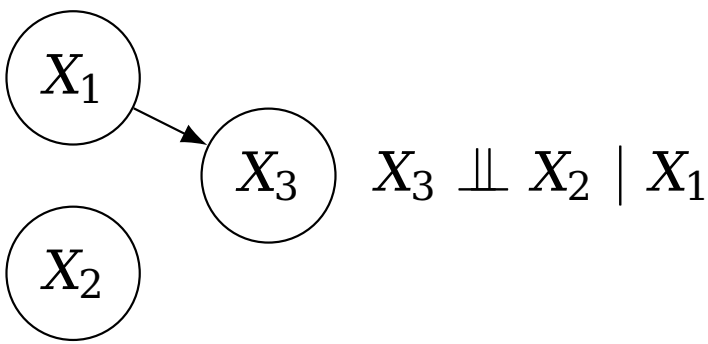
$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1)$$

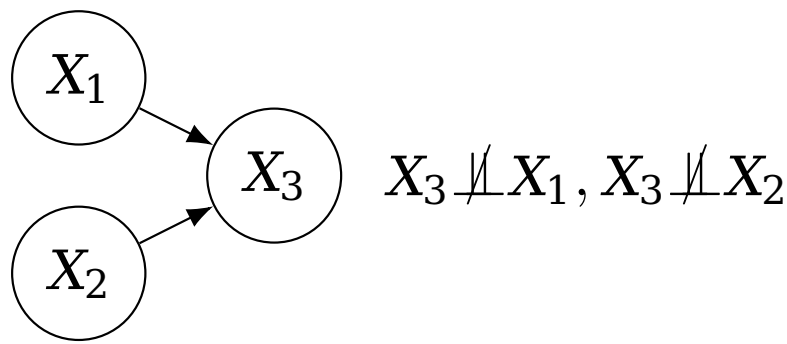
Causal Graphs

Add in X_3: independence case



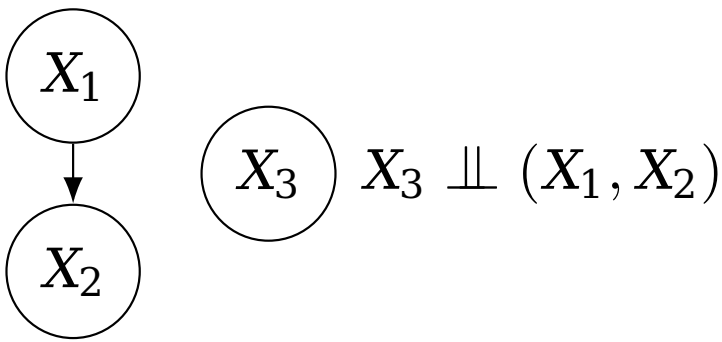


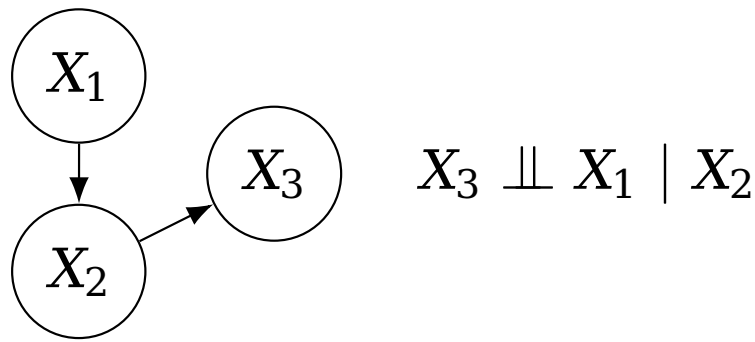


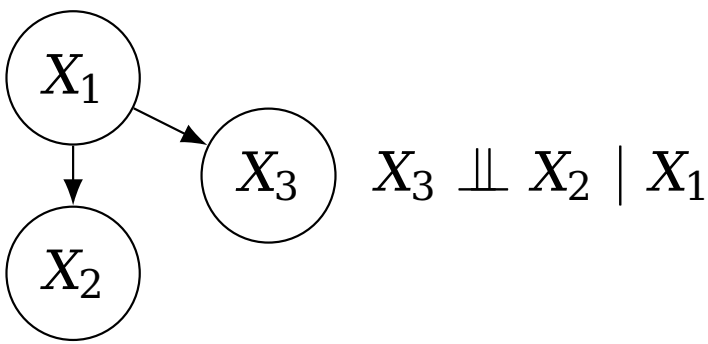


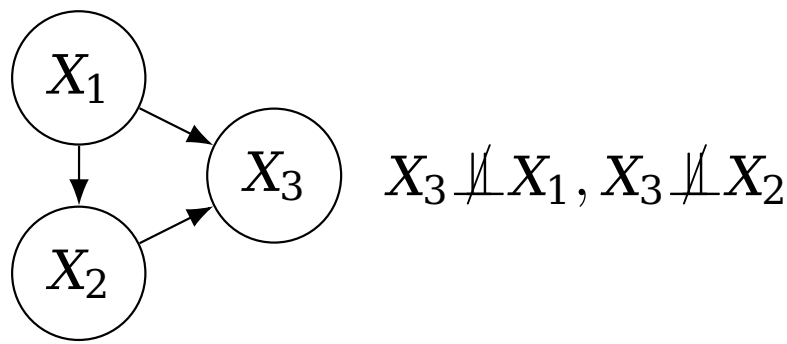
Causal Graphs

Add in X_3: dependence case









Causal Graphs

Compatibility: The probability distribution P is *compatible* with graph G if P admits the factorization implied by G.

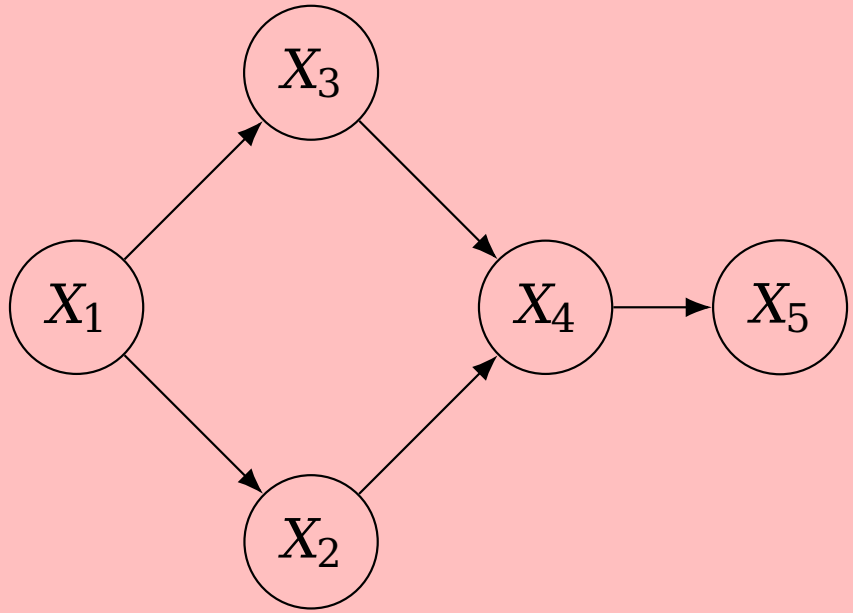
- note the G does not define P, merely the chain rule factorization that P admits;
- termed '*Markov compatibility*'; P is *Markov with respect to* G, that is, we may deduce from G that

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

say, but we do not know the forms or values of the individual terms.

Causal Graphs

Example:



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

Note

We need so ensure that all conditional densities are well-defined, that is, we must condition on values that carry are in the *support* of the *marginal density* for the conditioning variables. For example, we can only compute

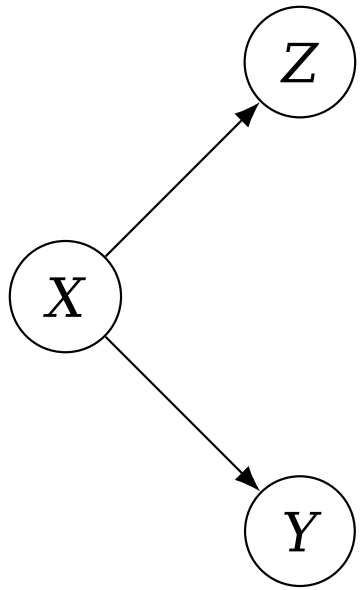
$$f_{X_3|X_1,X_2}(x_3|x_1,x_2)$$

for (x_1, x_2) such that

$$f_{X_1,X_2}(x_1,x_2) > 0.$$

Structural models

When we write



what precisely (mechanistically) does the symbol $\longrightarrow$ mean ?

Structural models

One interpretation is via a *structural* interpretation:

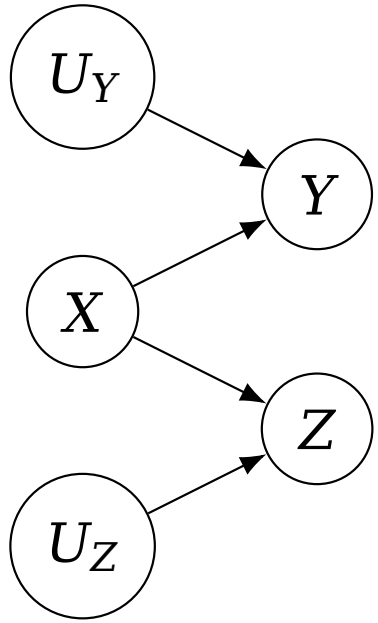
- generate X independently,
- generate Y and Z independently as functions of the realized X, for example

$$Y = 3X$$

$$Z = 4X + 9$$

Structural models

X, U_Z, U_Y
independent



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

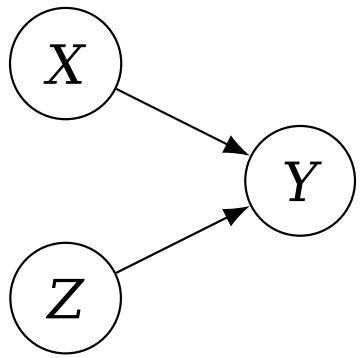
$$Z = g_2(X, U_Z)$$

For example

$$Y = X + U_Y$$

$$Z = X + U_Z$$

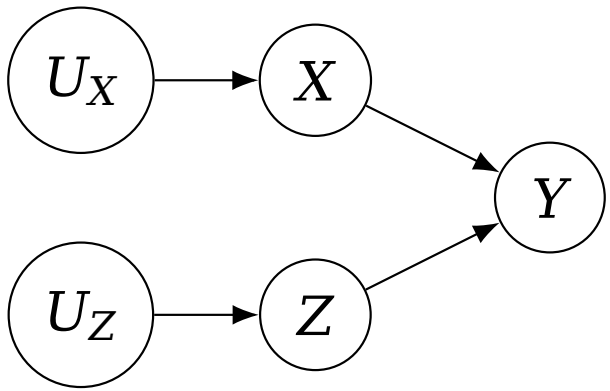
Structural models



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

If we fixed $X = x$ and $Z = z$, we would know $Y = g(x, z)$ *precisely*.

Structural models



so that $X = g_1(U_X)$, $Z = g_2(U_Z)$, and

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

Structural models

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

- 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

Variables that have no 'causes' (ancestors) are termed *exogenous*; variables that have at least one cause are termed *endogenous*.

d-separation

Consider three disjoint sets of nodes

$$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$$

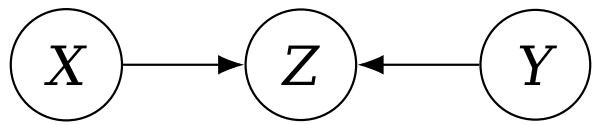
of DAG G. To assess whether

$$X \perp\!\!\!\perp Y \mid Z \qquad \forall X \in \mathcal{X}, Y \in \mathcal{Y}, Z \in \mathcal{Z}$$

for any distribution compatible with the DAG, we must assess whether Z *'blocks'* paths from X to Y.

d-separation

Consider the *collider* ('inverted fork') graph

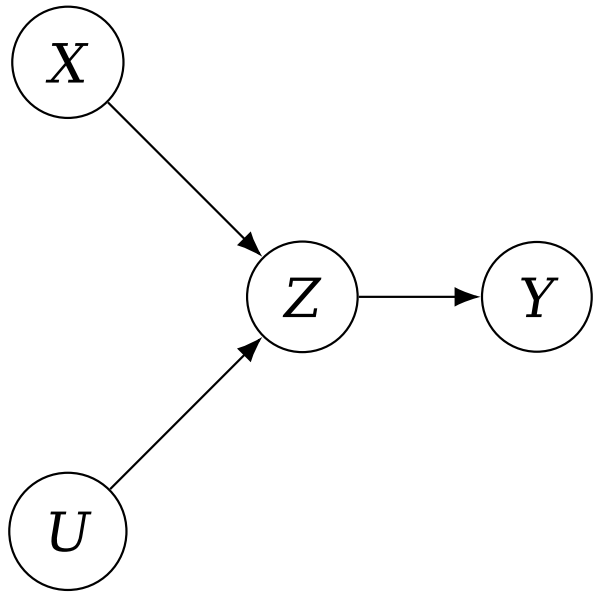


Z is a collider on this path, and the path from X to Y is blocked at Z.

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



- Z is a *collider* on XZU
- Z is *not a collider* on XZY.

Unconditional d-separation

A path is *unconditionally open*[1] if there is no collider on the path; if there is a collider, the path is *closed*.

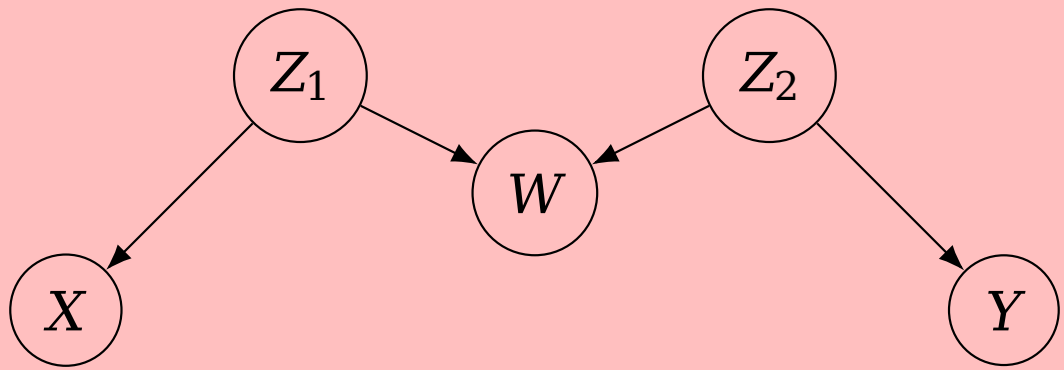
Variables X and Y are *d-separated* if there is no open path between them; if there is an open path, the two variables are *d-connected*.

[1] or *unblocked*, or *active*

Unconditional d-separation

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

- Z_1 family income
- Z_2 genetic risk
- W parental diabetes
- X low educational attainment
- Y diabetes of subject



Unconditional d-separation

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

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

$$f_{Z_1}(z_1) f_{Z_2}(z_2) f_{W|Z_1,Z_2}(w|z_1, z_2) f_{X|Z_1}(x|z_1) f_{Y|Z_2}(y|z_2)$$

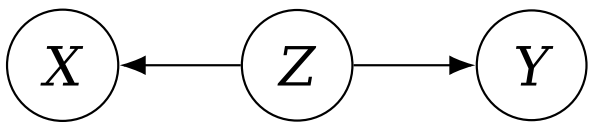
and X and Y are *independent*:

- integrate out w, then z_1, then z_2.

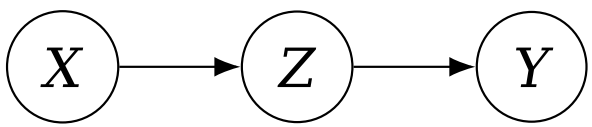
Conditional d-separation

We can consider similar statements obtained after *conditioning* on a variable.

For a non-collider Z: consider conditioning on Z



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

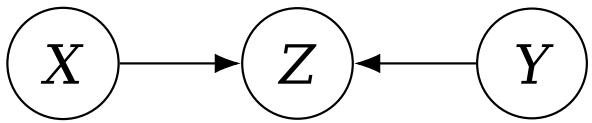


$X \perp\!\!\!\perp Y \mid Z$

Conditioning *blocks* the path, and yields $X \perp\!\!\!\perp Y$.

Conditional d-separation

For a collider Z: consider conditioning on Z

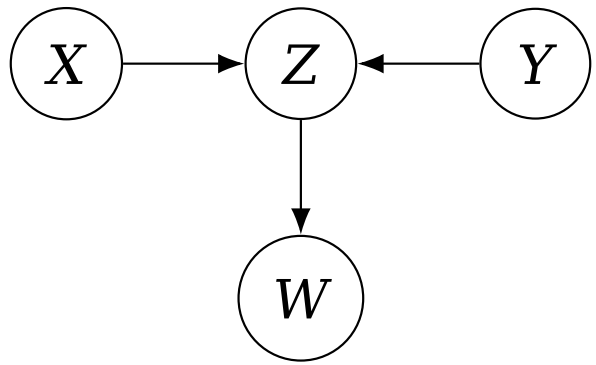


$$X \not\perp\!\!\!\perp Y \mid Z$$

Conditioning *unblocks* the path, and this yields $X \not\perp\!\!\!\perp Y$ conditionally.

Conditional d-separation

However, consider



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

Here X and Y are unconditionally independent as Z is a collider.

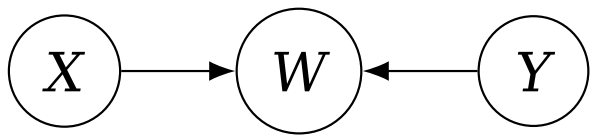
Conditional d-separation

However

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

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

Therefore we have that



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

and W is a collider in the marginalized model.

Conditional d-separation

Therefore

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

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

Conditional d-separation

Suppose S is a set of variables.

- S *blocks* a path from X to Y 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* from X to Y, then X and Y are *d-separated*.

Conditional d-separation

- If S d-separates X and Y, $X \perp\!\!\!\perp Y \mid S$,

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

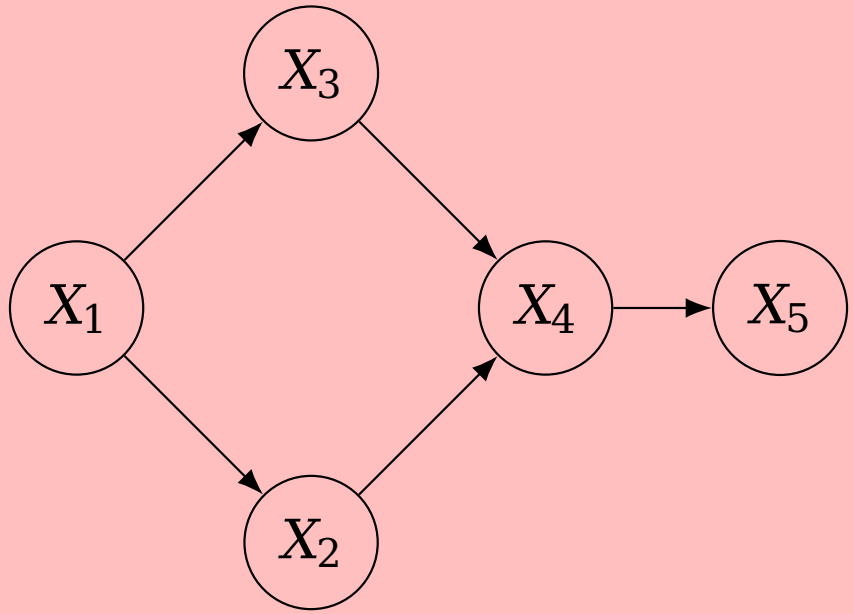
- If S does not d-separate X and Y, then X and Y may be dependent, and

$$f_{X|Y,S}(x|y,s)$$

  cannot be made independent of y in general.

Conditional d-separation

Example:

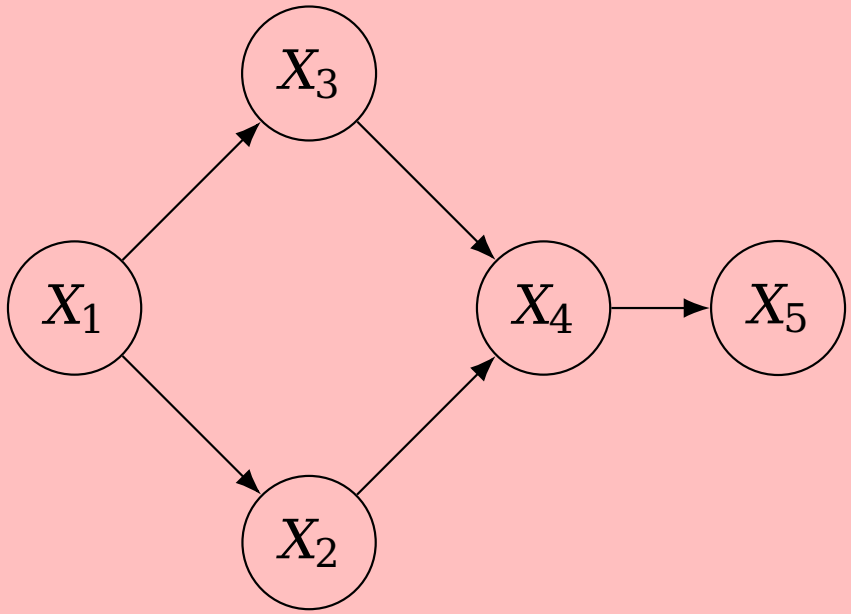


$\{X_2\}$ and $\{X_3\}$ are d-separated by $\{X_1\}$, and $X_2 \perp\!\!\!\perp X_3 \mid X_1$.

- there are two paths between X_2 and X_3;
  - $X_2X_1X_3$: blocked by conditioning on X_1.
  - $X_2X_4X_3$: 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;

Conditional d-separation

Selection bias: Conditioning on the common effect of two causes renders the two causes dependent; this is known as *selection bias* or *Berkson bias*.

Conditional d-separation

Example: Hospital cohort

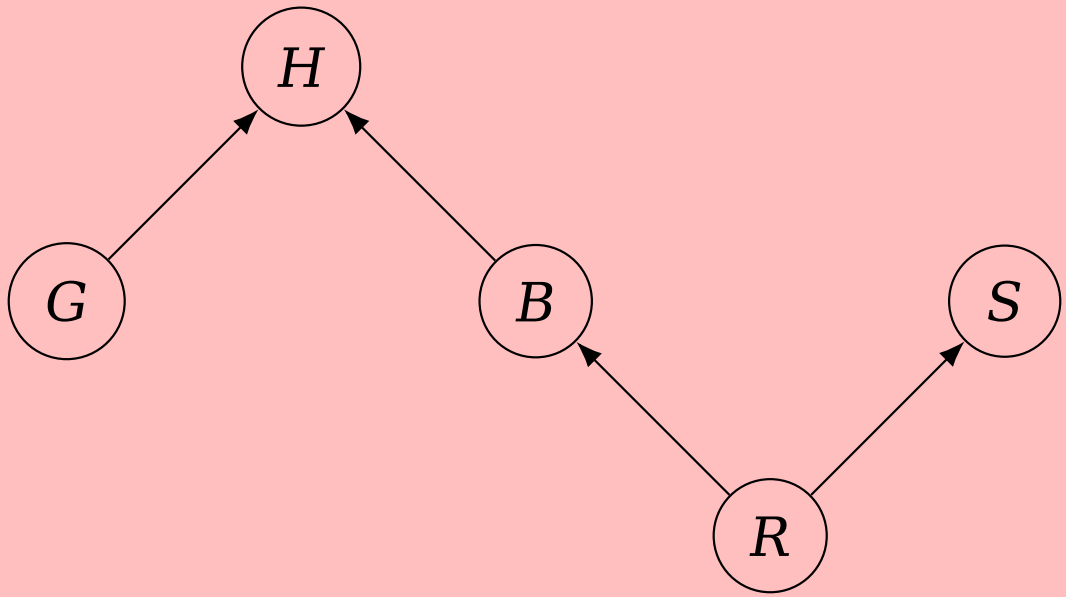
Consider the DAG for binary variables

- S – smoking status
- G – glioma
- H – hospitalization
- B – broken bone
- R – tendency for reckless behaviour

We wish to study the association between S and G in a hospitalized cohort.

Conditional d-separation

Example: Hospital cohort



The joint model $f_{R,B,S,G,H}(r,b,s,g,h)$ factorizes as

$$f_R(r)f_G(g)f_{B|R}(b|r)f_{S|R}(s|r)f_{H|B,G}(h|b,g)$$

Conditional d-separation

Example: Hospital cohort

In this graph, we have one path from S to G

- this is an undirected path,
- the path is blocked at collider H.

Therefore $G \perp\!\!\!\perp S$,

However, conditioning on H renders G and S *dependent*. That is, if we carry out a study in the general population, no association between will be detected, however, if we only look in hospitalized subjects, an association will be detected.

Conditional d-separation

Example: HIV study

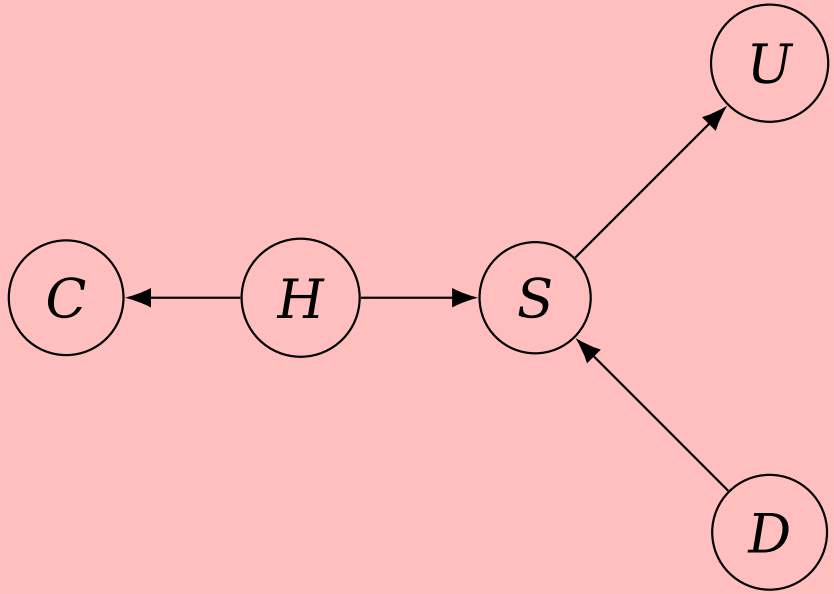
Consider the DAG for variables

- C – CD4 count (continuous)
- D – assignment of new HIV drug (binary)
- H – underlying HIV severity (binary)
- S – symptoms (binary)
- U – follow up indicator (binary, $U = 1$ implies that participant stays in trial)

We wish to examine the impact of D on C by studying patients who stay in the trial.

Conditional d-separation

Example: HIV study



The joint model $f_{H,D,C,S,U}(h,d,c,s,u)$ factorizes as

$$f_H(h)f_D(d)f_{C|H}(c|h)f_{S|H,D}(s|h,d)f_{U|S}(u|s)$$

Conditional d-separation

Example: HIV study

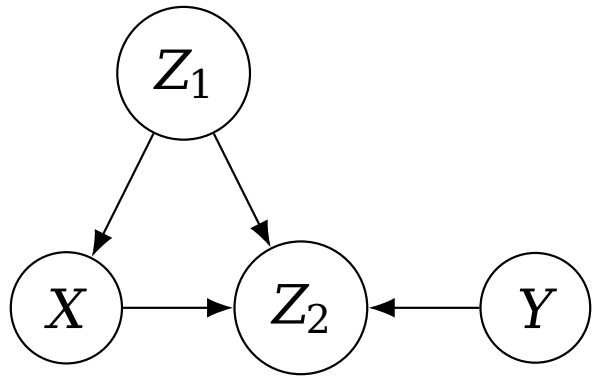
In this graph, we have one path from D to C

- this is an undirected path,
- the path is blocked at collider S.

Therefore $D \perp\!\!\!\perp C$, but conditioning on U, which is a descendant of S, renders D and C *dependent*.

Therefore, regressing C on D in the subgroup where $U = 1$ (that is, individuals who are followed in the trial) will indicate an association.

Conditional d-separation



Here $X \perp\!\!\!\perp Y$: there are two paths between X and Y

- $X \to Z_1 \to Z_2 \to Y$
- $X \to Z_2 \to Y$

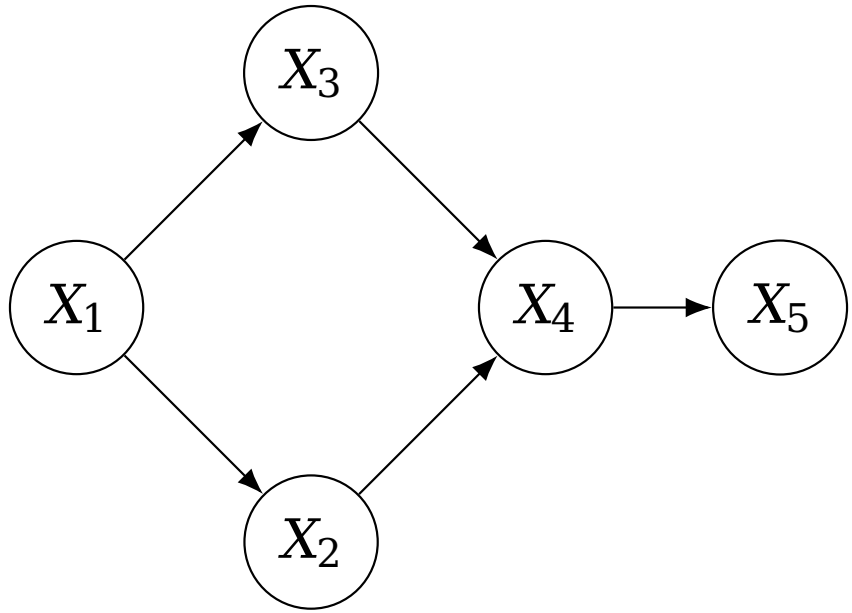
both blocked by collider Z_2. Therefore $X \not\perp\!\!\!\perp Y \mid \{Z_2\}$.

Conditional d-separation

If X and Y are d-separated by S then $X \perp\!\!\!\perp Y \mid S$ for all distributions compatible with G; conversely, if they are not d-separated, then X and Y are dependent given S for at least on distribution compatible with G.

Interventions

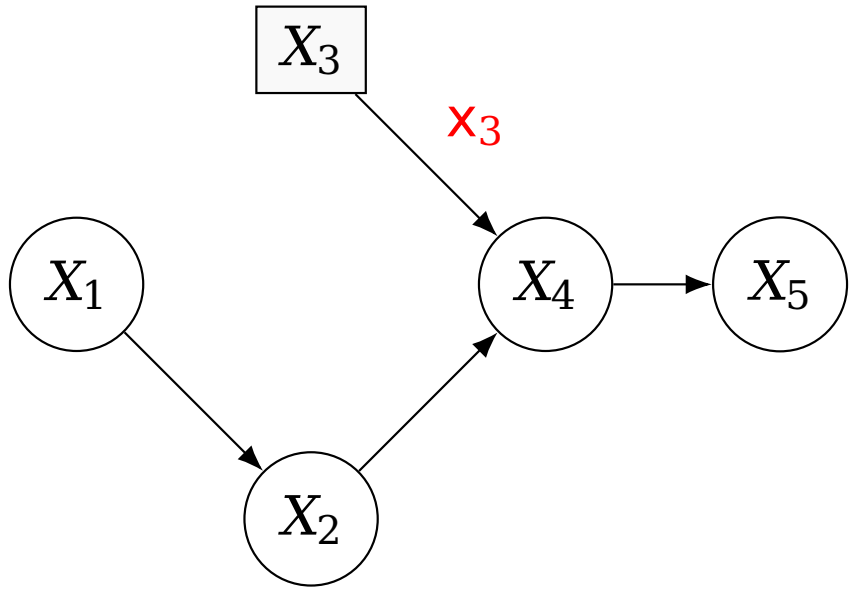
Consider the DAG



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

Interventions

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



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

where $f^*_{X_3}(.)$ is a *degenerate* distribution at x_3. X_1 is *no longer a cause* of X_3.

Interventions

Note

We note the distinction between the distributions

$$f_{X_1,X_2,X_4,X_5|X_3}(x_1, x_2, x_4, x_5|\mathsf{x}_3) = \frac{f_{X_1,X_2,X_3,X_4,X_5}(x_1, x_2, \mathsf{x}_3, x_4, x_5)}{f_{X_3}(\mathsf{x}_3)}$$

which arises from the *original* DAG, and

$$f^*_{X_1,X_2,X_4,X_5|X_3}(x_1, x_2, x_4, x_5|\mathsf{x}_3) = \frac{f^*_{X_1,X_2,X_3,X_4,X_5}(x_1, x_2, \mathsf{x}_3, x_4, x_5)}{f^*_{X_3}(\mathsf{x}_3)}$$

which arises from the *intervention* DAG.

Interventions

Note

In the causal literature, the distinction is sometimes acknowledged using the *'do' operator*

$$f_{X_1,X_2,X_4,X_5}(x_1, x_2, x_4, x_5 \mid \mathrm{do}(X_3) = \mathsf{x}_3)$$

is the same as

$$f^*_{X_1,X_2,X_4,X_5|X_3}(x_1, x_2, x_4, x_5 \mid \mathsf{x}_3)$$

This notation was introduced by J. Pearl.

Interventions

Intervening on variable X to set the level to x has the effect of

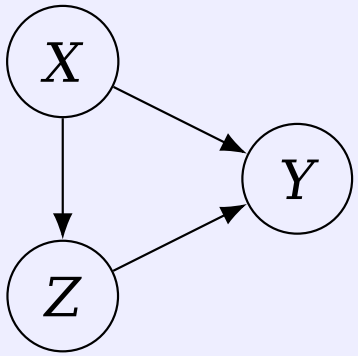
- removing all *incoming* arrows to X
- switching the marginal for X to the *degenerate distribution* $f_X^*(.)$

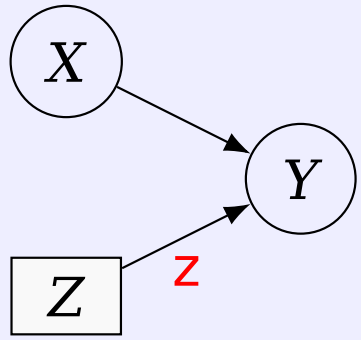
$$f_X^*(x) = \mathbb{1}_{\{x\}}(x) \quad x \in \mathbb{R}.$$

Interventions

Note

In the earlier example





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

Graphical representation of bias

We aim to understand the causal effect of

- treatment Z
- outcome Y,

We study the paths from Z to Y in the DAG describing the joint distribution of all variables. The paths maybe

- directed or undirected
- open or closed.

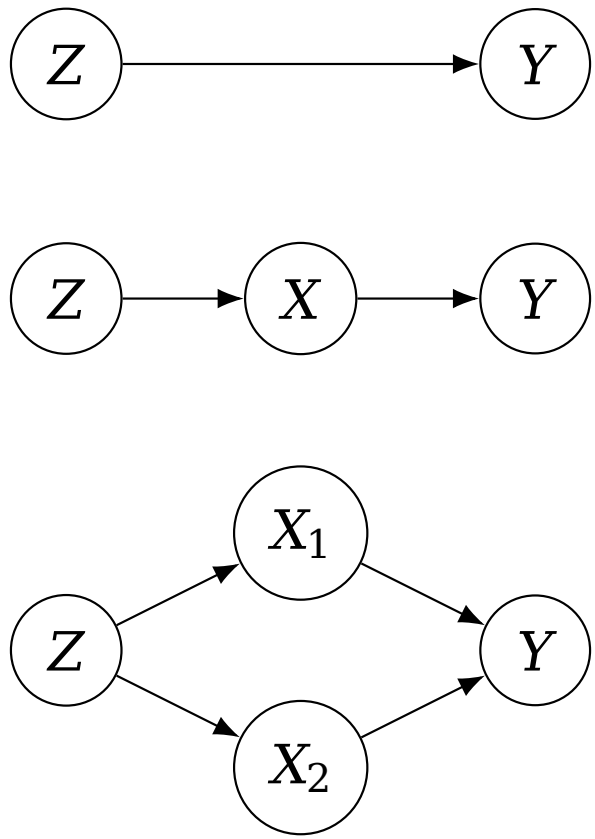
Graphical representation of bias

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 type of path is known as a *biasing* path.
- by *association*, we typically mean some form of *correlation* (or *partial correlation*).

Graphical representation of bias

The *association* between Z and Y is *unconditionally unbiased* (or *marginally unbiased*) for the effect of Z on Y if the only *open* paths between them are *directed* paths.



Graphical representation of bias

Variables S are *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.

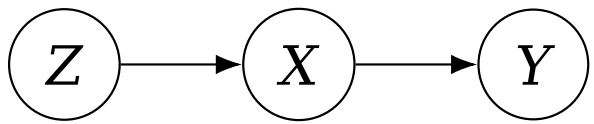
- S is *minimally* sufficient if no proper subset of S is sufficient.

The set of parents of nodes in S is always sufficient, but may not be minimally sufficient.

Conditioning on descendants of Z

Conditioning on descendants of Z

(i) *blocks* directed paths



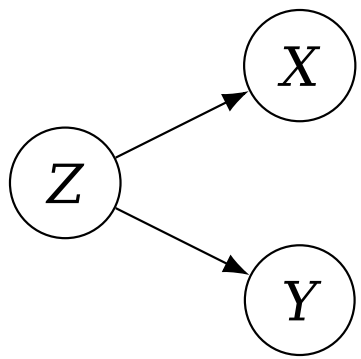
$Z \perp\!\!\!\perp Y \mid X$ but $Z \not\perp\!\!\!\perp Y$

Conditioning on descendants of Z

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

- collider case, selection bias case.

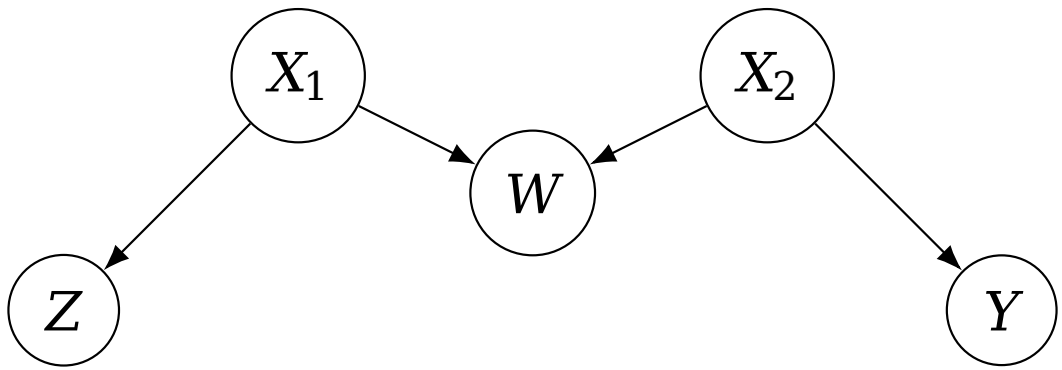
(iii) may be unnecessary in statistical terms: for example



Conditioning on X will not affect bias.

The backdoor criterion

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.

The backdoor criterion

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*.

The backdoor criterion

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

(i) S contains no descendant of Z, and

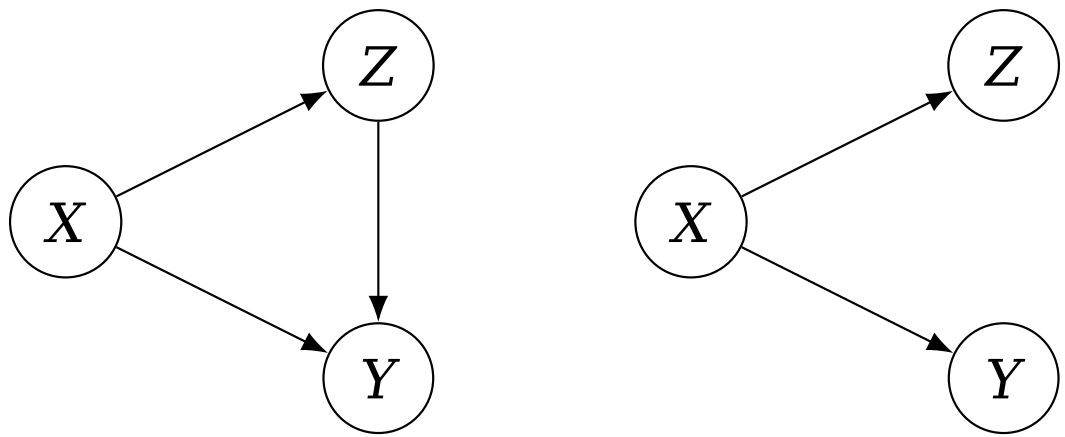
(ii) there is no open backdoor path from Z to Y after conditioning on S.

Conditioning on S allows identification of the causal effect of Z on Y.

Confounding

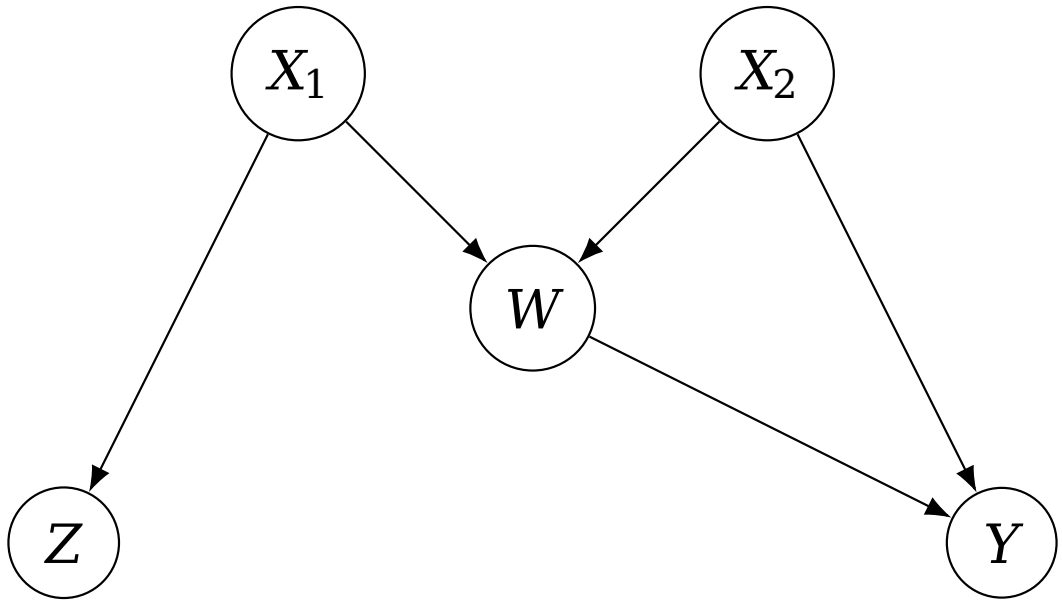
A *confounding path* between Z and Y is a biasing path (that is, an undirected open path) that ends with an arrow into Y.

Variables on a confounding path are termed *confounders*.



X is a confounder in both cases.

Confounding



W is a collider on the path from Z to Y

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

and hence this path is blocked.

Confounding

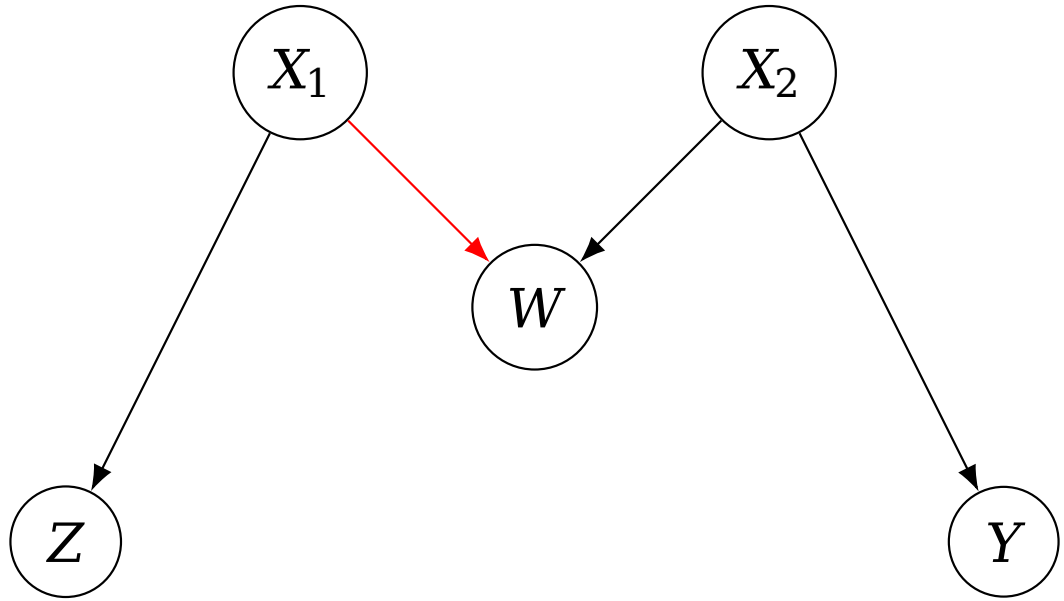
However unconditional on W, the effect of Z on Y is confounded by the backdoor path,

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

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

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

Confounding

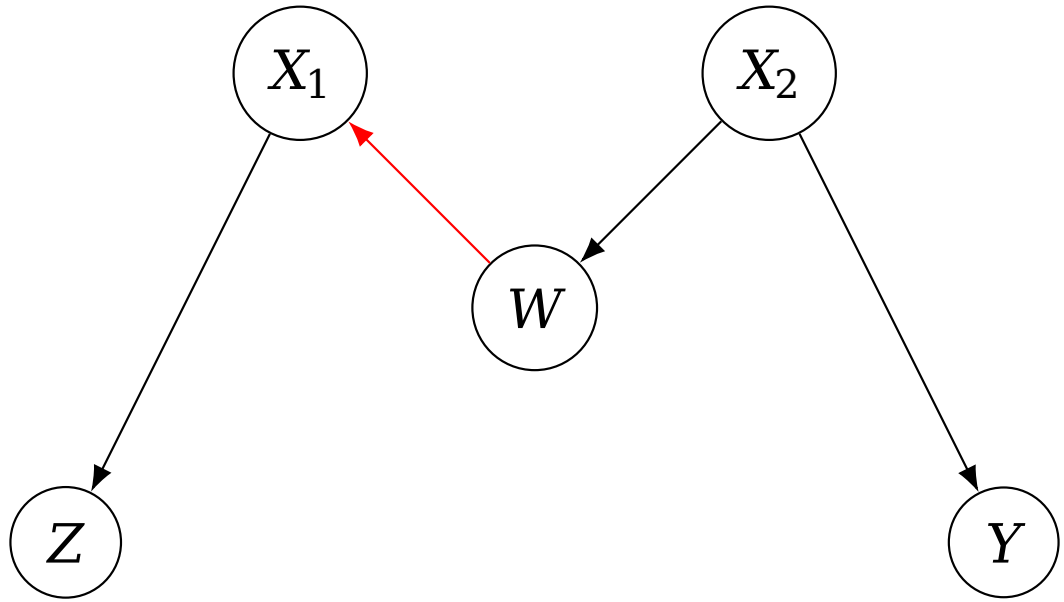


Conditioning on W *opens* the confounding path. Therefore $Z \perp\!\!\!\perp Y$ (as there is no open path between them), but

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

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.

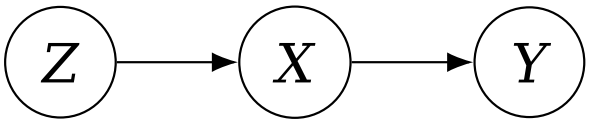
Direct and indirect effects

- *Direct effect:* A direct effect of Z on Y (relative to X) 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.

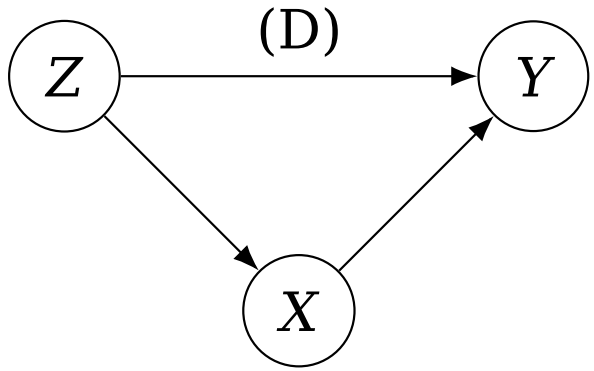
In this formulation, X is termed an *intermediate* or *mediator* variable.

Note that X may be ignored as a mediator, and merely treated as a third variable.

Direct and indirect effects



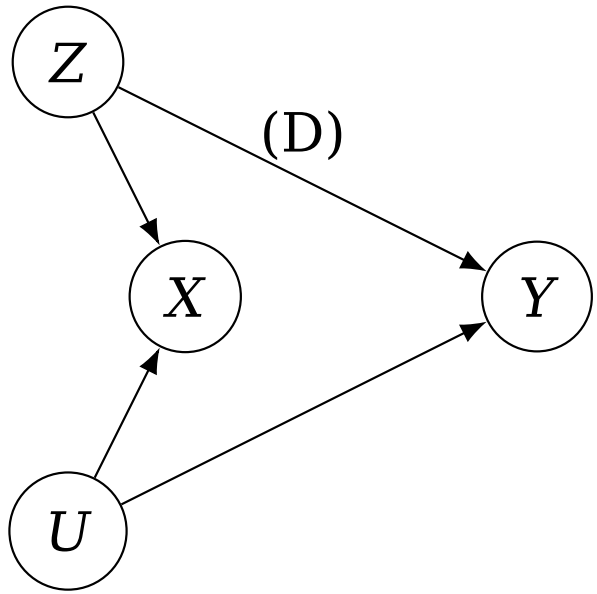
Indirect effect



Direct (D) & Indirect effect

Direct effect is confounded

Direct and indirect effects



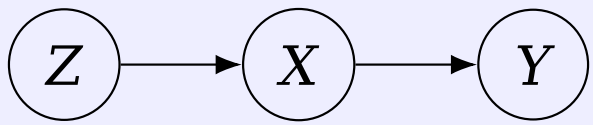
No indirect effect

Direct effect is not confounded

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

Direct and indirect effects

Note

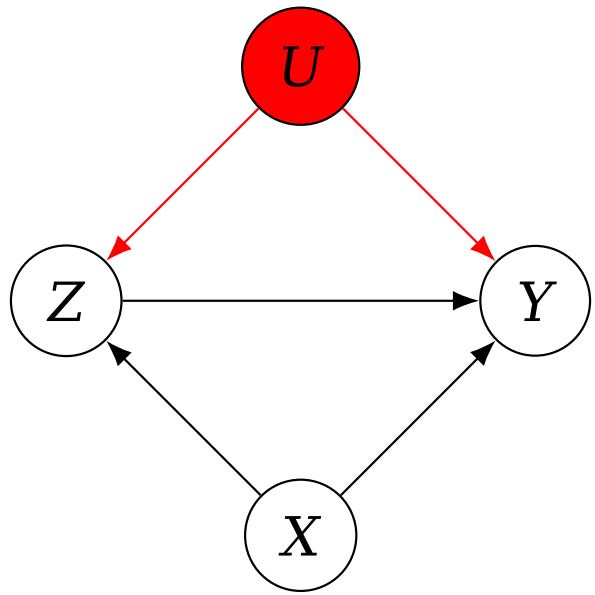


If we ignore X as a mediator:

- *controlled* direct effect: consider $X = x$ held constant.
- *natural* direct effect: consider $Z = z$ held constant, with X taking multiple values.

Unmeasured confounding

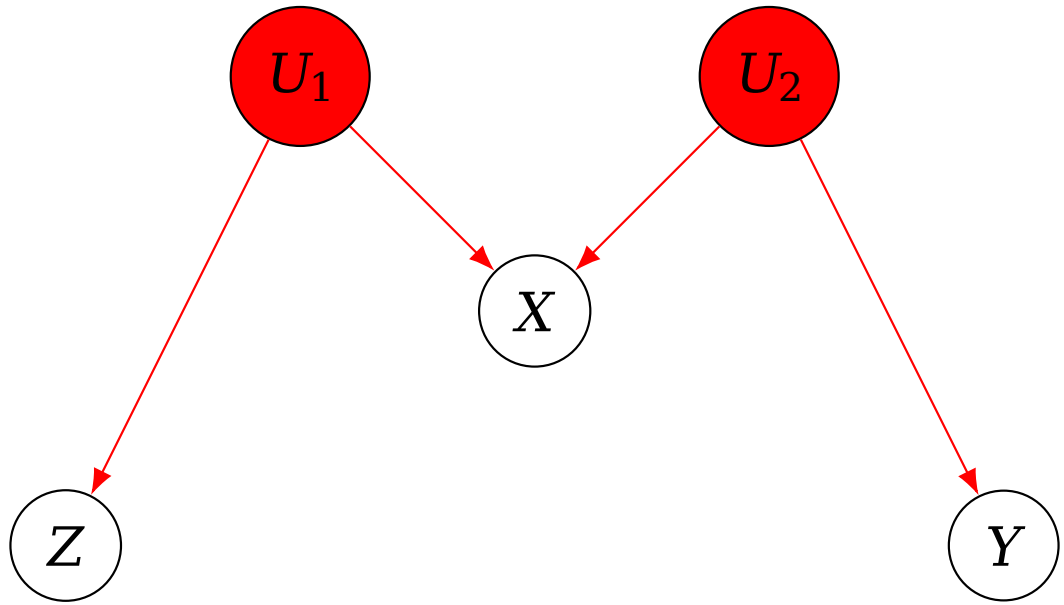
Suppose that in reality there is a further variable U that is a confounder, but is unmeasured in the observed data.



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

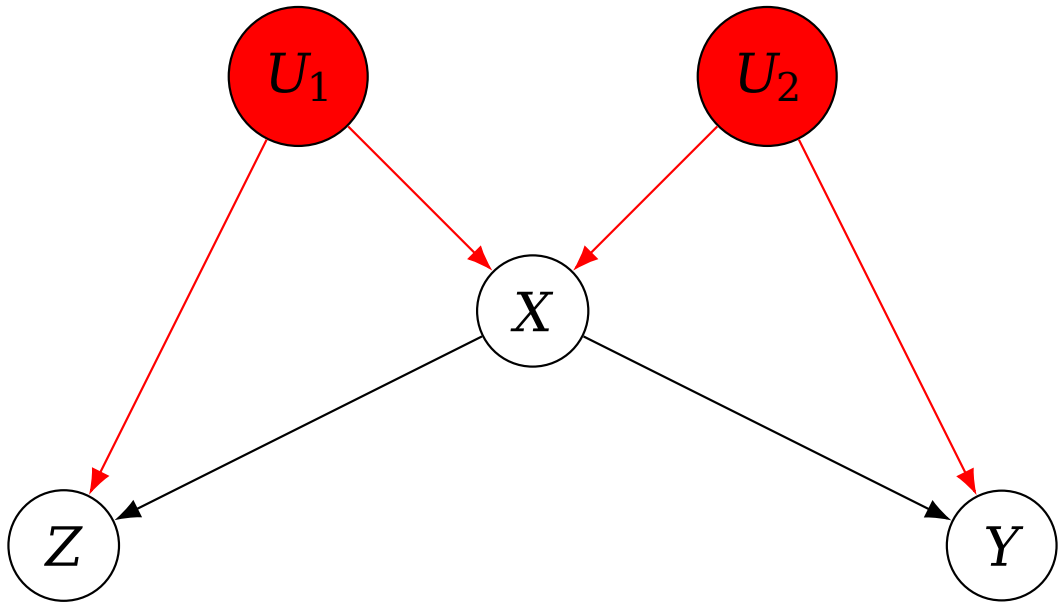
With two unmeasured confounders:



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

However with the same unmeasured confounders:



In the modelled DAG, $Y \perp\!\!\!\perp Z \mid X$; however, conditioning on X *opens* the *hidden* path.

Part 3
Causal Effects

Causal Effects

The *causal effect* of variable Z on variable Y is the amount to which an *intervention* to change Z modifies some aspect

- expected value
- quantile
- distribution

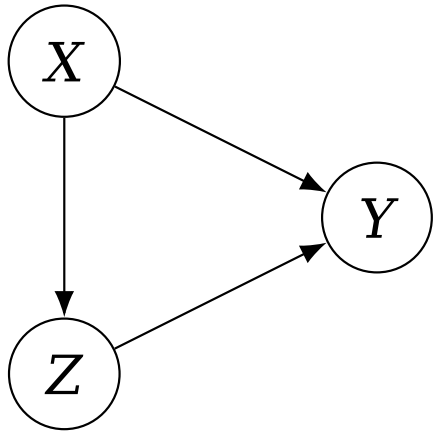
of Y.

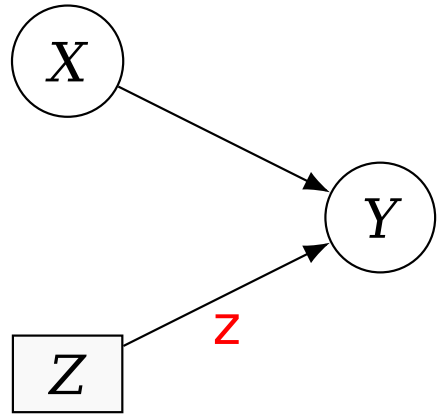
The intervention changes Z from z_0 to z_1, say; in the earlier notation, we consider the intervention model

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

evaluated for different values of z.

Causal Effects





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

The effect of intervening on Z is to *remove all inbound arrows* to node Z, and to fix the value of Z to z. We then consider features pertaining to

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

Causal Effects

Note

Usually we reserve the term ‘causal effect’ for cases where the treatment

- is a *single* specific quantity;
- is genuinely *manipulable* by intervention;
- *precedes the outcome* temporally.

Potential outcome notation

The *counterfactual* or *potential outcome* notation is widely used to formulate causal inference questions.

Let $Y(z)$ denote the random variable recording the (*potential* or *counterfactual*) outcome Y that would be observed if there is an intervention to set $Z = z$.

Potential outcome notation

For any individual, there is therefore a family of potential outcomes

$$\{Y(z) : z \in Z \subseteq \mathcal{Z}\}.$$

For example, if $Z \in \{0, 1\} \equiv Z$, then

$$Y(0): \text{ outcome if intervention sets } z = 0$$

$$Y(1): \text{ outcome if intervention sets } z = 1$$

In practice, it is sufficient to consider a *countable* set Z.

Potential outcome notation

In most cases, the intervention is a *hypothetical* one, and we utilize data arising from a study where Z is observed as part of some stochastic mechanism.

It is reasonable (in fact, necessary) to assume that for observed outcome Y and observed treatment Z, we have

$$Y = \sum_{z \in \mathcal{Z}} Y(z) \mathbb{1}_{\{z\}}(Z)$$

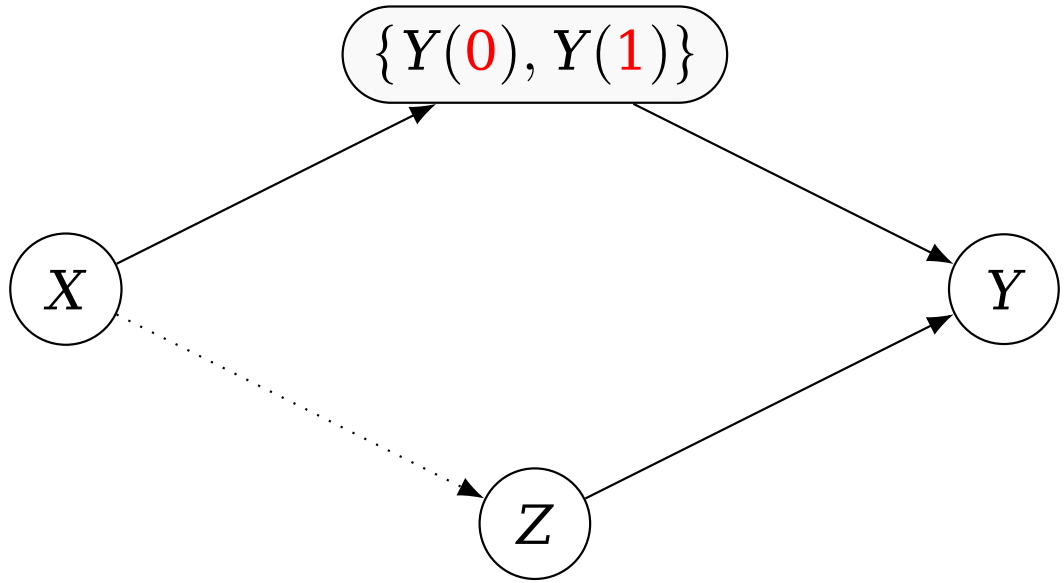
(with probability 1). That is, the observed outcome is identical to the potential outcome arising from the counterfactual treatment that matches the observed treatment.

Potential outcome notation

In the binary case, we may write

$$Y = (1 - Z)Y(0) + ZY(1)$$

and assume *strong ignorability*



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

Potential outcome notation

Note

- Y and $Y(z)$ are different random variables.
- We will typically only observe a single treatment for each individual, so only one of the potential outcomes will be observed.
- Modelling the joint distribution of $Y(z)$ for multiple values of z for a single individual will be challenging.
- Cannot have *unmeasured confounding*, that is, X must be an exhaustive list of confounders.

Potential outcome notation

We consider the data being generated as follows:

- an individual is selected at random, and brings their characteristics $X \sim f_X$;
- the effect of treatment on this individual is to be modelled; the distribution of each potential outcome $Y(z)$, conditional on X, is considered;
- as soon as treatment is assigned/observed, the relevant counterfactual distribution is selected.

Potential outcome notation

Consider the distribution

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

According to earlier assumptions, we should have that

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

that is, conditional on X, the effect on Y of the intervention to set $Z = z$ is the same as if Z were stochastically assigned.

Potential outcome notation

Then marginally, for each z

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

to yield

$$f_{Y(0)}(y) \qquad f_{Y(1)}(y)$$

say.

Potential outcome notation

To describe the causal effect, we may consider *causal contrasts*

- $Y(1) - Y(0)$
- $Y(z_1) - Y(z_0)$
- $\log Y(1) - \log Y(0)$

and so on.

These quantities are random variables, so it is more common to express the causal effect through summaries of $f_{Y(z)}(y)$

- Moments: $\mathbb{E}[Y(z)]$
- Quantiles

Experimental studies

In an *experimental study* precisely the right kind of 'intervention' to study causal contrast is made.

A simple form of experimental study proceeds as follows: $2n$ individuals are sampled from a homogeneous population.

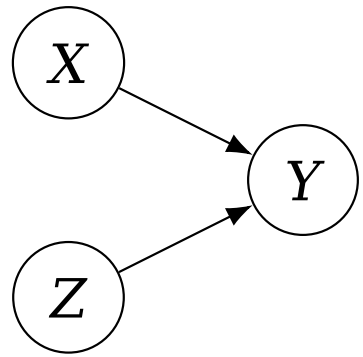
- n are assigned treatment $Z = 0$;
- n are assigned treatment $Z = 1$

irrespective of the individual characteristics of the subjects; the assignment of Z is *independent* of X.

Finally, the outcome Y is measured for each of the $2n$ subjects.

Experimental studies

Here, for each half of the study, the data generating mechanism is as follows.



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

Here we do not need to distinguish the data generating mechanism from the hypothetical intervention, by independence.

Experimental studies

We may also consider a *randomized* version of this study; for each of the individuals in the study, we assign $Z = z$, independently of X, according to the distribution

$$f_Z(z) = p^z(1-p)^{1-z} \quad z \in \{0, 1\}.$$

that is, an individual receives

- $Z = 0$ with probability $1 - p$
- $Z = 1$ with probability p

for some $0 < p < 1$.

Experimental studies

Let N_1 denote the total number of individuals for whom $Z = 1$

$$N_1 = \sum_{i=1}^{2n} \mathbb{1}_{\{1\}}(Z_i)$$

so that

$$N_1 \sim \mathit{Binomial}(2n, p).$$

In the study, we observe $N_1 = n_1$.

Experimental studies

In both non-randomized and randomized studies, the same distributional factorization pertains.

We will denote this distribution

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

with superscript $\mathcal{E}$ indicating the *experimental* assumption.

Experimental studies

Suppose we wish to estimate the difference in outcome (on average) between those individuals assigned $Z = 1$ and those assigned $Z = 0$.

The causal contrast of interest is then

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

which is known as the *average treatment effect* (ATE). The quantity

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

is termed the *average potential outcome* (APO).

Experimental studies

In terms of the above distributions, we may write

$$\mathbb{E}[Y(z)] \equiv \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \int y \; f^{\mathcal{E}}_{Y|Z}(y|z) \; dy$$

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

by independence of X and Z. Multiplying top and bottom by $f^{\mathcal{E}}_{Z}(z)$, we have

$$\mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \frac{\iint y \; f^{\mathcal{E}}_{Y|X,Z}(y|x,z) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x) \; dy \; dx}{f^{\mathcal{E}}_{Z}(z)}$$

Experimental studies

We may write the double integral as a triple integral: with a slight abuse of notation,

$$\frac{\iiint \mathbb{1}_{\{z\}}(z) y \; f^{\mathcal{E}}_{Y|X,Z}(y|x,z) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x) \; dy \; dx \; dz}{\iiint \mathbb{1}_{\{z\}}(z) f^{\mathcal{E}}_{Y|X,Z}(y|x,z) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x) \; dy \; dx \; dz}$$

Thus

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

Experimental studies

Now using the sample data, we may use moment-based estimation to estimate numerator and denominator:

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{X,Y,Z}[\mathbb{1}_{\{z\}}(Z)Y] = \frac{1}{2n}\sum_{i=1}^{2n}\mathbb{1}_{\{z\}}(Z_i)Y_i$$

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{X,Y,Z}[\mathbb{1}_{\{z\}}(Z)] = \frac{1}{2n}\sum_{i=1}^{2n}\mathbb{1}_{\{z\}}(Z_i)$$

and hence estimate the ratio by

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|Z}[Y|Z=z] = \frac{\sum_{i=1}^{2n}\mathbb{1}_{\{z\}}(Z_i)Y_i}{\sum_{i=1}^{2n}\mathbb{1}_{\{z\}}(Z_i)}. \tag{1}$$

Experimental studies

This follows as we actually have a sample from

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

In this estimator:

- numerator is merely the sum of the Y_is for all those individuals who received treatment $Z = z$;
- denominator is merely the number of individuals who received treatment $Z = z$.

Thus, we are merely estimating the quantity $\mathbb{E}[Y(z)]$ by taking the *sample mean* in the group for which $Z = z$.

Experimental studies

For the binary case, the formulae simplify

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|Z}[Y|Z=1] = \frac{\sum_{i=1}^{2n} Z_i Y_i}{\sum_{i=1}^{2n} Z_i} = \frac{1}{N_1}\sum_{i=1}^{2n} Z_i Y_i$$

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|Z}[Y|Z=0] = \frac{\sum_{i=1}^{2n} (1-Z_i) Y_i}{\sum_{i=1}^{2n} (1-Z_i)} = \frac{1}{N_0}\sum_{i=1}^{2n} (1-Z_i) Y_i.$$

Experimental studies

Note that we know that

$$f_Z^{\mathcal{E}}(z) = p^z(1-p)^{1-z}$$

so we can consider the alternative estimator

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

That is

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

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

and the estimator of the ATE is

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

Experimental studies

In a randomized study of size n, the estimator from (1) may be written

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i) Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i)} \equiv \sum_{i=1}^{n} W_i(z) Y_i$$

where

$$W_i(z) = \frac{\mathbb{1}_{\{z\}}(Z_i)}{\sum_{j=1}^{n} \mathbb{1}_{\{z\}}(Z_j)}.$$

Experimental studies

Note that

$$\mathbb{E}_Z^{\mathcal{E}}[W_i(z)] = \frac{1}{n}$$

and

$$0 \leqslant W_i(z) \leqslant 1 \qquad \sum_{i=1}^{n} W_i(z) = 1.$$

Experimental studies

Note also that for this estimator we can define

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{z\}}(Z_i) = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

so that

$$\hat{\mathbb{E}}^{\mathcal{E}}_{Y|Z}[Y|Z=1] = \frac{1}{n\hat{p}}\sum_{i=1}^{n} Z_i Y_i$$

$$\hat{\mathbb{E}}^{\mathcal{E}}_{Y|Z}[Y|Z=0] = \frac{1}{n(1-\hat{p})}\sum_{i=1}^{n} (1-Z_i) Y_i$$

Experimental studies

Both estimators (1) and (2) are *unbiased* for the APO.

$$(1)\quad \widehat{\mu}_n(1) = \frac{1}{n\widehat{p}} \sum_{i=1}^{n} Z_i Y_i \qquad \widehat{\mu}_n(0) = \frac{1}{n(1-\widehat{p})} \sum_{i=1}^{n} (1-Z_i) Y_i$$

$$(2)\quad \widetilde{\mu}_n(1) = \frac{1}{np} \sum_{i=1}^{n} Z_i Y_i \qquad \widetilde{\mu}_n(0) = \frac{1}{n(1-p)} \sum_{i=1}^{n} (1-Z_i) Y_i$$

This results in estimators that are unbiased for the ATE:

$$\widehat{\delta}_n = \frac{1}{n\widehat{p}} \sum_{i=1}^{n} Z_i Y_i - \frac{1}{n(1-\widehat{p})} \sum_{i=1}^{n} (1 - Z_i) Y_i = \frac{1}{n} \sum_{i=1}^{n} \frac{(Z_i - \widehat{p})}{\widehat{p}(1-\widehat{p})} Y_i$$

$$\widetilde{\delta}_n = \frac{1}{np} \sum_{i=1}^{n} Z_i Y_i - \frac{1}{n(1-p)} \sum_{i=1}^{n} (1 - Z_i) Y_i = \frac{1}{n} \sum_{i=1}^{n} \frac{(Z_i - p)}{p(1-p)} Y_i$$

The only difference between the estimators is whether we use $\widehat{p}$ or p to represent the treatment probability.

Experimental studies

It transpires that

$$\lim_{n \longrightarrow \infty} n\text{Var}[\widehat{\delta}_n] < \lim_{n \longrightarrow \infty} n\text{Var}[\widetilde{\delta}_n]$$

that is, estimator $\widehat{\delta}_n$ is (asymptotically) more *efficient*.

That is, *it is better to estimate* p rather than use its known value.

Observational studies

In an *observational study* we do not intervene to assign treatment to subjects, we observe it as part of the data collection process.

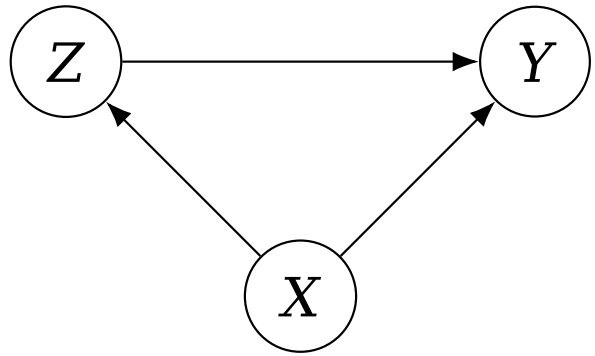
We denote the data generating mechanism

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

In the observational setting, there may be several possible proposed data generating mechanisms, but critically we may consider 'causes' of treatment Z.

Observational studies

A DAG of interest involves a backdoor path from Z to Y



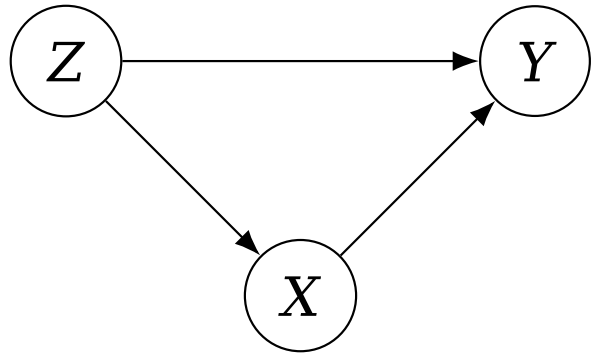
$$f_X^{\mathcal{O}}(x) f_{Z|X}^{\mathcal{O}}(z|x) f_{Y|X,Z}^{\mathcal{O}}(y|x,z)$$

There is an open backdoor (confounding) path $Z \to X \to Y$ and therefore there is a possibility of bias.

To get at the causal effect of Z on Y, we must block the backdoor path by *conditioning* on X.

Observational studies

Note that we might have the following DAG:



$$f_Z^{\mathcal{O}}(z) f_{X|Z}^{\mathcal{O}}(x|z) f_{Y|X,Z}^{\mathcal{O}}(y|x,z)$$

There are two paths from Z to Y: the direct path, and the path $Z \to X \to Y$, which is again blocked by *conditioning* on X.

Here X is a mediator on the indirect path; we might be interested in both the direct and indirect effects of Z on Y.

Observational studies

Suppose we try to estimate the causal effect of Z on Y in the confounding case. First, consider

$$\mathbb{E}^{\mathcal{O}}_{Y|Z}[Y|Z = \mathsf{z}]$$

which would be the equivalent of the earlier causal quantity (the APO at treatment z); however, here note that, in the observed data, $Z = \mathsf{z}$ is not achieved by intervention as in the experimental case.

Observational studies

We have, as in the earlier calculation

$$\mathbb{E}^{\mathcal{O}}_{Y|Z}[Y|Z = \mathsf{z}] = \iint y \; f^{\mathcal{O}}_{Y|X,Z}(y|x,\mathsf{z}) f^{\mathcal{O}}_{X|Z}(x|\mathsf{z}) \; dy \; dx$$

$$= \frac{\iiint \mathbb{1}_{\{\mathsf{z}\}}(z) y \; f^{\mathcal{O}}_{Y|X,Z}(y|x,z) f^{\mathcal{O}}_{X|Z}(x|z) f^{\mathcal{O}}_{Z}(z) dy \; dx \; dz}{\iiint \mathbb{1}_{\{\mathsf{z}\}}(z) \; f^{\mathcal{O}}_{Y|X,Z}(y|x,z) f^{\mathcal{O}}_{X|Z}(x|z) f^{\mathcal{O}}_{Z}(z) dy \; dx \; dz}$$

where again the indicator function $\mathbb{1}_{\{\mathsf{z}\}}(z)$ reduces the contribution to the dz integrals to the point evaluation at $z = \mathsf{z}$.

Hence, as before, we have

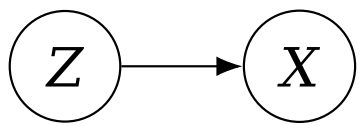
$$\mathbb{E}^{\mathcal{O}}_{Y|Z}[Y|Z = \mathsf{z}] = \frac{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[\mathbb{1}_{\{\mathsf{z}\}}(Z)Y]}{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[\mathbb{1}_{\{\mathsf{z}\}}(Z)]}.$$

Observational studies

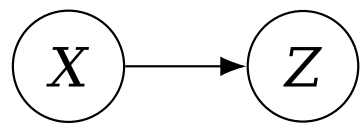
Note that by the chain rule factorization, we must have

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

This result re-iterates that a DAG does not define a unique *joint* distribution:



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



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

Observational studies

Thus $\mathbb{E}^{\mathcal{O}}_{Y|Z}[Y|Z = z]$ can be rewritten

$$\frac{\iiint \mathbb{1}_{\{z\}}(z) y \; f^{\mathcal{O}}_{Y|X,Z}(y|x,z) f^{\mathcal{O}}_{Z|X}(z|x) f^{\mathcal{O}}_{X}(x) dy \; dx \; dz}{\iiint \mathbb{1}_{\{z\}}(z) \; f^{\mathcal{O}}_{Y|X,Z}(y|x,z) f^{\mathcal{O}}_{Z|X}(z|x) f^{\mathcal{O}}_{X}(x) dy \; dx \; dz}$$

This can be contrasted with the earlier formula for the APO

$$\mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \frac{\iiint \mathbb{1}_{\{z\}}(z) y \; f^{\mathcal{E}}_{Y|X,Z}(y|x,z) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x) dy \; dx \; dz}{\iiint \mathbb{1}_{\{z\}}(z) f^{\mathcal{E}}_{Y|X,Z}(y|x,z) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x) \; dy \; dx \; dz}$$

Observational studies

Now we can legitimately assume

$$f_X^{\mathcal{O}}(x) \equiv f_X^{\mathcal{E}}(x)$$

as this distribution describes the population characteristics.

Also, with the proposed data generating distribution given by the confounding DAG, we have

$$f_{Y|X,Z}^{\mathcal{O}}(y|x,z) \equiv f_{Y|X,Z}^{\mathcal{E}}(y|x,z).$$

Observational studies

However, in general

$$f^{\mathcal{O}}_{Z|X}(z|\mathbf{x}) \neq f^{\mathcal{E}}_{Z}(z) \quad \forall(\mathbf{x}, z)$$

and so evidently

$$\mathbb{E}^{\mathcal{O}}_{Y|Z}[Y|Z = z] \neq \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z].$$

Observational studies

Thus, if we consider using moment-based estimation

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

$$\widehat{\mathbb{E}}^{\mathcal{O}}_{X,Y,Z}[\mathbb{1}_{\{z\}}(Z)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{z\}}(Z_i)$$

and then estimate the ratio by

$$\widehat{\mathbb{E}}^{\mathcal{O}}_{Y|Z}[Y|Z=z] = \frac{\sum_{i=1}^{n}\mathbb{1}_{\{z\}}(Z_i)Y_i}{\sum_{i=1}^{n}\mathbb{1}_{\{z\}}(Z_i)}. \tag{3}$$

the estimator will in general be biased for the APO.

Observational studies

The bias arises as $Z \not\perp\!\!\!\perp X$;

- this implies that, in the sample data, we cannot treat Z as if it were assigned independently of X;
- different observed values of Z will (in general) have different associated distributions of X, as $f_{X|Z}(x|z)$ changes as z changes;

Observational studies

- subpopulations identified by different values of z are *not comparable*; the characteristics of individuals in different subpopulations are different;
- if X also affects Y, we cannot simply compare the outcomes for different observed Z values, as individuals with different Z values have different X characteristics.

Model-based estimation

If we *know* that

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

say, then as $f^{\mathcal{E}}_X(x) = f^{\mathcal{O}}_X(x)$ it follows by iterated expectation that

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

where

$$\begin{aligned}\mathbb{E}^{\mathcal{O}}_{Y|Z}[Y|Z = z] &= \iint y\ f^{\mathcal{O}}_{Y|X,Z}(y|x,z) f^{\mathcal{O}}_X(x)\ dy\ dx \\ &= \int \mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X,Z=z] f^{\mathcal{O}}_X(x)\ dx.\end{aligned}$$

Model-based estimation

Then a moment-based estimator of the APO is

$$\widehat{\mu}_{\text{OR}}(z) \equiv \widehat{\mathbb{E}}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \frac{1}{n}\sum_{i=1}^{n} \mu(X_i, z).$$

and in the binary case, the corresponding estimator of the ATE is

$$\widehat{\delta}_{\text{OR}} = \widehat{\mu}_{\text{OR}}(1) - \widehat{\mu}_{\text{OR}}(0) = \frac{1}{n}\sum_{i=1}^{n} \{\mu(X_i, 1) - \mu(X_i, 0)\}.$$

The subscript OR indicates *outcome regression*.

Model-based estimation

This approach is termed a *model-based* analysis; note that it requires correct specification of the $\mu(x, z)$ function; if we mistakenly assume

$$\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[Y|X, Z] = \mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X, Z] = m(X, Z)$$

then the resulting estimators, for example

$$\widehat{\delta}_{\text{OR}} = \frac{1}{n} \sum_{i=1}^{n} \{m(X_i, 1) - m(X_i, 0)\}.$$

are, in general, biased.

Model-based estimation

Note that we can afford some mis-specification: for example, if Z is binary, we can always write

$$\mu(x, z) = \mu_0(x) + z\mu_1(x) \qquad \text{TRUE}$$

$$m(x, z) = m_0(x) + zm_1(x) \qquad \text{MODEL}$$

in which case the estimator

$$\frac{1}{n}\sum_{i=1}^{n} m_1(X_i)$$

is unbiased for the ATE provided

$$\mathbb{E}_X^{\mathcal{O}}[m_1(X)] = \mathbb{E}_X^{\mathcal{O}}[\mu_1(X)] \equiv \mathbb{E}_X^{\mathcal{E}}[\mu_1(X)].$$

That is, we can *mis-specify* $m_0(x)$.

Model-based estimation

In the binary case, if

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

then we may consider an alternate estimator

$$\widehat{\delta}^{*}_{\text{OR}} = \frac{1}{n\widehat{p}}\sum_{i=1}^{n} Z_i \mu(X_i, Z_i) - \frac{1}{n(1-\widehat{p})}\sum_{i=1}^{n}(1-Z_i)\mu(X_i, Z_i)$$

which estimates the mean separately in the two subgroups defined by $Z = 1$ and $Z = 0$ separately. That is,

$$\widehat{\delta}^{*}_{\text{OR}} = \frac{1}{n}\sum_{i=1}^{n} \frac{(Z_i - \widehat{p})}{\widehat{p}(1-\widehat{p})}\mu(X_i, Z_i)$$

Model-based estimation

Note

In the model-based approach, we must have

$$\mu(\mathbf{x}, z)$$

specified precisely. In practice, however, we will propose *parametric* models, for example

$$\mu(\mathbf{x}, z; \beta, \psi) = \mu_0(\mathbf{x}; \beta) + z\mu_1(\mathbf{x}; \psi)$$

and then hope to estimate (β, ψ) from the observed data.

In general, this parametric model must be *completely* correctly specified for consistent estimation of the ATE.

Model-based estimation

Note

For example, in the linear model case with

$$\mu(\mathbf{x}, z; \beta, \psi) = \mathbf{x}_\beta \beta + z\, \mathbf{x}_\psi \psi$$

we may only consistently estimate β and ψ, and hence the ATE, if this mean model is correctly specified.

Model-based estimation

Note

It is no longer sufficient to specify

$$\mu_1(\mathbf{x};\beta) = \mathbf{x}_\psi \psi$$

correctly as the *'treatment contributed'* expected response, correct specification of

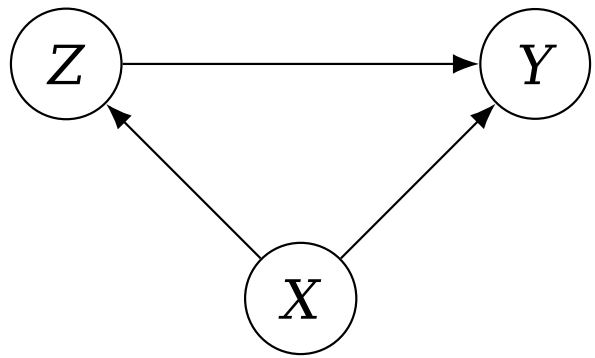
$$\mu_0(x;\beta) = \mathbf{x}_\beta \beta$$

as the *'treatment free'* expected response is also necessary.

Part 4
Causal Adjustment Methods

Causal Adjustment Methods

We have seen that regression methods can recover causal quantities of interest in observational studies, provided there is *correct specification* of the regression model, even if there is a *confounding* of the direct effect.



Model-based estimation will be successful if

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X,Z]$$

is correctly specified.

Causal Adjustment Methods

Recall that, under the previous assumptions

$$\mu(z) = \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \iint y \; f^{\mathcal{E}}_{Y|X,Z}(y|x, z) f^{\mathcal{E}}_{X}(x) \; dy \; dx$$

$$\equiv \iint y \; f^{\mathcal{O}}_{Y|X,Z}(y|x, z) f^{\mathcal{O}}_{X}(x) \; dy \; dx$$

$$\equiv \int \mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X = x, Z = z] f^{\mathcal{O}}_{X}(x) \; dx$$

that is, the treatment data are ignored.

This is known as a *G-computation* formula; it yields estimates of APO $\mu(z)$ under correct specification of

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X = x, Z = z]$$

Causal Adjustment Methods

If correct specification cannot be guaranteed, we must seek other adjustment approaches.

The key complication that prevents use of the observational data is that

$$Z \not\perp\!\!\!\perp X$$

so that the treatment-indexed subgroups are *incomparable* due to their different X characteristics.

How can we break the dependence ?

Matching

Suppose first that the confounder X is *degenerate* at $x = x_0$, that is

$$\Pr[X = x_0] = 1.$$

Then (trivially)

$$f^{\mathcal{O}}_{Z|X}(z|x) \equiv f^{\mathcal{O}}_{Z}(z) \quad \forall z,\ x = x_0$$

and

$$\mathbb{E}^{\mathcal{O}}_{Y|Z}[Y|Z = z] = \int y\ f^{\mathcal{O}}_{Y|X,Z}(y|x_0, z)\ dy = \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z].$$

Matching

Now suppose X takes values on the finite set

$$\mathcal{X} = \{x_1, \ldots, x_J\}$$

with

$$f_X^{\mathcal{O}}(x) = f_X^{\mathcal{E}}(x)$$

determining the distribution of X.

Matching

Define the *'local'* APO at $x = x_j$, for $j = 0, 1, 2, \ldots$, by

$$\mathbb{E}^{\mathcal{E},j}_{Y|Z}[Y|Z = z] = \int y \; f^{\mathcal{O}}_{Y|X,Z}(y|x_j, z) \; dy = \mu_j(z)$$

say. We may estimate this quantity using sample-based estimation using the estimator

$$\widehat{\mu}_j(z) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{x_j\}}(X_i)\mathbb{1}_{\{z\}}(Z_i)Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\{x_j\}}(X_i)\mathbb{1}_{\{z\}}(Z_i)}$$

This is the sample mean in the treatment group with $Z = z$ in the population stratum $\mathcal{X}_j$ for which $X = x_j$.

Matching

Within stratum $\mathcal{X}_j$, for the binary treatment case, the *local ATE* is estimated by

$$\widehat{\delta}_{\text{MATCH},j} = \widehat{\mu}_j(1) - \widehat{\mu}_j(0)$$

This is an unbiased estimator of the local ATE, that is, the ATE in the stratum $\mathcal{X}_j$.

Matching

Finally, we can estimate the (global) ATE using a weighted combination of the local estimators.

$$\widehat{\delta}_{\text{MATCH}} = \sum_{j=1}^{J} \widehat{w}_j \ \widehat{\delta}_{\text{MATCH},j}$$

where

$$\widehat{w}_j = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_j\}}(X_i)$$

estimates the probability of observing X in stratum $\mathcal{X}_j$.

Matching

This is a *matching* estimator:

- the local estimators are constructed by *matching* separately on the x_j;
- in the matched stratum, the only difference between individuals is their treatment status;
- in a matched subsample, we can *directly* compare the outcomes for the treatment-indexed subgroups.
- we combine the local estimators into a global estimator.

Matching

Note

The local estimators rely on having a large enough subsample size in the stratum $\mathcal{X}_j$, for all the targeted treatment values, to allow the estimators to exhibit good behaviour.

The matching approach can also be applied if X is vector-valued; however, again the subsample size can deplete as the dimension of X increases.

Matching

If X is *continuous*, then exact matching cannot be used. However, we can define similar strata

$$\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_J$$

that form a partition of $\mathcal{X}$, and then assume that

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

for $j = 1, 2, \ldots, J$, where $f_j(z)$ does not depend on x.

Then, the matching estimator $\widehat{\delta}_{\text{MATCH}}$ can still be used.

Matching

Note

- If X is scalar, we can define the strata by using quantiles of the observed data.
- Defining the strata may not be straightforward when X is vector-valued.
- The assumption that, *within* a stratum, $f^{\mathcal{O}}_{Z|X}(z|x)$ does not depend on x is quite a strong one.

Matching

Model-based matching estimators may also be constructed: for example, could write

$$\mathbb{E}^{\mathcal{O},j}_{Y|Z}[Y|X=\mathrm{x},Z=z]=\mu(\mathrm{x},z;\beta_j,\psi_j)\qquad \mathrm{x}\in\mathcal{X}_j$$

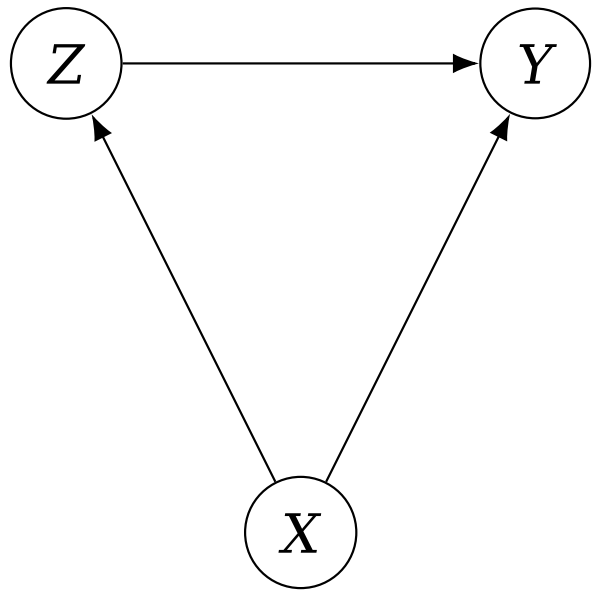
which then leads to an estimator

$$\widehat{\mu}_j(\mathsf{z})=\frac{\sum_{i=1}^{n}\mathbb{1}_{\{\mathrm{x}_j\}}(X_i)\mathbb{1}_{\{\mathsf{z}\}}(Z_i)\mu(X_i,\mathsf{z};\widehat{\beta}_j,\widehat{\psi}_j)}{\sum_{i=1}^{n}\mathbb{1}_{\{\mathrm{x}_j\}}(X_i)\mathbb{1}_{\{\mathsf{z}\}}(Z_i)}$$

after estimating β_j and ψ_j using data within stratum $\mathcal{X}_j$.

Constructing Balance

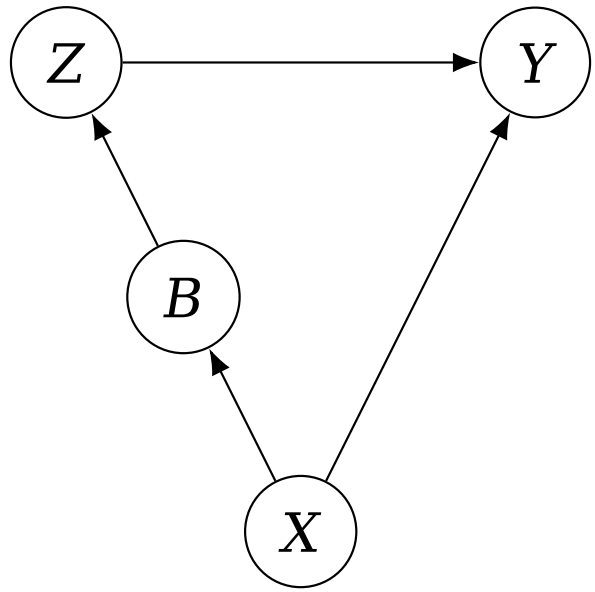
Consider the basic confounding set up:



We know that conditioning on X blocks the confounding path.

Constructing Balance

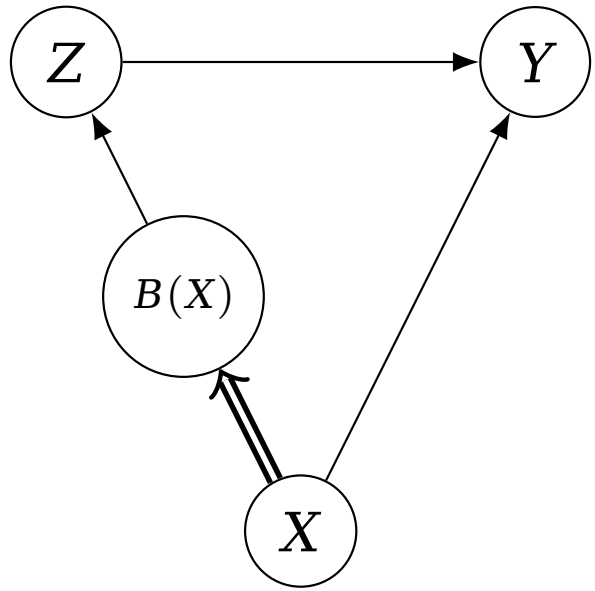
Now suppose we could find a new variable B so that



Conditioning on B also *blocks* the confounding path.

Constructing Balance

We can define B as a *deterministic* function of X: $B \equiv B(X)$



For example, X is vector-valued, B is a scalar summary of X.

Constructing Balance

Note that

- if we can find such a B, then

$$Z \perp\!\!\!\perp X \mid B$$

- there is no arrow from B to Y, and

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

  as conditioning on (X, Z) blocks the paths from B to Y.

Constructing Balance

If we can find such a B, then we can consider performing 'local' analyses for the causal effect within strata of B;

- within a given stratum $\mathcal{B}_j$ of B, Z and X are independent;
- we can directly compare the Ys for different Z values for the subjects falling within $\mathcal{B}_j$, and be sure that the X values for those subjects are suitably *matched*.
- ideally, B can be a *low-dimensional* summary (a bit like a sufficient statistic) even if X is *high-dimensional*.

Constructing Balance

We need to find B such that $Z \perp\!\!\!\perp X \mid B$, that is, for all (x, z, b),

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

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

provided the conditional densities are well-defined.

Note that if $B = B(X)$ *deterministically*, then

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

Constructing Balance

Suppose that $Z \in \{0, 1\}$. Then we *must* have that

$$f_{Z|X}(z|x) \equiv Bernoulli(p(x))$$

where $0 \leqslant p(x) \leqslant 1$ a probability that may depend on x.

That is,

$$\Pr[Z = z|X = x] = \{p(x)\}^z\{1 - p(x)\}^{1-z} \quad z = 0, 1$$

Constructing Balance

Similarly, for any proposed B, we must also have

$$f_{Z|X,B}(z|x,b) \equiv Bernoulli(q(x,b))$$

where $0 \leqslant q(x,b) \leqslant 1$ a probability that may depend on (x,b). That is,

$$\Pr[Z = z|X = x, B = b] = q(x,b)^z(1 - q(x,b))^{1-z} \quad z = 0,1$$

We must now ensure that

$$\Pr[Z = z|X = x, B = b] = \Pr[Z = z|B = b] \qquad \forall(x,z,b).$$

Constructing Balance

We can do this by requiring

$$q(x, b) \equiv q(b)$$

for all (x, b). That is, we must ensure

$$\Pr[Z = z | X = x, B = b] = q(b)^z (1 - q(b))^{1-z} \quad z = 0, 1.$$

Constructing Balance

Define the function

$$B(x) = \Pr[Z = 1 | X = x]$$

with random variable version $B(X) = \Pr[Z = 1 | X]$, and set

$$q(b) = b$$

so that

$$\Pr[Z = z | X = x, B = b] = b^z (1 - b)^{1-z} = \Pr[Z = z | B = b].$$

Constructing Balance

Hence by construction

$$f_{Z|X,B}(z|x,b) = f_{Z|B}(z|b) = b^z(1-b)^{1-z}$$

that is, if we consider the 'contour'

$$\mathcal{X}_b = \{x : B(x) = b\}$$

then

$$\Pr[Z = 1|X = x] = b \qquad \forall x \in \mathcal{X}_b.$$

Thus

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

Constructing Balance

The function $B(x)$ contains all the relevant information extracted from X to determine the conditional distribution of Z given X.

Note that $B(X)$ is a *scalar* random variable, whatever the dimension of X; if $X = (X_1, X_2, X_3)$ say, we might have that

$$f_{Z|X}(1|x) = \Pr[Z = 1|X = x] = \frac{\exp\{x_1 + 2x_2x_3^2\}}{1 + \exp\{x_1 + 2x_2x_3^2\}} \equiv B(x).$$

Many triples (x_1, x_2, x_3) yield the *same* probability.

Constructing Balance

In the binary case, the random variable $B = B(X)$ is known as the *propensity score*; this is denoted

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

by Rosenbaum and Rubin (1983).

Constructing Balance

We have for the joint pdf

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

$$= \begin{cases} f_X(x) b^z (1-b)^{1-z} & b = B(x) \\ 0 & \text{otherwise} \end{cases}$$

as

$$f_{B|X}(b|x) = \begin{cases} 1 & b = B(x) \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_{Z|X,B}(z|x,b) = f_{Z|X}(z|x) = b^z (1-b)^{1-z}.$$

Constructing Balance

Thus if

$$\mathcal{X}_b = \{x : B(x) = b\}$$

then for all (z, b)

$$\begin{aligned} f_{Z,B}(z, b) &= \int_{\mathcal{X}_b} f_{X,Z,B}(x, z, b)\, dx \\ &= \int_{\mathcal{X}_b} b^z (1-b)^{1-z} f_X(x)\, dx \\ &= b^z (1-b)^{1-z} \int_{\mathcal{X}_b} f_X(x)\, dx \end{aligned}$$

Constructing Balance

Therefore

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

$$= \begin{cases} \dfrac{f_X(x)}{\displaystyle\int_{\mathcal{X}_b} f_X(t)\ dt} & x \in \mathcal{X}_b \\ 0 & \text{otherwise} \end{cases}$$

as the b terms cancel. Thus

$$f_{X|Z,B}(x|z,b) \equiv f_{X|B}(x|b).$$

Constructing Balance

Note

Note that for the binary case

$$e(X) = \Pr[Z = 1|X] \equiv \mathbb{E}_{Z|X}[Z|X].$$

Constructing Balance

We also have that B^* is another balancing score *if and only if* B is a function of B^*

- that is, if and only if B is 'coarser' than B^*,

$$B^* = b^* \quad \Longrightarrow \quad B = b\,.$$

Constructing Balance

To see this, suppose first that $B^* = b^*$ implies $B = b$. Then, by iterated expectation

$$\begin{aligned}
\Pr[Z = 1 | B^* = b^*] &= \mathbb{E}_{Z|B^*}[Z | B^* = b^*] \\
&= \mathbb{E}_{B|B^*}\left[\mathbb{E}_{Z|B,B^*}[Z|B, B^* = b^*] \middle| B^* = b^*\right] \\
&= \mathbb{E}_{B|B^*}\left[\mathbb{E}_{Z|B}[Z|B = b] \middle| B^* = b^*\right] \\
&= \mathbb{E}_{Z|B}[Z|B = b] \\
&= \Pr[Z = 1|B = b]
\end{aligned}$$

so therefore B^* is a balancing score, as B is a balancing score.

Constructing Balance

Conversely, suppose B^* is a balancing score, that is

$$\Pr[Z = 1 | X = x, B^*(x) = b^*] = \Pr[Z = 1 | B^*(x) = b^*]$$

Consider two values x_1 and x_2. We have that

$$\Pr[Z = 1 | B^*(x_1) = b^*] = \Pr[Z = 1 | B^*(x_2) = b^*]$$

that is, if both x_1 and x_2 map to b^* under $B^*(.)$, then the two probabilities must be equal by the balancing assumption.

Constructing Balance

As B^* is a balancing score, we have also that

$$\Pr[Z = 1|B^*(x_1) = b^*] = \Pr[Z = 1|X = x_1, B^*(x_1) = b^*]$$

$$\Pr[Z = 1|B^*(x_2) = b^*] = \Pr[Z = 1|X = x_2, B^*(x_2) = b^*]$$

But as for all x

$$\Pr[Z = 1|X = x, B^*(x) = b^*] \equiv \Pr[Z = 1|X = x]$$

this implies that

$$\Pr[Z = 1|X = x_2] = \Pr[Z = 1|X = x_2]$$

and hence that $B(x_1) = B(x_2)$, as required.

Constructing Balance

The balancing construction extends *beyond* the case of binary treatments: suppose Z is *continuous*, and that

$$f_{Z|X}(z|x)$$

is some conditional density for Z given X in the same (observational) model.

Suppose that we have for some function $B = B(X)$

$$f_{Z|X}(z|x) \equiv f_{Z|B}(z|B(x)) \qquad \forall (x, z)$$

Then directly

$$f_{Z|X,B}(z|x, b) \equiv f_{Z|B}(z|b) \qquad \forall (x, z), b = B(x).$$

Constructing Balance

Example:

Suppose

$$Z \mid X_1 = x_1, X_2 = x_2 \sim Normal(x_1 + x_2, \sigma^2).$$

Then define

$$B(x) \equiv B(x_1, x_2) = x_1 + x_2$$

so that

$$Z \mid X_1 = x_1, X_2 = x_2, B = b \sim Normal(b, \sigma^2).$$

which does not depend on (x_1, x_2).

Constructing Balance

For the binary case

$$e(X) = \Pr[Z = 1|X] \equiv \mathbb{E}_{Z|X}[Z|X]$$

which suggests another possible balancing score construction involves inspection of

$$B(X) = \mathbb{E}[Z|X].$$

This will not necessarily yield *independence*, but it may yield (partial) *uncorrelatedness*, that is

$$\text{Cov}[X, Z|B] = 0.$$

Constructing Balance

Note

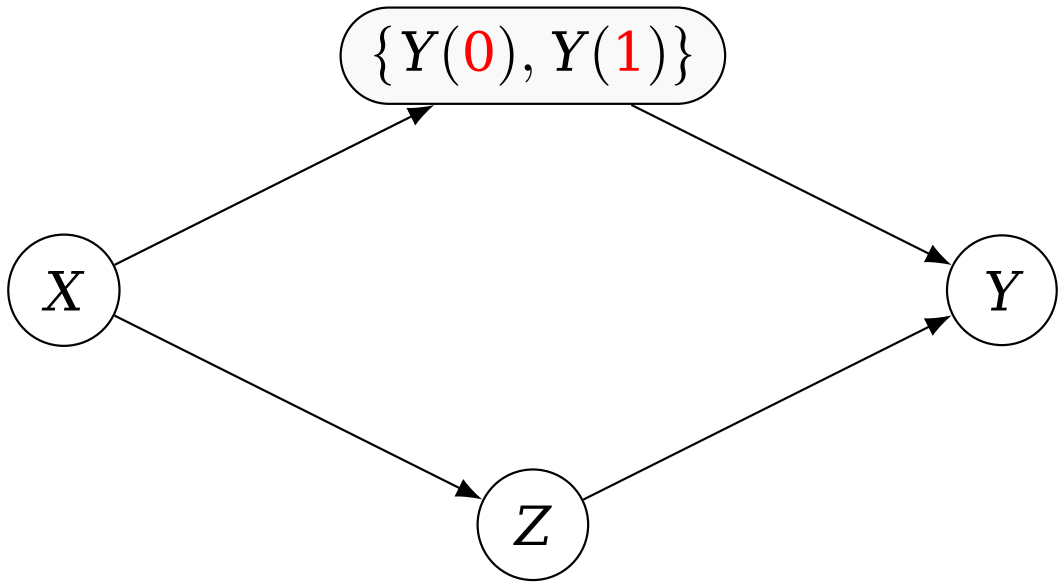
For the moment, we will assume that $e(X)$ or $B(X)$ is *known precisely*.

In practice, we will typically have to

- assume a *parametric* model and rely on correct specification to ensure consistent estimation of the propensity score parameters and values, or
- use *advanced approaches* (machine learning, flexible, adaptive approaches) to obtain the propensity score function.

Constructing Balance

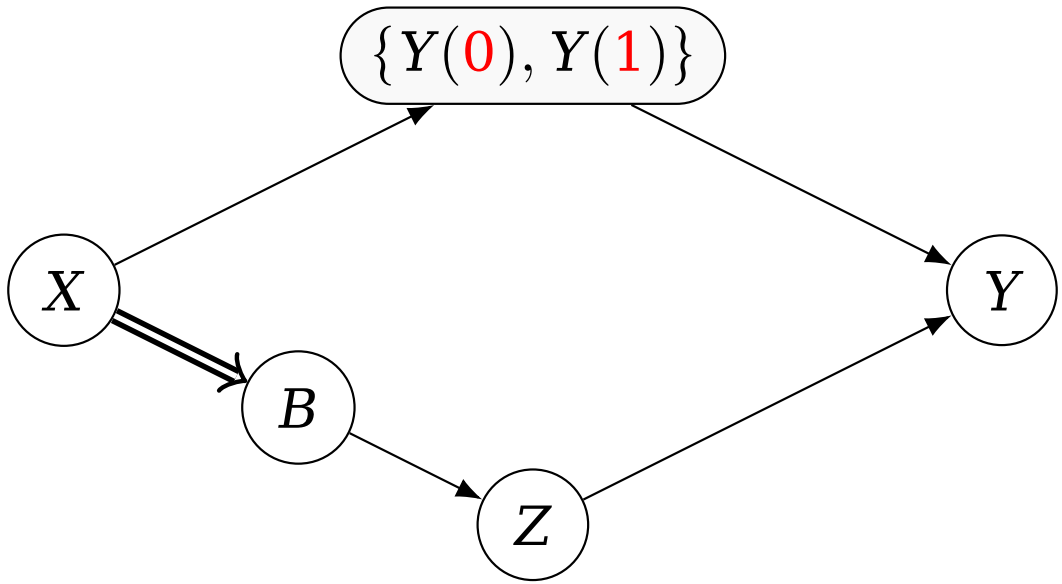
Recall the assumption of *strong ignorability*:



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

Constructing Balance

This can be considered in the balancing score case:



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

Adjustment via the Propensity Score

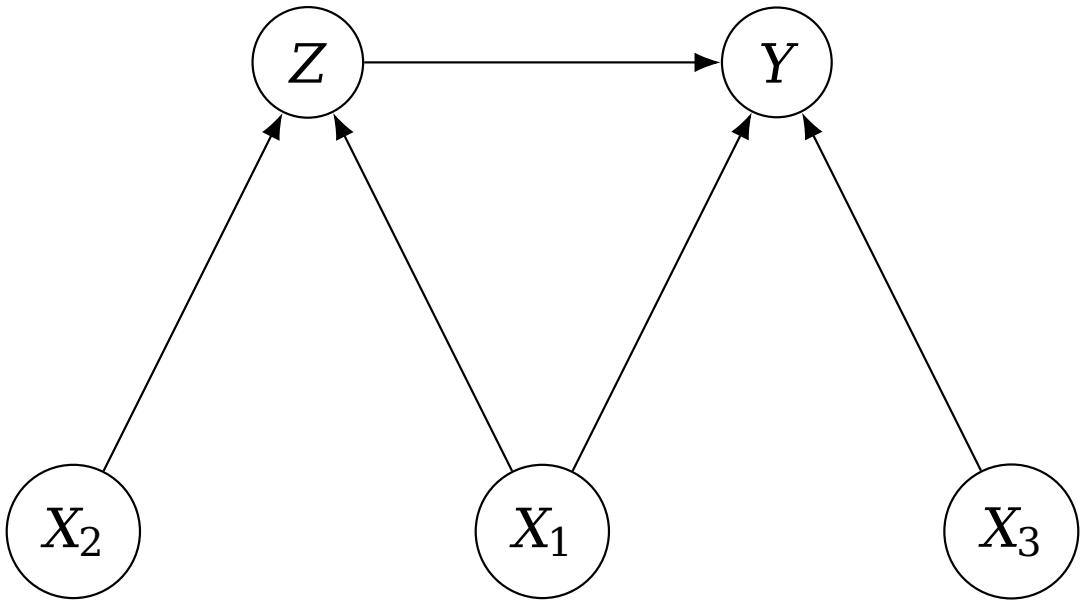
Adjustment methods based on balancing scores can be developed; the balancing score is used to block backdoor (confounding) paths.

We will focus mainly on the binary case, and the propensity score

$$e(X) = \Pr[Z = 1|X].$$

Adjustment via the Propensity Score

The basic set up we consider is the following:



Adjustment via the Propensity Score

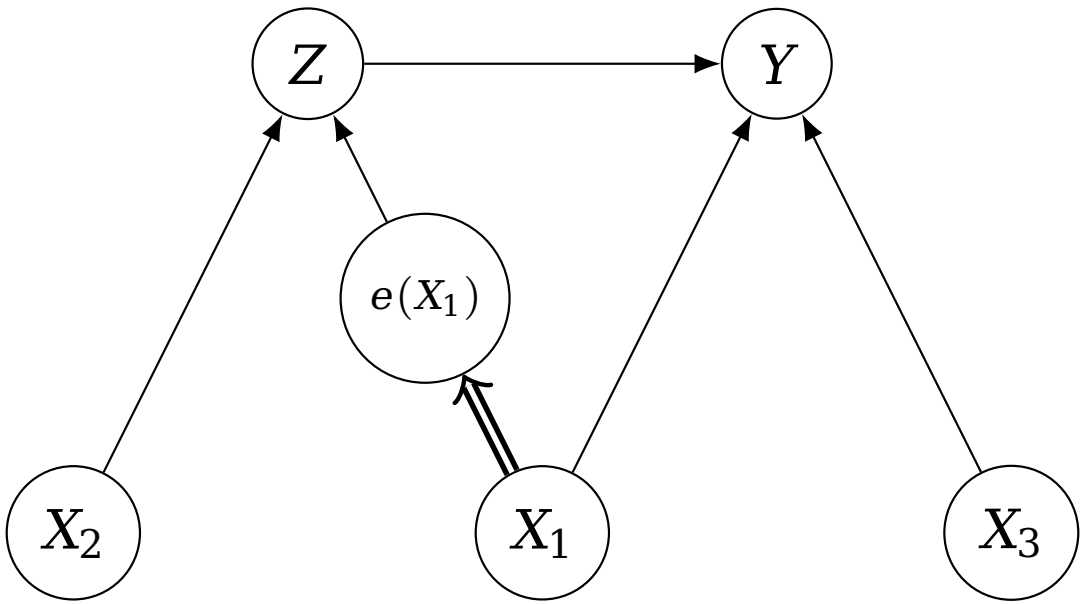
There are three types of covariate:

- X_1 *confounders*
- X_2 *instruments* (pure predictors of treatment)
- X_3 *pure predictors of outcome*

There are no paths connecting X_1, X_2 and X_3, and there is no unmeasured confounding.

Adjustment via the Propensity Score

Including the propensity score retains the previous feature:



Adjustment via the Propensity Score

Note

The propensity score does not need to be a function X_2 or X_3; making it depend *only* on X_1 is sufficient to block the back-door path.

This is the case even though X_2 is a cause of Z; that is, even though

$$f_{Z|X_1,X_2}(z|x_1,x_2) \equiv f_{Z|e(X_1),X_2}(z|e(x_1),x_2)$$

and

$$Z \not\perp\!\!\!\perp X_2 \mid e(X_1)$$

we can still base the propensity score only on X_1 (the confounders).

Stratification and Matching

The propensity score $e(X)$ is a *scalar* random variable irrespective of the dimension of X. We may construct an estimator of the APO by noting

$$\begin{aligned}\mu(z) = \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z=z] &= \mathbb{E}^{\mathcal{E}}_{X}[\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[Y|X,Z=z]] \\ &= \mathbb{E}^{\mathcal{E}}_{X}[\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[Y|X,e(X),Z=z]] \\ &= \mathbb{E}^{\mathcal{O}}_{X}[\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X,e(X),Z=z]]\end{aligned}$$

and, after conditioning on $e(X)$, X and Z are independent, and so sample-based estimation can be utilized.

Stratification and Matching

For fixed values of $e(X)$, we may directly compare

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

for different values of z.

For fixed value e_1, let

$$\mathcal{X}_{e_1} = \{x : e(x) = e_1\}$$

and define a 'local' APO estimator as

$$\widehat{\mu}_{e_1}(z) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\mathcal{X}_{e_1}}(X_i)\mathbb{1}_{\{z\}}(Z_i)Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\mathcal{X}_{e_1}}(X_i)\mathbb{1}_{\{z\}}(Z_i)}$$

Stratification and Matching

We can construct a *stratification* estimator by considering strata of the propensity score.

Consider a partition constructed using

$$\mathcal{X}_j = \{x : e(x) \in \mathcal{E}_j\} \quad j = 1, \ldots, J$$

where $\mathcal{E}_1, \ldots, \mathcal{E}_J$ exhaustively cover the interval $(0, 1)$.

Stratification and Matching

We can define local estimator

$$\widehat{\mu}_{\mathcal{X}_j}(z) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\mathcal{X}_j}(X_i)\mathbb{1}_{\{z\}}(Z_i)Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\mathcal{X}_j}(X_i)\mathbb{1}_{\{z\}}(Z_i)}$$

and global estimator

$$\widehat{\mu}(z) = \sum_{j=1}^{J} \widehat{\mu}_{\mathcal{X}_j}(z)\Pr[X \in \mathcal{X}_j]$$

Note

To construct such estimators of the ATE, say

$$\widehat{\mu}(1) - \widehat{\mu}(0)$$

we require that sufficient data for the different values of z are available with the strata. That is, in any propensity score stratum, we require that there are a large enough number of subjects with both $Z = 0$ and with $Z = 1$.

This is termed an *overlap* condition.

Stratification and Matching

We can also consider *matching* on the propensity score;

- recall that we argued previously that two individuals that had precisely the same X value but different Z values could be directly compared as they were 'matched';
- we can extend this argument to the propensity score – two individuals with the same $e(X)$ value are also considered matched.

Stratification and Matching

There are many ways to carry out matching in practice (where matching on exact values) is not feasible:

- *caliper* matching: two individuals i (with $Z_i = 1$) and j (with $Z_j = 0$) are considered matched if

$$\sqrt{(e(x_i) - e(x_j))^2} < c$$

  for some constant c.

- *1:1 nearest case matching*: for individual i with $Z_i = 1$ we find the individual j with $Z_j = 0$ such that

$$\sqrt{(e(x_i) - e(x_j))^2}$$

  is *minimized*.

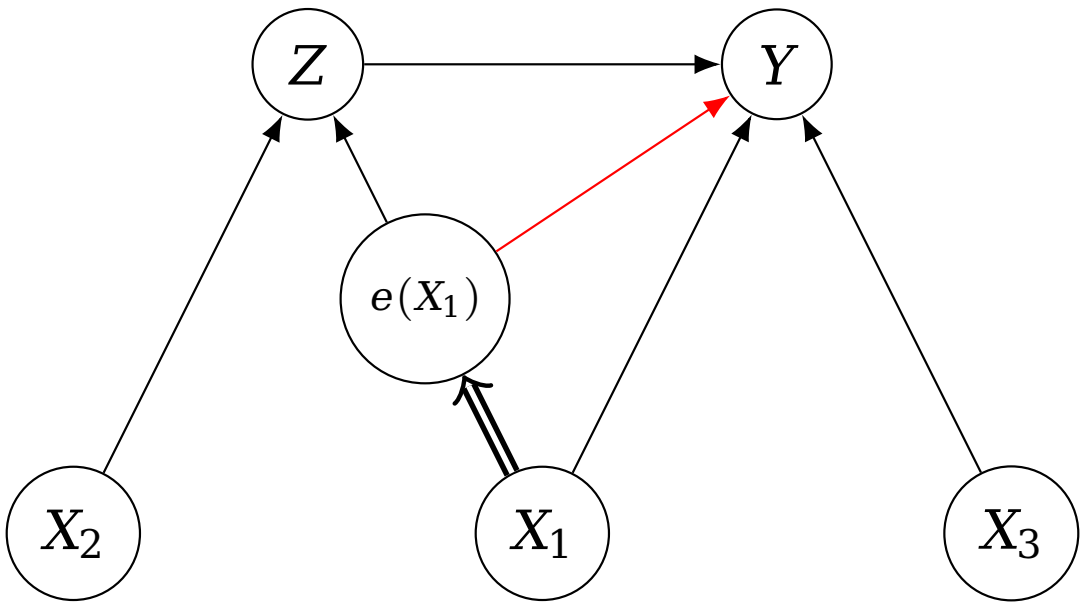
Stratification and Matching

- *1:M matching*: for individual i with $Z_i = 1$ we find the M individuals in the data set with $Z = 0$ such that the distances between their propensity score values and $e(x_i)$ *are the M smallest*.

The statistical properties of matching estimators are not always straightforward to establish.

Propensity Score Regression

Conditioning on $e(X)$ can be achieved using *regression* methods; we consider the model inspired by the DAG



Propensity Score Regression

We may consider the regression model

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

which, as

$$X \perp\!\!\!\perp Z \mid e(X)$$

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

Propensity Score Regression

Example:

Suppose that we have the following data generating model:

- Confounders: $(X_1, X_2)^\top \sim \mathit{Normal}_2((1,1)^\top, \Sigma)$ 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: $Z|X_1, X_2 \sim \mathit{Bernoulli}(e(X_1, X_2))$, where

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

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

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

Propensity Score Regression

Example:

We consider fitting the parametric model

$$m(\mathbf{x}, z; \beta, \psi) = (\beta_0 + \beta_1 \mathbf{x}_1) + z\psi_0$$

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

Propensity Score Regression

```
#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
```

Propensity Score Regression

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.

Propensity Score Regression

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 final term that depends on the propensity score.

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.

Propensity Score Regression

```
#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
```

Propensity Score Regression

Example:

We now have

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

and so correct estimation of ψ_0 has been recovered.

Propensity Score Regression

Example:

Now suppose

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

and we try the same strategy, using the propensity score regression model

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

Propensity Score Regression

```
#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
```

Propensity Score Regression

```
#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
```

Propensity Score Regression

```
#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
```

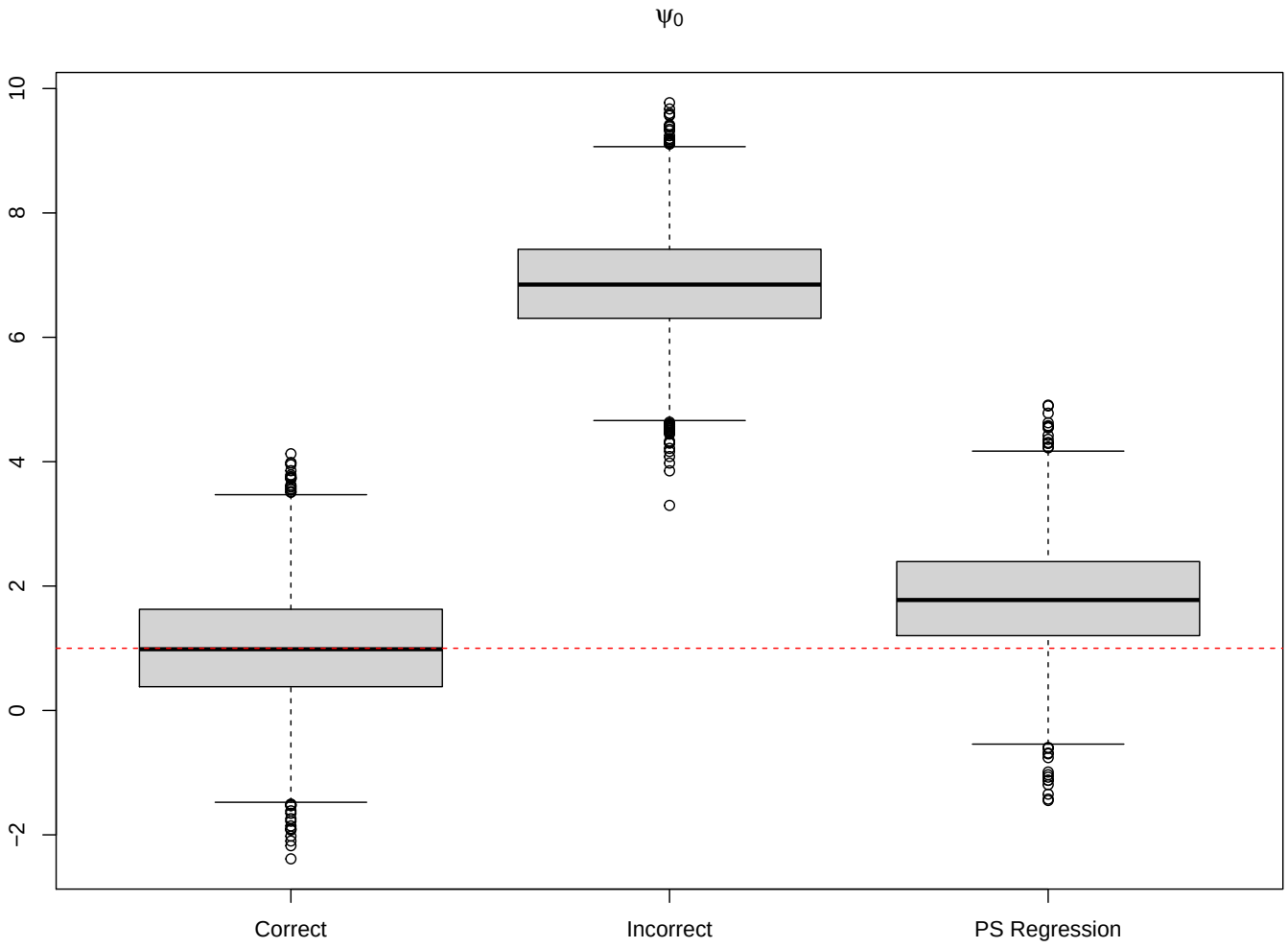
Propensity Score Regression

Example:

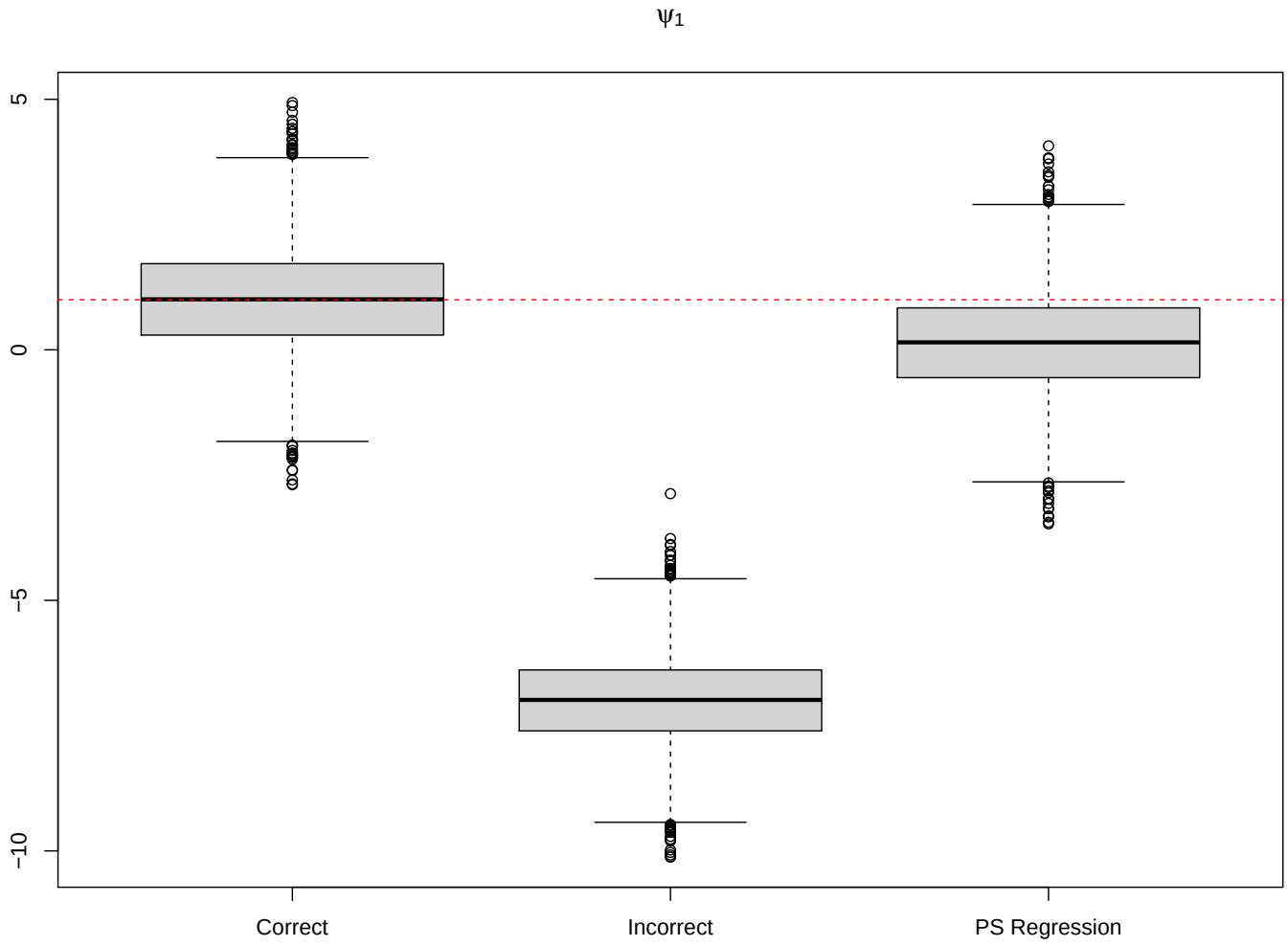
Hard to conclude anything due to the inherent variability, but it seems that including the propensity score does improve the estimation of (ψ_0, ψ_1, ψ_2).

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

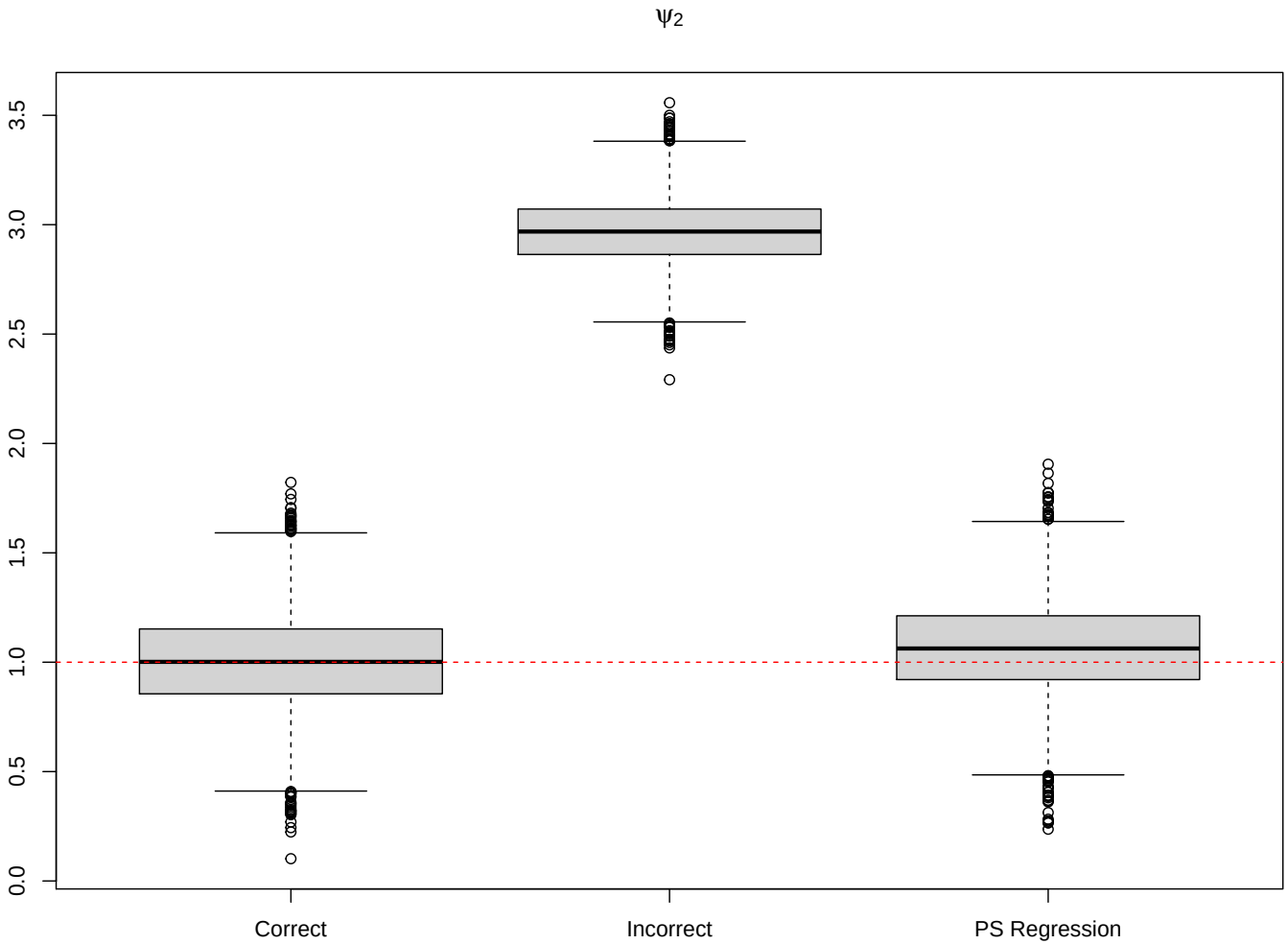
Propensity Score Regression



Propensity Score Regression



Propensity Score Regression



Propensity Score Regression

Example:

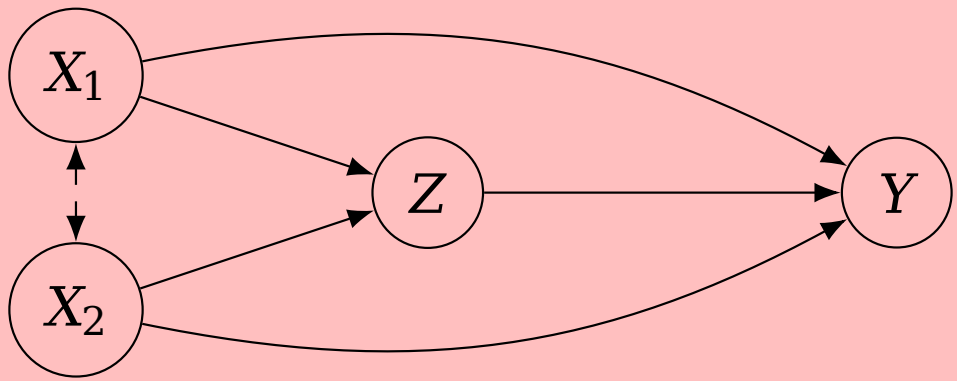
This confirms that including the propensity score *does* improve the estimation of (ψ_0, ψ_1, ψ_2), even if the treatment-free model component is incorrectly specified.

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

Propensity Score Regression

Example:

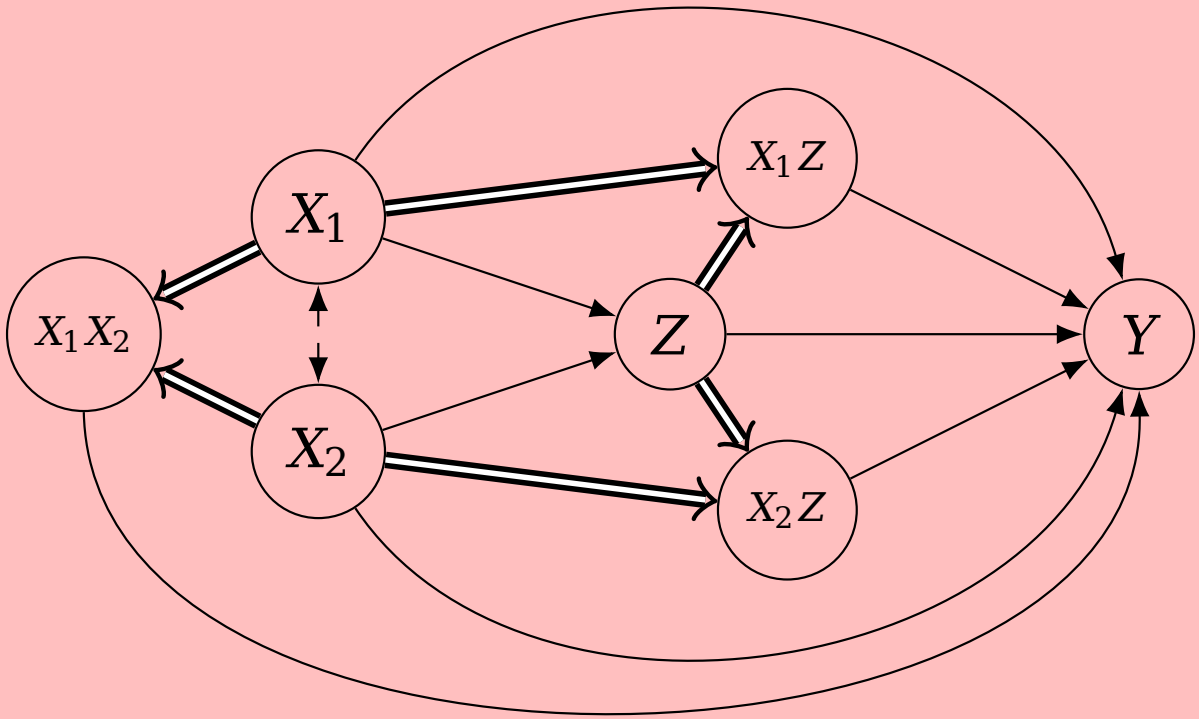
Here is a version of the DAG for the data generating model



Propensity Score Regression

Example:

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



Propensity Score Regression

Example:

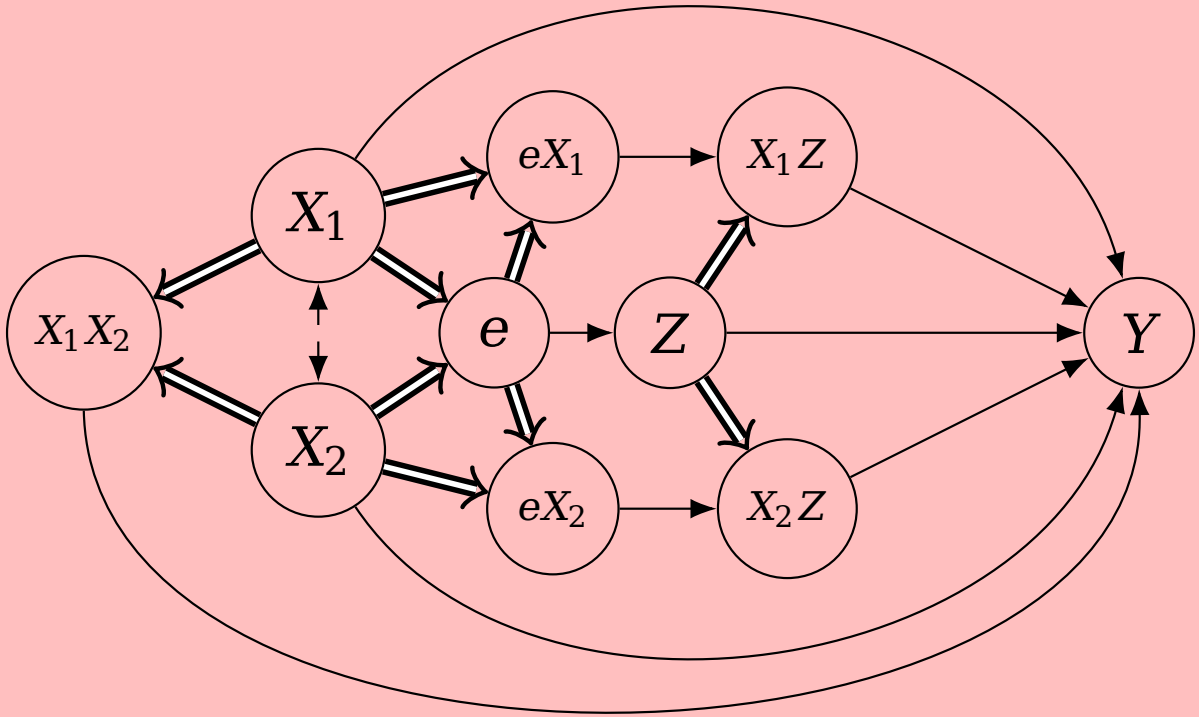
We need to block the open paths via the interactions. This can be achieved by using the model

$$m(\mathbf{x}, z; \beta, \psi, \phi) = (\beta_0 + \beta_1 \mathbf{x}_1) + z(\psi_0 + \psi_1 \mathbf{x}_1 + \psi_2 \mathbf{x}_2) + e(\mathbf{x}_1, \mathbf{x}_2)(\phi_0 + \phi_1 \mathbf{x}_1 + \phi_2 \mathbf{x}_2)$$

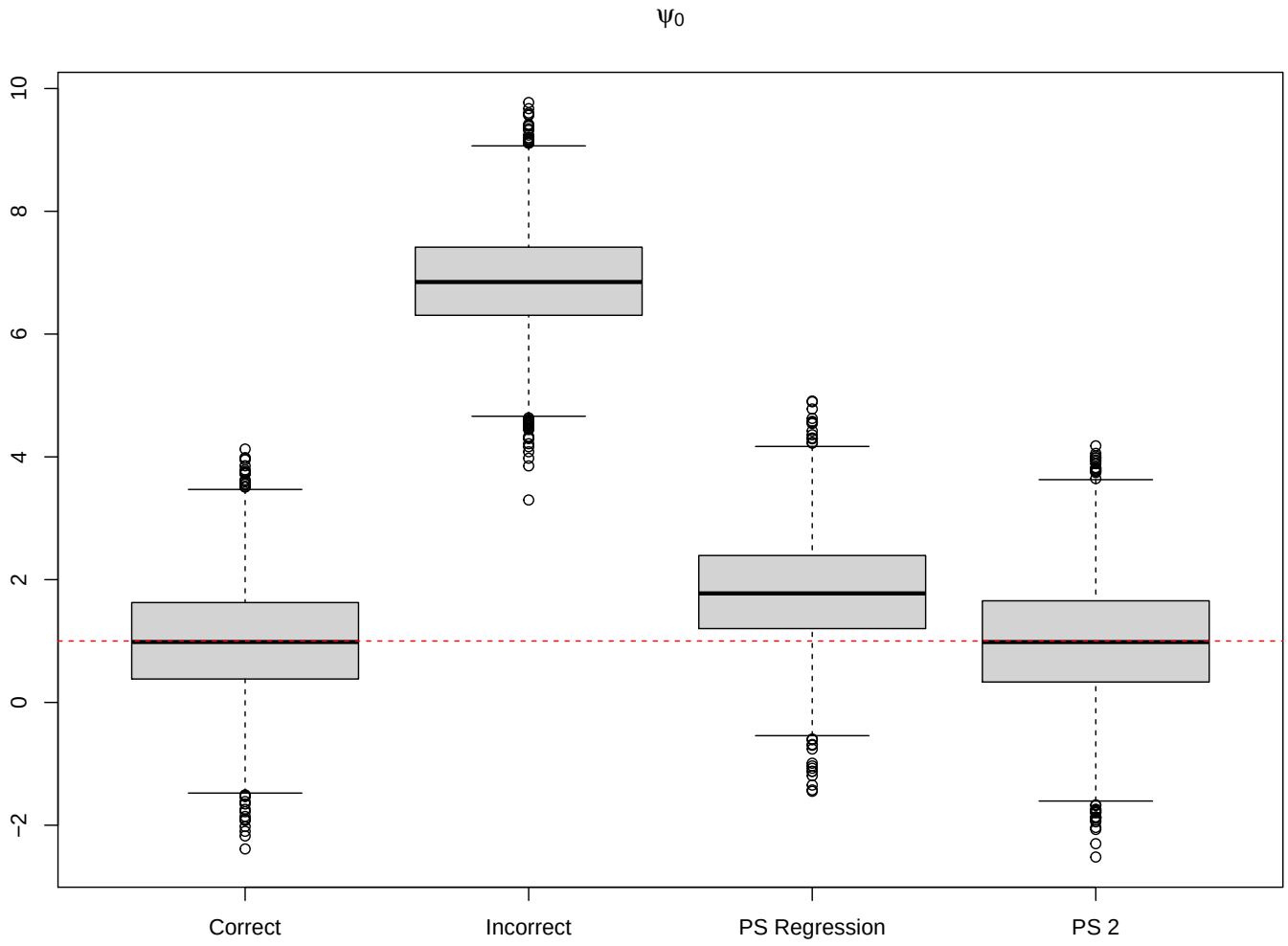
Conditioning on $e(X), e(X)X_1$ and $e(X)X_2$ blocks the paths.

Propensity Score Regression

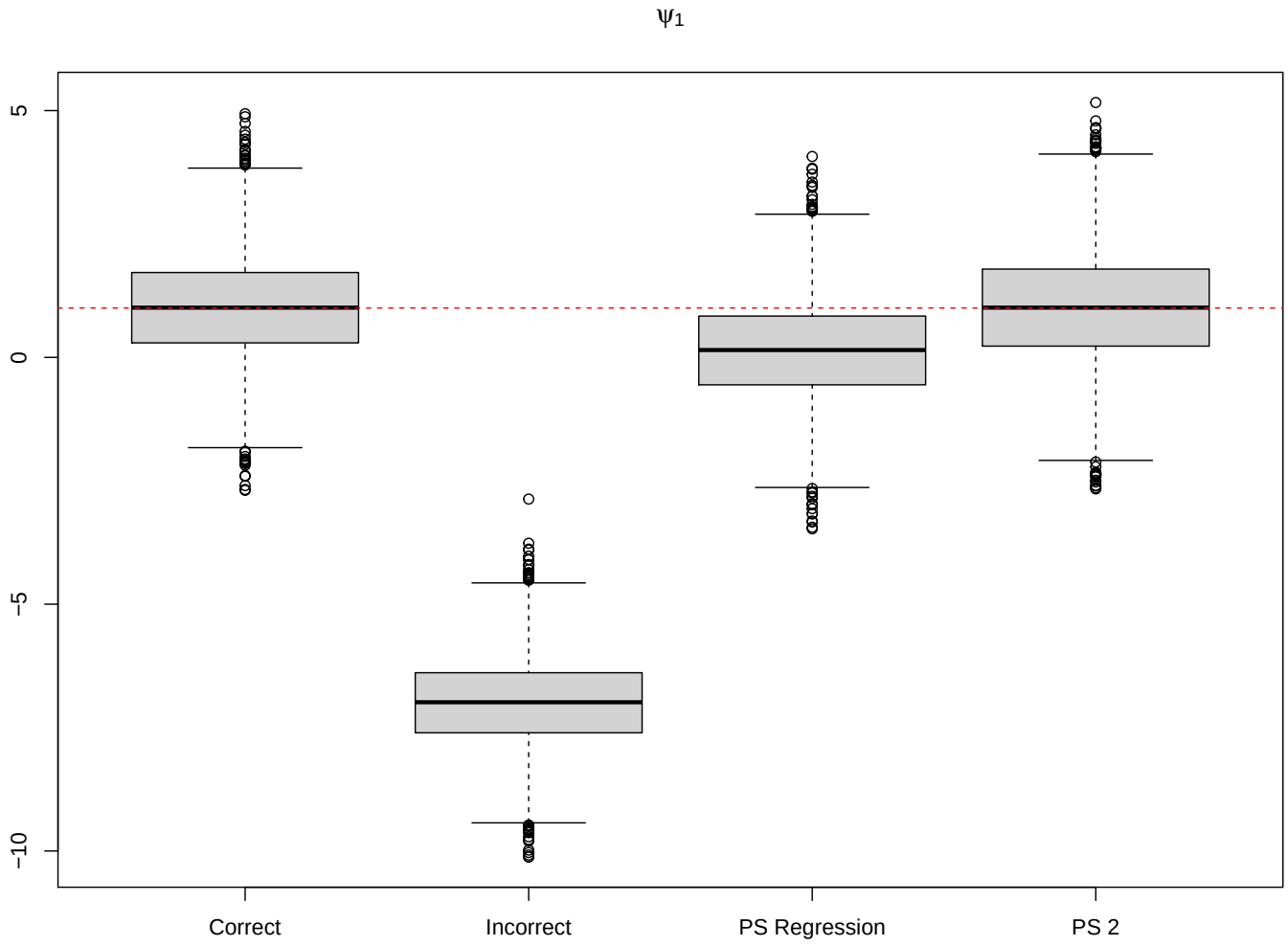
Example:



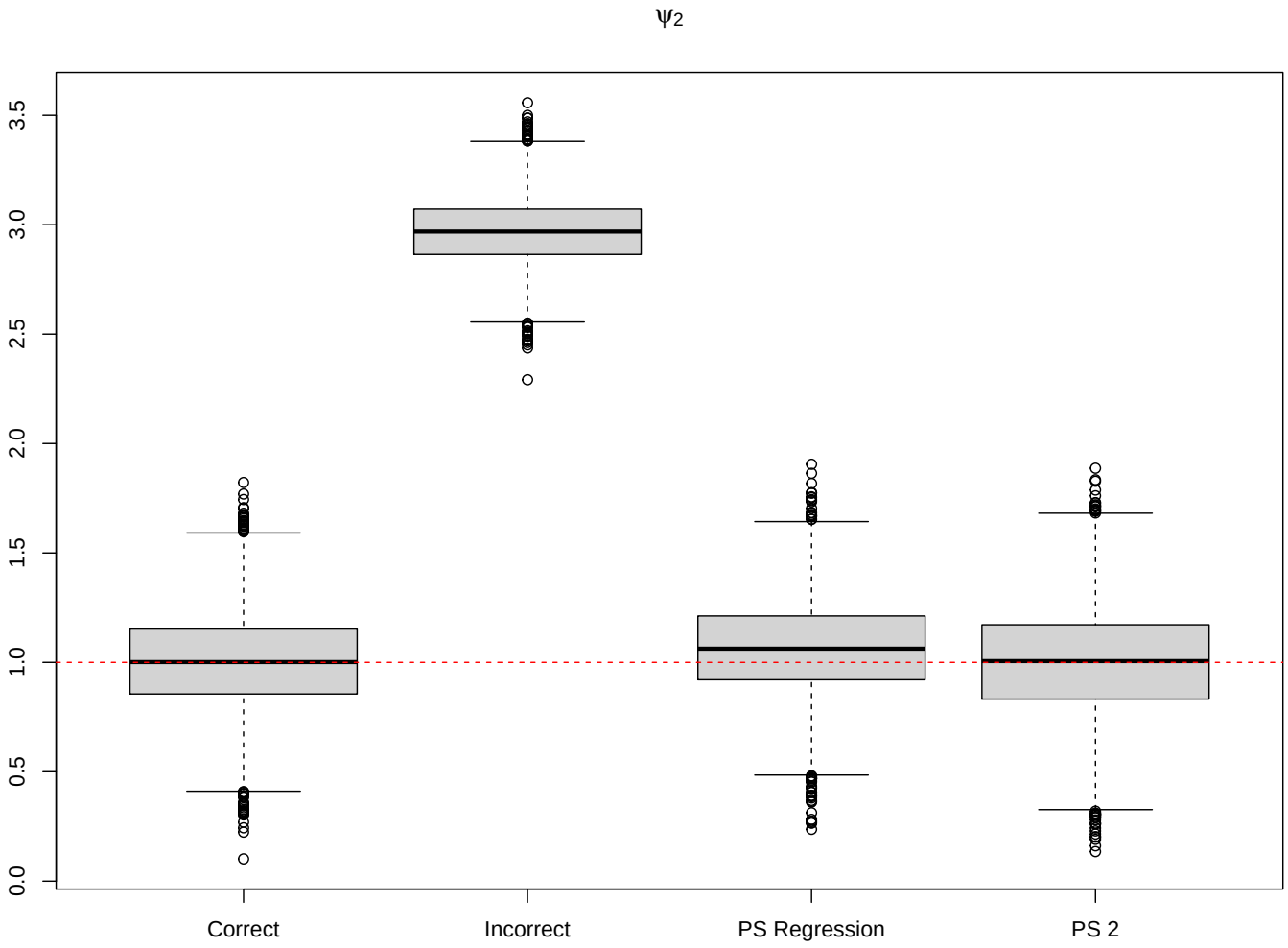
Propensity Score Regression



Propensity Score Regression



Propensity Score Regression



Propensity Score Regression

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.

Propensity Score Regression

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

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

Propensity Score Regression

Example:

Note, however that 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

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

This cannot in general be estimated correctly using

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

Propensity Score Regression

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.

Inverse Probability Weighting

We have from a previous result that

$$\mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z=z] = \frac{\iiint \mathbb{1}_{\{z\}}(z) y \; f^{\mathcal{E}}_{Y|X,Z}(y|x,z) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x) dy \; dx \; dz}{\iiint \mathbb{1}_{\{z\}}(z) f^{\mathcal{E}}_{Y|X,Z}(y|x,z) f^{\mathcal{E}}_{Z}(z) f^{\mathcal{E}}_{X}(x) \; dy \; dx \; dz}$$

and also that

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

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

Inverse Probability Weighting

Using the 'importance sampling' (or *change of measure*) result, we therefore have that

$$\mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \frac{\iiint \mathbb{1}_{\{z\}}(z) y \dfrac{f^{\mathcal{E}}_Z(z)}{f^{\mathcal{O}}_{Z|X}(z|x)} \; f^{\mathcal{O}}_{X,Y,Z}(x,y,z) dy \; dx \; dz}{\iiint \mathbb{1}_{\{z\}}(z) \dfrac{f^{\mathcal{E}}_Z(z)}{f^{\mathcal{O}}_{Z|X}(z|x)} f^{\mathcal{O}}_{X,Y,Z}(x,y,z) \; dy \; dx \; dz}$$

Inverse Probability Weighting

That is

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

or equivalently

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

Inverse Probability Weighting

We therefore have the estimator

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

This is the *Inverse Probability Weighting* (IPW) estimator.

Inverse Probability Weighting

We may also write

$$\widehat{\mu}_{\text{IPW}}(z) = \sum_{i=1}^{n} W_i(z) Y_i$$

where

$$W_i(z) = \frac{\frac{\mathbb{1}_{\{z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}}{\sum_{j=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_j)}{f^{\mathcal{O}}_{Z|X}(Z_j|X_j)}}$$

is a weight, where

$$\sum_{i=1}^{n} W_i(z) = 1 \qquad \mathbb{E}^{\mathcal{O}}_{X,Z}[W_i(z)] = \frac{1}{n}.$$

Inverse Probability Weighting

In the binary case, we have

$$W_i(0) = \frac{\frac{(1-Z_i)}{(1-e(X_i))}}{\sum_{j=1}^{n} \frac{(1-Z_j)}{(1-e(X_j))}} \qquad W_i(1) = \frac{\frac{Z_i}{e(X_i)}}{\sum_{j=1}^{n} \frac{Z_j}{e(X_j)}}$$

where

$$\mathbb{E}^{\mathcal{O}}_{X_i,Z_i}\left[\frac{Z_i}{e(X_i)}\right] = \mathbb{E}^{\mathcal{O}}_{X_i}\left[\frac{e(X_i)}{e(X_i)}\right] = 1$$

by iterated expectation.

Inverse Probability Weighting

Note that an alternative estimator that utilizes the fact that for all i

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

is

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

Inverse Probability Weighting

Note

The two estimators solve two slightly different estimating equations:

- For $\widetilde{\mu}_{\text{IPW}}(1)$:

$$\sum_{i=1}^{n} \left(\frac{Z_i}{e(X_i)} Y_i - \mu(1) \right) = 0$$

  i.e. reweights the datum Y_i.

- For $\widehat{\mu}_{\text{IPW}}(1)$:

$$\sum_{i=1}^{n} \frac{Z_i}{e(X_i)} \left(Y_i - \mu(1) \right) = 0$$

  i.e. reweights the residual $(Y_i - \mu(1))$.

Inverse Probability Weighting

Note

These equations illustrate how IPW operates; it creates a *re-weighted* data set, say for $i = 1, \ldots, n$,

$$(X_i^*, Y_i^*, Z_i^*) = (X_i, Y_i, Z_i) \quad w.p. \left(\frac{Z_i}{e(X_i)} + \frac{1 - Z_i}{1 - e(X_i)} \right)$$

which represent a sample from a pseudo-population in which

$$X^* \perp\!\!\!\perp Z^*.$$

The new data set does not suffer from confounding.

Corrected from original slides.

Inverse Probability Weighting

Note

- $\widehat{\mu}_{\text{IPW}}(z)$ and $\widetilde{\mu}_{\text{IPW}}(z)$ are *unbiased* estimators of $\mu(z)$ by construction.
- In these estimators

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

plays a critical role; this is the function that determines the propensity score.

- There is an important requirement that

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

for any (x, z). This is termed a *positivity* requirement.

Inverse Probability Weighting

Note

- Positivity requires that for all z under consideration

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

that is, in the binary case, we *do not* have that

$$\Pr[Z = z|X = x] = 1$$

for any x. This is sometimes termed the *experimental treatment assignment* (ETA) assumption; that is, no individual receives treatment (or no treatment) with certainty.

Augmentation

The IPW estimators rely on knowledge (and correct specification) of $f^{\mathcal{O}}_{Z|X}(z|x)$, but are otherwise model-free.

Suppose that we have knowledge of the conditional model

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

We could use this model for *outcome regression*.

Augmentation

However, note that

$$\begin{aligned}
\mu(z) &= \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y \mid Z = z] = \mathbb{E}^{\mathcal{E}}_{X}\left[\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[Y \mid X, Z = z]\right] \\
&= \mathbb{E}^{\mathcal{E}}_{X}\left[\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[(Y - \mu(X,Z) + \mu(X,Z) \mid X, Z = z]\right] \\
&= \mathbb{E}^{\mathcal{E}}_{X}\left[\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[(Y - \mu(X,Z)) \mid X, Z = z]\right] \\
&\qquad + \mathbb{E}^{\mathcal{E}}_{X}\left[\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[\mu(X,Z) \mid X, Z = z]\right] \\
&= \mathbb{E}^{\mathcal{E}}_{X}\left[\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[(Y - \mu(X,Z)) \mid X, Z = z]\right] + \mathbb{E}^{\mathcal{E}}_{X}[\mu(X,z)]
\end{aligned}$$

as the internal integrand of the second term does not depend on Y.

Augmentation

Now under the standard assumption, we can write

$$\mathbb{E}_X^{\mathcal{E}}[\mu(X, z)] \equiv \mathbb{E}_X^{\mathcal{O}}[\mu(X, z)]$$

as for outcome regression. Secondly, using the IPW idea, we can re-write

$$\mathbb{E}_X^{\mathcal{E}}\left[\mathbb{E}_{Y|X,Z}^{\mathcal{E}}[(Y - \mu(X, Z)) \mid X, Z = z]\right]$$

as

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

Augmentation

This suggests the alternative moment-based estimator

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

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

Augmentation

Analogous to the earlier forms, we also have a second AIPW estimator

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

where as before

$$W_i(z) = \frac{\frac{\mathbb{1}_{\{z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)}}{\sum_{j=1}^{n} \frac{\mathbb{1}_{\{z\}}(Z_j)}{f^{\mathcal{O}}_{Z|X}(Z_j|X_j)}}.$$

Augmentation

Note

Note that

$$\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\mathbb{E}^{\mathcal{O}}_{Y|X,Z}\left[\frac{\mathbb{1}_{\{z\}}(Z)}{f^{\mathcal{O}}_{Z|X}(Z|X)}(Y-\mu(X,Z)) \mid X, Z=z\right]\right]=0$$

as the internal conditional expectation is zero, so the first term in both $\widetilde{\mu}_{\text{AIPW}}(z)$ and $\widehat{\mu}_{\text{AIPW}}(z)$ has expectation zero .

Note also that

$$\mathbb{E}^{\mathcal{O}}_{X}[\mu(X,z)]=\mu(z)$$

identical to the outcome regression estimator.

Augmentation

Note

The advantage of the AIPW estimator is that it has variance that is *no greater* than the IPW estimator, that is

$$\text{Var}[\widetilde{\mu}_{\text{AIPW}}(z)] \leqslant \text{Var}[\widetilde{\mu}_{\text{IPW}}(z)]$$

and

$$\text{Var}[\widehat{\mu}_{\text{AIPW}}(z)] \leqslant \text{Var}[\widehat{\mu}_{\text{IPW}}(z)]$$

Augmentation

Note

However, note that

$$\text{Var}[\widehat{\mu}_{\text{OR}}(z)] \leqslant \text{Var}[\widetilde{\mu}_{\text{AIPW}}(z)] \leqslant \text{Var}[\widetilde{\mu}_{\text{IPW}}(z)]$$

and

$$\text{Var}[\widehat{\mu}_{\text{OR}}(z)] \leqslant \text{Var}[\widehat{\mu}_{\text{AIPW}}(z)] \leqslant \text{Var}[\widehat{\mu}_{\text{IPW}}(z)]$$

that is, using augmentation we can improve on the IPW estimator, but we cannot improve on the OR estimator.

Importantly, these results follow provided the proposed function $\mu(x, z)$ is *correctly specified*.

Augmentation

The real advantage of AIPW estimators is that can still give consistent estimation even if $\mu(x, z)$ is *mis-specified*.

With mean model $m(x, z)$ we have the two estimators

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

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

Augmentation

Write for the expectation of the second term

$$M(z) = \mathbb{E}_X^{\mathcal{O}}[m(X, z)]$$

and consider the first term of $\widetilde{\mu}_{\text{AIPW}}(z)$. We have that

$$\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\mathbb{E}_{Y|X,Z}^{\mathcal{O}}\left[\frac{\mathbb{1}_{\{z\}}(Z)}{f_{Z|X}^{\mathcal{O}}(Z|X)}(Y - m(X,Z)) \,\middle|\, X, Z = z\right]\right]$$

$$= \mathbb{E}_{X,Z}^{\mathcal{O}}\left[\frac{\mathbb{1}_{\{z\}}(Z)}{f_{Z|X}^{\mathcal{O}}(Z|X)}(\mu(X,Z) - m(X,Z)) \,\middle|\, X, Z = z\right]$$

Augmentation

Provided the model $f^{\mathcal{O}}_{Z|X}(z|x)$ is *correctly specified*, if we perform iterated expectation by conditioning on X, we have that this expectation is equal to

$$\mathbb{E}^{\mathcal{O}}_X[(\mu(X, z) - m(X, z))]$$

and hence we have

$$\begin{aligned}\mathbb{E}[\widetilde{\mu}_{\text{AIPW}}(z)] &= \mathbb{E}^{\mathcal{O}}_X[(\mu(X, z) - m(X, z))] + \mathbb{E}^{\mathcal{O}}_X[m(X, z)] \\ &= \mathbb{E}^{\mathcal{O}}_X[\mu(X, z)] - M(z) + M(z) \\ &= \mu(z).\end{aligned}$$

The same result holds for $\widehat{\mu}_{\text{AIPW}}(z)$.

Augmentation

Note also that if $f^{\mathcal{O}}_{Z|X}(z|x)$ is *mis-specified*, we still have that

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}\left[\frac{\mathbb{1}_{\{z\}}(Z)}{f^{\mathcal{O}}_{Z|X}(Z|X)}(Y - m(X,Z)) \mid X, Z = z\right] = 0$$

provided $m(x,z) = \mu(x,z)$.

Augmentation

Hence we have that both AIPW estimators are unbiased provided *either*

$$m(x, z)$$

or

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

is correctly specified. This phenomenon is known as *double robustness*.

If *both* models are correctly specified, then we have the *optimal* IPW estimator.

Augmentation

Note

In Monte Carlo, the 'augmentation' trick is known as the use of *antithetic variables*. Writing

$$\mathbb{E}^{\mathcal{E}}_{Y|X,Z}[Y \mid X, Z = z]$$

$$= \mathbb{E}^{\mathcal{E}}_{Y|X,Z}[(Y - \mu(X,Z)) \mid X, Z = z] + \mu(X, z)$$

allows us to introduce estimators of the first and second terms that are *negatively correlated*, thereby potentially reducing the variance of the combined estimator overall.

Augmentation

Consider $\widetilde{\mu}(z)$ in the binary case. Write

$$R_i = \frac{\mathbb{1}_{\{z\}}(Z_i)}{f^{\mathcal{O}}_{Z|X}(Z_i|X_i)} = \frac{Z_i}{e(X_i)}z + \frac{(1-Z_i)}{1-e(X_i)}(1-z)$$

so that

$$\widetilde{\mu}(z) = \frac{1}{n}\sum_{i=1}^{n}\{R_i Y_i + (1-R_i)\mu(X_i, z)\}$$

and that

$$\text{Var}[\widetilde{\mu}(z)] = \frac{1}{n}\,\text{Var}[RY + (1-R)\mu(X, z)]$$

where this calculation is carried out with respect to the observational distribution, $\mathcal{O}$.

Augmentation

Note that

$$\text{Var}[RY + (1-R)\mu(X, z)]$$

$$= \text{Var}[RY] + \text{Var}[(1-R)\mu(X, z)] + 2\text{Cov}[RY, (1-R)\mu(X, z)].$$

For the second term

$$\text{Var}[(1-R)\mu(X, z)] = \mathbb{E}[(1-R)^2\{\mu(X, z)\}^2]$$

as by iterated expectation

$$\mathbb{E}_{R|X}[(1-R)|X] = 0 \quad \Longrightarrow \quad \mathbb{E}_{R,Z}[(1-R)\mu(X, z)] = 0.$$

Augmentation

Similarly by iterated expectation

$$\mathrm{Cov}[RY, (1-R)\mu(X, z)] = \mathbb{E}[R(1-R)\{\mu(X, z)\}^2]$$

Therefore

$$\begin{aligned}
&\mathrm{Var}[RY + (1-R)\mu(X, z)] \\
&\quad = \mathrm{Var}[RY] + \mathbb{E}[(1-R)^2\{\mu(X, z)\}^2] + 2\mathbb{E}[R(1-R)\{\mu(X, z)\}^2] \\
&\quad = \mathrm{Var}[RY] + \mathbb{E}[(1-R^2)\{\mu(X, z)\}^2]
\end{aligned}$$

Augmentation

However

$$\mathbb{E}_{R|X}[(1-R^2)\{\mu(X,z)\}^2|X] = \{\mu(X,z)\}^2\mathbb{E}_{R|X}[(1-R^2)\mid X]$$

$$= \{\mu(X,z)\}^2\mathbb{E}_{Z|X}\left[1-\left(\frac{\mathbb{1}_{\{z\}}(Z)}{f^{\mathcal{O}}_{Z|X}(Z|X)}\right)^2\middle|X\right]$$

$$= \{\mu(X,z)\}^2\mathbb{E}_{Z|X}\left[1-\frac{\mathbb{1}_{\{z\}}(Z)}{\left(f^{\mathcal{O}}_{Z|X}(Z|X)\right)^2}\middle|X\right]$$

$$= \{\mu(X,z)\}^2\left(1-\frac{1}{f^{\mathcal{O}}_{Z|X}(z|X)}\right)$$

$$\leqslant 0 \qquad \text{(w.p. 1.)}$$

Augmentation

Therefore

$$\mathrm{Var}[RY + (1-R)\mu(X, z)] \leqslant \mathrm{Var}[RY]$$

and hence

$$\mathrm{Var}[\widetilde{\mu}_{\mathrm{AIPW}}] \leqslant \mathrm{Var}[\widetilde{\mu}_{\mathrm{IPW}}].$$

Similar result for $\widehat{\mu}(z)$.

Augmentation

Note

This variance result *can* hold if $\mu(x, z)$ is mis-specified as the same argument follows for any estimator

$$\widetilde{\mu}(z) = \frac{1}{n}\sum_{i=1}^{n}\{R_i Y_i + (1 - R_i)m(X_i, z)\}$$

We have that

$$\begin{aligned}\text{Var}[RY + (1 - R)m(X, z)] = \ & \text{Var}[RY] + \mathbb{E}[(1 - R)^2\{m(X, z)\}^2] \\ & + 2\mathbb{E}[R(1 - R)\mu(X, z)m(X, z)]\end{aligned}$$

Augmentation

Note

Thus we get variance reduction over the IPW estimator if

$$\mathbb{E}[(1-R)^2\{m(X,z)\}^2] \geqslant -2\mathbb{E}[R(1-R)\mu(X,z)m(X,z)]$$

which will hold if $m(X,z)$ and $\mu(X,z)$ are sufficiently positively correlated.

Augmentation

Note

If *neither* of the models

$$\mu(x, z) \qquad e(x) = f^{\mathcal{O}}_{Z|X}(1|x)$$

is correctly specified, then the AIPW estimator is *biased*. If we instead use

$$m(x, z) \qquad g(x)$$

for these two models, the expectation of $\widetilde{\mu}(1)$ is

$$\mathbb{E}^{\mathcal{O}}_X \left[\frac{e(X)}{g(X)} (\mu(X, 1) - m(X, 1)) \right] + \mathbb{E}^{\mathcal{O}}_X \left[m(X, 1) \right]$$

Augmentation

Note

The bias is therefore

$$\mathbb{E}_X^{\mathcal{O}}\left[\left(\frac{e(X)}{g(X)}-1\right)(\mu(X,1)-m(X,1))\right]$$

Augmentation

Note

In the inverse weighting estimators, it has been assumed that the model

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

is known *precisely*. This can be replaced by a parametric model

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

with α then estimated using maximum likelihood or other methods. The IPW estimators then proceed using the fitted values

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

AIPW via regression

Consider the binary treatment case, and the model

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y \mid X = x, Z = z] = \mu(x, z) + \phi_0 \frac{1 - z}{1 - e(x)} + \phi_1 \frac{z}{e(x)}$$

for parameters ϕ_0 and ϕ_1. We consider estimating these parameters using ordinary least squares. Let

$$R_{0i} = \frac{1 - Z_i}{1 - e(X_i)} \qquad R_{1i} = \frac{Z_i}{e(X_i)}$$

with corresponding observed values r_{0i} and r_{1i}.

AIPW via regression

The OLS score equations are

$$\frac{\partial}{\partial \phi_0} \; : \quad -2\sum_{i=1}^{n} r_{0i}\left(y_i - \mu(\mathbf{x}_i, z_i) - \phi_0 r_{0i} - \phi_1 r_{1i}\right) = 0$$

$$\frac{\partial}{\partial \phi_1} \; : \quad -2\sum_{i=1}^{n} r_{1i}\left(y_i - \mu(\mathbf{x}_i, z_i) - \phi_0 r_{0i} - \phi_1 r_{1i}\right)$$

and we may solve these directly to obtain

$$\widehat{\phi}_z = \frac{\sum_{i=1}^{n} r_{zi}\left(y_i - \mu(\mathbf{x}_i, z_i)\right))}{\sum_{i=1}^{n} r_{zi}^2} \qquad z = 0, 1.$$

AIPW via regression

Predictions from this fit for the $z = 0, 1$ cases are

$$\mu(x_i, 0) + \widehat{\phi}_0 \frac{1}{1 - e(x_i)} \qquad \mu(x_i, 1) + \widehat{\phi}_1 \frac{1}{e(x_i)}$$

respectively.

To obtain an estimates of $\mu(0)$ and $\mu(1)$, we consider

$$\widetilde{\mu}_{\text{AOR}}(0) = \frac{1}{n} \sum_{i=1}^{n} \left(\mu(x_i, 0) + \widehat{\phi}_0 \frac{1}{1 - e(x_i)} \right).$$

and

$$\widetilde{\mu}_{\text{AOR}}(1) = \frac{1}{n} \sum_{i=1}^{n} \left(\mu(x_i, 1) + \widehat{\phi}_1 \frac{1}{e(x_i)} \right).$$

AIPW via regression

Plugging in the estimates of the ϕs, we obtain

$$\widetilde{\mu}_{\text{AOR}}(0) = \frac{1}{n}\sum_{i=1}^{n}\mu(x_i,0) + \frac{\sum_{i=1}^{n}\dfrac{1}{1-e(x_i)}}{\sum_{i=1}^{n} r_{0i}^2}\frac{1}{n}\sum_{i=1}^{n} r_{0i}(y_i - \mu(x_i,0))$$

and

$$\widetilde{\mu}_{\text{AOR}}(1) = \frac{1}{n}\sum_{i=1}^{n}\mu(x_i,1) + \frac{\sum_{i=1}^{n}\dfrac{1}{e(x_i)}}{\sum_{i=1}^{n} r_{1i}^2}\frac{1}{n}\sum_{i=1}^{n} r_{1i}(y_i - \mu(x_i,1)).$$

AIPW via regression

Now

$$\frac{\sum_{i=1}^{n} \frac{1}{e(X_i)}}{\sum_{i=1}^{n} R_{1i}^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{e(X_i)}}{\frac{1}{n}\sum_{i=1}^{n} R_{1i}^2} \xrightarrow{p} \frac{\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)}\right]}{\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\frac{Z^2}{\{e(X)\}^2}\right]} = 1.$$

as

$$\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\frac{Z^2}{\{e(X)\}^2}\right] \equiv \mathbb{E}_{X,Z}^{\mathcal{O}}\left[\frac{Z}{\{e(X)\}^2}\right] = \mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)}\right]$$

AIPW via regression

Therefore

$$\widetilde{\mu}_{\text{AOR}}(1) = \frac{1}{n}\sum_{i=1}^{n}\mu(X_i, 1) + \frac{1}{n}\sum_{i=1}^{n} R_{1i}(Y_i - \mu(X_i, 1)) + \mathrm{o}_p(1)$$

$$= \widetilde{\mu}_{\text{AIPW}}(1) + \mathrm{o}_p(1).$$

Similarly

$$\widetilde{\mu}_{\text{AOR}}(0) = \widetilde{\mu}_{\text{AIPW}}(0) + \mathrm{o}_p(1)$$

This approach to IPW estimation is known as *augmented outcome regression* (AOR).

AIPW via regression

To estimate the ATE

$$\delta = \mu(1) - \mu(0)$$

we can also use augmented outcome regression based on the mean model

$$\mu(\mathbf{x}, z) + \phi\left(\frac{z}{e(\mathbf{x})} - \frac{(1-z)}{(1-e(\mathbf{x}))}\right)$$

and take the difference between the fitted values to obtain the estimate, $\widehat{\delta}_{\text{AOR}}$.

AIPW via regression

Note

The variance of the estimator $\widetilde{\mu}(1)$ is given by

$$\frac{1}{n}\text{Var}_{X,Y,Z}^{\mathcal{O}}\left[\frac{ZY}{e(X)}\right]$$

and under the correct specification of $e(x)$, we have

$$\text{Var}_{X,Y,Z}^{\mathcal{O}}\left[\frac{ZY}{e(X)}\right] = \mathbb{E}_{X,Y,Z}^{\mathcal{O}}\left[\frac{Z^2Y^2}{\{e(X)\}^2}\right] - \{\mu(1)\}^2$$

AIPW via regression

Note

Now

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}\left[\frac{Z^2Y^2}{\{e(X)\}^2}\right] = \mathbb{E}^{\mathcal{O}}_{X,Z}\left[\frac{Zv(X,Z)}{\{e(X)\}^2}\right] = \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{v(X,1)}{e(X)}\right]$$

where

$$v(x,z) = \mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y^2|X=x,Z=z].$$

AIPW via regression

Note

Thus the magnitude of the variance depends on

$$\mathbb{E}_X^{\mathcal{O}}\left[\frac{v(X,1)}{e(X)}\right]$$

and if $v(x,z) \equiv v$, a constant, then this equals

$$v\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)}\right].$$

In general, although we have assumed positivity ($e(x) > 0$ for all x) we have no guarantee that the expectation in this expression is *finite*; even if it is finite, it may be large due to the reciprocation of $e(X)$.

AIPW via regression

Note

This feature can affect all IPW estimators.

- it is sometimes assumed that $e(x)$ must be bounded away from zero;
- alternatively, it is common to *truncate* the propensity score values such that either data for which, for some $\epsilon > 0$

$$e(x_i) < \epsilon$$

are omitted, or to use

$$e_\epsilon(x_i) = \max\{e(x_i), \epsilon\}.$$

AIPW via regression

Note

Note that

$$\widetilde{\delta}_{\text{IPW}} \equiv \widetilde{\mu}_{\text{IPW}}(1) - \widetilde{\mu}_{\text{IPW}}(0) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-e(X_i)}$$

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

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

AIPW via regression

Note

Note also that

$$e(X_i)(1 - e(X_i)) \equiv \mathrm{Var}_{Z_i|X_i}[Z_i|X_i]$$

so in fact

$$\widetilde{\mu}_{\mathrm{IPW}}(1) - \widetilde{\mu}_{\mathrm{IPW}}(0) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i - e(X_i)}{\mathrm{Var}_{Z_i|X_i}[Z_i|X_i]}\right) Y_i$$

which resembles the earlier formulae for the randomized experimental case.

AIPW via regression

Note

Finally, note the variance of $\widetilde{\delta} = \widetilde{\mu}(1) - \widetilde{\mu}(0)$ is

$$\frac{1}{n}\text{Var}^{\mathcal{O}}_{X,Y,Z}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)Y\right].$$

Now, in this expression, the variance term can be written

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)^2 Y^2\right] - \delta^2$$

as $\widetilde{\delta}$ is unbiased for δ.

AIPW via regression

Note

Using the previous notation, we have that the first term is

$$\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{Z}{e(X)}-\frac{(1-Z)}{1-e(X)}\right)^2 v(X,Z)\right]$$

and if $v(X,Z)\equiv v$, a constant, this reduces to

$$v\,\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{Z}{e(X)}-\frac{(1-Z)}{1-e(X)}\right)^2\right].$$

AIPW via regression

Note

Finally, the expectation simplifies

$$\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{Z}{e(X)}-\frac{(1-Z)}{1-e(X)}\right)^2\right]=\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{Z}{e(X)}\right)^2\right]+\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{(1-Z)}{1-e(X)}\right)^2\right]-2\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\frac{Z(1-Z)}{e(X)(1-e(X))}\right]$$

However, $Z(1-Z)=0$ w.p. 1, so the third term is zero.

AIPW via regression

Note

Hence as $Z^2 = Z$ and $(1-Z)^2 = (1-Z)$ w. p. 1,

$$\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)^2\right] = \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{e(X)}\right] + \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{1-e(X)}\right]$$

$$= \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{e(X)(1-e(X)}\right]$$

$$= \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{\mathrm{Var}^{\mathcal{O}}_{Z|X}[Z|X]}\right]$$

AIPW via regression

Note

Therefore, the variance of $\widetilde{\delta}_{\text{IPW}}$ is

$$\frac{v}{n}\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{\text{Var}_{Z|X}^{\mathcal{O}}[Z|X]}\right] - \frac{\delta^2}{n}$$

G-estimation

Earlier we saw the idea of *propensity score regression*, where we construct a model of the form

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

which is potentially useful as

$$X \perp\!\!\!\perp Z \mid e(X).$$

G-estimation

In the binary treatment, linear model case, we saw that if the data generating model is

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

then the propensity score regression model

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

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

G-estimation

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

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

analytically using the usual approaches.

G-estimation

However, note that subtracting the third component from the second, we obtain the equivalent system

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

which has an identical solution.

The second component takes the form

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

G-estimation

Notice first that if the mean model is *correctly specified*

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

with $\phi = \mathbf{0}$, that is, the true model is nested inside the fitted model, then β and ψ will be consistently estimated, and we will observe

$$\widehat{\phi} \xrightarrow{p} 0$$

as $n \longrightarrow \infty$; indeed, for finite n, the expected value of $\widehat{\phi}$ is zero.

G-estimation

Now suppose the mean model is *mis-specified*, but that

(i) the propensity score model $e(x)$ is correctly specified;

(ii) the random quantity

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

is *functionally independent* of Z_i, that is, the dependence of the mean model on Z_i is correctly specified, and the effect of Z_i is captured via

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

G-estimation

Then we have that

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))\mathbf{X}_2^\top(Y - \mathbf{X}_1\beta - Z\,\mathbf{X}_2\psi - e(X)\mathbf{X}_2\phi)] = \mathbf{0}$$

as, using iterated expectation, we have first that

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[(Y - \mathbf{X}_1\beta - Z\,\mathbf{X}_2\psi - e(X)\mathbf{X}_2\phi)|X, Z] = h(\mathbf{X}; \beta, \psi, \phi)$$

where

$$\begin{aligned} h(\mathbf{x}; \beta, \psi, \phi) &= (\mathbf{x}_0\beta_{\text{TRUE}} + z\,\mathbf{x}_2\psi) - (\mathbf{x}_1\beta + z\,\mathbf{x}_2\psi + e(x)\mathbf{x}_2\phi)) \\ &= \mathbf{x}_0\beta_{\text{TRUE}} - (\mathbf{x}_1\beta + e(x)\mathbf{x}_2\phi)). \end{aligned}$$

G-estimation

That is, $h(\mathbf{X}; \beta, \psi, \phi)$ is *functionally independent* of Z. Then

$$\begin{aligned}
&\mathbb{E}^{\mathcal{O}}_{Z|X}[(Z - e(X))\mathbf{X}_2^\top h(X; \beta, \psi, \phi)|X] \\
&\qquad = \mathbf{X}_2^\top h(X; \beta, \psi, \phi)\mathbb{E}^{\mathcal{O}}_{Z|X}[(Z - e(X))|X] \\
&\qquad = \mathbf{0}
\end{aligned}$$

by the correct specification of $e(X)$, so the overall expectation is zero.

Thus, this is an *unbiased* estimating equation and therefore the solutions to the resulting equation are *consistent* for the true values.

G-estimation

This is another form of *double robustness*; inference for ψ is correct if *either*

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

(or both) is correctly specified, *provided* the expectation

$$\mathbb{E}^{\mathcal{O}}_{\varepsilon|X,Z}[\varepsilon \mid X, Z]$$

does not depend on Z.

G-estimation

Note

Under correct specification of the propensity score, the G-estimation procedure is robust to mis-specification of the treatment-free mean model

$$\mathbf{x}_1\beta$$

so in fact we may re-write the G-estimating equation by combining the two terms that do not depend on Z, and omitting the nuisance parameter ϕ from the procedure.

Note

That is, consider the reduced form

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

This form still leads to double robustness by identical arguments.

G-estimation

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

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

and in the simplest case with ψ one-dimensional

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

say.

G-estimation

The estimating equation invokes the moment requirement

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))(Y - Z\psi_0)] = 0$$

which is a form of *orthogonality* statement, that is

$(Z - e(X))$ is *uncorrelated* with $(Y - Z\psi_0)$.

G-estimation

In this case we can solve explicitly to obtain

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

with corresponding estimator

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

G-estimation

Using standard arguments, we have that as $n \longrightarrow \infty$

$$\frac{\sum_{i=1}^{n}(Z_i - e(X_i))Y_i}{\sum_{i=1}^{n} Z_i(Z_i - e(X_i))} \xrightarrow{p} \frac{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y]}{\mathbb{E}^{\mathcal{O}}_{X,Z}[Z(Z - e(X))]}.$$

and note that in the denominator, by iterated expectation

$$\mathbb{E}^{\mathcal{O}}_{X,Z}[Z(Z - e(X))] = \mathbb{E}^{\mathcal{O}}_{X}\left[\mathbb{E}^{\mathcal{O}}_{Z|X}[Z(Z - e(X))|X]\right].$$

Then, as $Z^2 = Z$ w.p. 1, we have

$$\mathbb{E}^{\mathcal{O}}_{Z|X}[Z(Z - e(X))|X] = \mathbb{E}^{\mathcal{O}}_{Z|X}[Z^2 - Ze(X)|X] = e(X)(1 - e(X))$$

G-estimation

Thus

$$\begin{aligned}\mathbb{E}^{\mathcal{O}}_{X,Z}[Z(Z - e(X)] &= \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))] \\ &\equiv \mathbb{E}^{\mathcal{O}}_{X}[\text{Var}^{\mathcal{O}}_{Z|X}[Z|X]]\end{aligned}$$

where the second line follows as

$$Z|X \sim \mathit{Bernoulli}(e(X)).$$

G-estimation

In the numerator

$$\begin{aligned}\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y] &= \mathbb{E}^{\mathcal{O}}_{X,Z}[(Z - e(X))\mu(X, Z)] \\ &= \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))\mu(X, 1) - (1 - e(X))e(X)\mu(X, 0)] \\ &= \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))(\mu(X, 1) - \mu(X, 0))] \\ &= \psi_0 \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))]\end{aligned}$$

as, here

$$\mu(X, 1) - \mu(X, 0) = \psi_0$$

with probability 1.

G-estimation

Therefore

$$\frac{\sum_{i=1}^{n}(Z_i - e(X_i))Y_i}{\sum_{i=1}^{n} Z_i(Z_i - e(X_i))} \xrightarrow{p} \frac{\psi_0 \mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]} = \psi_0.$$

and we have consistent estimation.

G-estimation

For the variance, note first that by the previous result

$$\widehat{\psi}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{(Z_i - e(X_i))}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]} Y_i + \mathrm{o}_p(1)$$

so we may compute the large sample variance by computing the variance of the statistic on the right hand side; this variance is

$$\frac{1}{n\{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]\}^2} \mathrm{Var}_{X,Y,Z}^{\mathcal{O}}[(Z - e(X))Y].$$

G-estimation

We have that

$$
\begin{aligned}
&\mathrm{Var}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y] \\
&= \mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))^2 Y^2] - \{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y]\}^2 \\
&= \mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))^2 Y^2] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))]\}^2 \\
&= \mathbb{E}^{\mathcal{O}}_{X,Z}[(Z - e(X))^2 v(X,Z)] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))]\}^2
\end{aligned}
$$

G-estimation

If $v(x,z) = v$ is a constant, then this becomes

$$v\mathbb{E}^{\mathcal{O}}_{X,Z}[(Z - e(X))^2] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_X[e(X)(1-e(X))]\}^2$$

but by iterated expectation

$$\mathbb{E}^{\mathcal{O}}_{X,Z}[(Z - e(X))^2] = \mathbb{E}^{\mathcal{O}}_X[e(X)(1-e(X))].$$

Thus

$$\begin{aligned}&\mathrm{Var}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y] \\ &\qquad = v\mathbb{E}^{\mathcal{O}}_X[e(X)(1-e(X))] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_X[e(X)(1-e(X))]\}^2.\end{aligned}$$

G-estimation

Therefore, combining all the elements, we conclude that the variance of $\widehat{\psi}_0$, obtained by G-estimation, satisfies

$$n\mathrm{Var}^{\mathcal{O}}[\widehat{\psi}_0] \longrightarrow v\frac{1}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]} - \psi_0^2.$$

Recall that in this model

$$\psi_0 = \mathbb{E}_X^{\mathcal{O}}[\mu(X,1) - \mu(X,0)] = \mu(1) - \mu(0)$$

so ψ_0 is the ATE.

G-estimation

We contrast this with the variance of the IPW estimator of the ATE obtained earlier: we had that as $\delta = \psi_0$,

$$n\text{Var}^{\mathcal{O}}[\widetilde{\delta}_{\text{IPW}}] = v\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)(1-e(X))}\right] - \psi_0^2$$

G-estimation

Now by Jensen's inequality

$$\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)(1-e(X))}\right] \geqslant \frac{1}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]}$$

and so it is evident that for n large enough

$$\mathrm{Var}^{\mathcal{O}}[\widetilde{\delta}_{\mathrm{IPW}}] > \mathrm{Var}^{\mathcal{O}}[\widehat{\psi}_0].$$

Recall, however, that the two methods make different assumptions: specifically, G-estimation requires the *correct specification* of the treatment effect model.

Note

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

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

and we can achieve similar comparisons with AIPW estimators.

G-estimation

Note

Throughout, we have assumed $e(X)$ is known precisely. More typically, we will propose a parametric model $e(x) \equiv e(x;\alpha)$, and then estimate α using a further estimation procedure.

For example, using *logistic regression*, we could solve

$$\sum_{i=1}^{n} \mathbf{x}_i^\top (z_i - e(x_i;\alpha)) = 0$$

where $\mathbf{x}_i$ is a row vector of the same dimension as α.

Having obtained $\widehat{\alpha}$, we then proceed with $e(x_i;\widehat{\alpha})$ in place of $e(x_i)$ in the earlier formulae, using a *plug-in* strategy.

G-estimation

Note

The plug-in approach will work provided the estimator of α is consistent; but

- Should we 'pay a penalty' for estimating α, that is, will the variance of the ATE estimators *increase* ?
- Do we need to account for the estimation of α ?

Part 5
Semiparametric Inference

Semiparametric Inference

The *semiparametric theory* of estimation that can be used to justify several of the previous causal inference methods.

We consider models that include

- parametric components (with a finite dimensional parameter of interest)
- nonparametric components (with an infinite dimensional parameter that is usually a nuisance parameter)

Example: Ordinary Least Squares

Ordinary least squares is a semiparametric approach:

$$\widehat{\theta} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - \mu(\mathbf{x}_i; \theta))^2$$

- parametric mean model $\mu(x; \theta)$, $\theta \in \mathbb{R}^d$ say
- nonparametric distributional assumption for the conditional distribution of Y given $X = x$.

Example: Ordinary Least Squares

If

$$\varepsilon_i = Y_i - \mu(X_i; \theta)$$

we assume only that

$$\int f_{\varepsilon|X}(t|x)\, dt = 1 \qquad \int t\, f_{\varepsilon|X}(t|x)\, dt = 0$$

and

$$\int t^2\, f_{\varepsilon|X}(t|x)\, dt < \infty$$

but make no other assumptions.

Linear models: a recap

Suppose we have a linear model

$$Y_i = \mathbf{x}_{i1}\beta + \mathbf{x}_{i2}\psi + \varepsilon_i$$

where

- $\mathbf{x}_{i1}$ is $1 \times r$
- β is $r \times 1$
- $\mathbf{x}_{i2}$ is $1 \times q$
- ψ is $q \times 1$

Let θ be the concatenation of β and ψ, so that θ is $p \times 1$ with $p = q + r$.

Linear models: a recap

In vector form, for a random sample of size n, we have

$$\mathbf{Y} = \mathbf{X}_1\beta + \mathbf{X}_2\psi + \varepsilon$$

We require *linear independence* of the columns of the combined matrix

$$[\mathbf{X}_1\ \mathbf{X}_2]\,.$$

Linear models: a recap

Note that

- $\mathbf{X}_1\beta$ is an arbitrary point in the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}_1$ – denote this linear subspace Λ.
- $\mathbf{X}_2\psi$ is an arbitrary point in the linear subspace of $\mathbb{R}^n$ spanned by the columns of $\mathbf{X}_2$.

Linear models: a recap

For OLS estimation we need to solve

$$\begin{pmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{pmatrix} (\mathbf{y} - \mathbf{X}_1\beta - \mathbf{X}_2\psi) = \mathbf{0}_p$$

We can write the left hand side

$$\begin{pmatrix} \mathbf{X}_1^\top \\ \mathbf{X}_2^\top \end{pmatrix} \varepsilon$$

which is an element in a p-dimensional space.

Linear models: a recap

Suppose that ψ is the *parameter of interest*. For *any* ψ, we may write the model

$$(\mathbf{Y} - \mathbf{X}_2\psi) = \mathbf{X}_1\beta + \varepsilon$$

and hence for *nuisance parameter* β deduce that

$$\widetilde{\beta} = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}\mathbf{X}_1^\top(\mathbf{y} - \mathbf{X}_2\psi).$$

This follows from the usual OLS result; note that

$$\widetilde{\beta} \equiv \widetilde{\beta}(\psi)$$

is a function of ψ.

Linear models: a recap

We then have that

$$\mathbf{X}_1\widetilde{\beta} = \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top(\mathbf{y} - \mathbf{X}_2\psi) = \mathbf{H}_1(\mathbf{y} - \mathbf{X}_2\psi)$$

say, with

$$\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top \qquad (n \times n).$$

Note that $\mathbf{H}_1$ is *symmetric*, and that

$$\mathbf{H}_1\mathbf{H}_1 = \mathbf{H}_1^\top\mathbf{H}_1 = \mathbf{H}_1.$$

Linear models: a recap

Note that

- the $n \times 1$ vector $\mathbf{X}_1\widetilde{\beta}$ is a *specific* point in the space Λ.
- for any ψ, $\mathbf{X}_1\widetilde{\beta}$ is the *closest* point in Λ to $(\mathbf{y} - \mathbf{X}_2\psi)$;
- the matrix $\mathbf{H}_1$ *projects* $(\mathbf{y} - \mathbf{X}_2\psi)$ onto Λ to compute $\mathbf{X}_1\widetilde{\beta}$.
- the '*residual*' vector after finding $\mathbf{X}_1\widetilde{\beta}$ is

$$\begin{aligned}(\mathbf{y} - \mathbf{X}_2\psi) - \mathbf{X}_1\widetilde{\beta} &= (\mathbf{y} - \mathbf{X}_2\psi) - \mathbf{H}_1(\mathbf{y} - \mathbf{X}_2\psi) \\ &= (\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)\end{aligned}$$

Linear models: a recap

- Note that for any values of β and ψ

$$\{(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)\}^\top \{\mathbf{X}_1\beta\} = (\mathbf{y} - \mathbf{X}_2\psi)^\top (\mathbf{I}_n - \mathbf{H}_1)^\top \mathbf{X}_1\beta$$
$$= 0$$

  where $\mathbf{X}_1\beta$ is an arbitrary point in Λ, as

$$\begin{aligned}(\mathbf{I}_n - \mathbf{H}_1)^\top \mathbf{X}_1 &= (\mathbf{I}_n - \mathbf{H}_1^\top)\mathbf{X}_1 \\ &= \mathbf{X}_1 - \mathbf{H}_1\mathbf{X}_1 && \text{as } \mathbf{H}_1^\top = \mathbf{H}_1 \\ &= \mathbf{X}_1 - \{\mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top\}\mathbf{X}_1 \\ &= \mathbf{X}_1 - \mathbf{X}_1 = \mathbf{0}.\end{aligned}$$

Linear models: a recap

Therefore the residual vector

$$(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)$$

is *orthogonal* to Λ.

Linear models: a recap

Returning to the estimation of ψ, we now have the reduced form model

$$(\mathbf{Y} - \mathbf{X}_2\psi) = \mathbf{X}_1\widetilde{\beta} + \varepsilon$$

or equivalently

$$(\mathbf{Y} - \mathbf{X}_2\psi) = \mathbf{H}_1(\mathbf{Y} - \mathbf{X}_2\psi) + \varepsilon$$

that is

$$(\mathbf{I}_n - \mathbf{H}_1)\mathbf{Y} = (\mathbf{I}_n - \mathbf{H}_1)\mathbf{X}_2\psi + \varepsilon.$$

which is a linear form in ψ that can be solved using OLS as usual.

Note that the estimating equation is

$$\mathbf{X}_2^\top(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi) = \mathbf{0}$$

again implying a need for orthogonality between the columns of $\mathbf{X}_2$ and the residual quantity $(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi)$.

Note also that for any β

$$(\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_2\psi) = (\mathbf{I}_n - \mathbf{H}_1)(\mathbf{y} - \mathbf{X}_1\beta - \mathbf{X}_2\psi).$$

as

$$(\mathbf{I}_n - \mathbf{H}_1)\mathbf{X}_1 = (\mathbf{I}_n - \mathbf{H}_1)^\top\mathbf{X}_1 = \mathbf{0}.$$

Hilbert spaces

Hilbert space: A (real) Hilbert space $\mathcal{H}$ is a vector space that

- has an associated *inner product* $\langle ., . \rangle$, such that for elements $h_1, h_2, h_3 \in \mathcal{H}$

$$\langle h_1, h_2 \rangle \in \mathbb{R}$$

can be computed and has certain properties.

- $\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$
- $\langle a h_1 + b h_2, h_3 \rangle = a \langle h_1, h_3 \rangle + b \langle h_2, h_3 \rangle$
- $\langle h, h \rangle \geqslant 0$

Hilbert spaces

- has a *metric* defined in terms of the inner product that defines the concept of *norm*, *distance* and hence *convergence* in the space.
  - *norm*:
$$\|h\| = \sqrt{\langle h, h \rangle}$$
  - *distance*:
$$d(h_1, h_2) = \|h_1 - h_2\| = \sqrt{\langle h_1 - h_2, h_1 - h_2 \rangle}$$

Hilbert spaces

We may deduce that for $h_1, h_2, h_3 \in \mathcal{H}$, for example

$$d(h_1, h_3) \leqslant d(h_1, h_2) + d(h_2, h_3)$$

$$\|h_1 + h_2\| \leqslant \|h_1\| + \|h_2\|$$

$$|\langle h_1, h_2 \rangle| \leqslant \|h_1\| \|h_2\|$$

- is *complete*: essentially, Cauchy sequences of elements of $\mathcal{H}$ converge to a limit point that is also in $\mathcal{H}$.

Hilbert spaces

We are most used to dealing with Hilbert spaces when the elements h are d-dimensional real vectors, with

$$\langle h_1, h_2 \rangle = h_1^\top h_2 = \sum_{j=1}^{d} h_{1j} h_{2j}$$

for which the usual concepts of Euclidean distance etc apply.

Hilbert spaces

Consider the Hilbert space $\mathcal{H}_q$ of q-dimensional functions of random variable Z, with

$$\mathbb{E}[h(Z)] = \mathbf{0}_q \qquad \mathbb{E}[\{h(Z)\}^\top h(Z)] < \infty$$

for $h \in \mathcal{H}_q$, with the *covariance inner product*

$$\langle h_1, h_2 \rangle = \mathbb{E}[\{h_1(Z)\}^\top h_2(Z)] \qquad h_1, h_2 \in \mathcal{H}_q.$$

so that $\|h\|^2 = \langle h, h \rangle$. We say that h_1 and h_2 are *orthogonal* if

$$\langle h_1, h_2 \rangle \equiv \mathbb{E}[h_1^\top h_2] = 0.$$

In this formulation, all expectations are taken with respect to the distribution of random variable Z.

Subspaces and Projections

Consider arbitrary Hilbert space $\mathcal{H}$. For k linearly independent functions $h_1, \ldots, h_k \in \mathcal{H}$ consider the *linear subspace*, $\mathcal{U}$ of $\mathcal{H}$ defined by

$$\mathcal{U} \equiv \left\{ u : u = \sum_{j=1}^{k} a_j h_j, (a_1, \ldots, a_k) \in \mathbb{R}^k \right\}.$$

$\mathcal{U}$ is the subspace *spanned* by $h_1, \ldots, h_k$. Points in the subspace are identified uniquely by the coefficients

$$(a_1, \ldots, a_k).$$

Subspaces and Projections

For each $h \in \mathcal{H}$, there is a *unique value* $u_0 \in \mathcal{U}$ such that

$$\|h - u_0\| < \|h - u\| \qquad \forall u \in \mathcal{U}, u \neq u_0.$$

that is, u_0 is the nearest point in $\mathcal{U}$ to h.

- u_0 is the *projection* of h onto $\mathcal{U}$, sometimes denoted

$$\Pi(h|\mathcal{U});$$

- the element $h - u_0 \in \mathcal{H}$ is *orthogonal* to $\mathcal{U}$, that is for all $u \in \mathcal{U}$

$$\mathbb{E}[(h - u_0)^\top u] = 0.$$

- $\|h\|^2 = \|u_0\|^2 + \|h - u_0\|^2$.

Subspaces and Projections

Suppose that $v \equiv v(Z)$ is an r-dimensional function of Z with

$$\mathbb{E}[v(Z)] = \mathbf{0}_r \qquad \mathbb{E}[\{v(Z)\}^\top v(Z)] < \infty.$$

Let

$$\mathcal{U}_v \equiv \{\mathbf{B}v : \mathbf{B} \text{ an arbitrary } q \times r \text{ real matrix}\}$$

define a q-dimensional linear subspace of $\mathcal{H}_q$ derived from v.

That is, all elements of $\mathcal{U}_v$ can be expressed

$$\mathbf{B}v$$

for some (non-stochastic) matrix $\mathbf{B}$.

For $h \in \mathcal{H}_q$, the projection of h onto $\mathcal{U}_v$ must take the form $\mathbf{B}_0 v$, with $\mathbf{B}_0$ satisfying

$$\langle h - \mathbf{B}_0 v, \mathbf{B} v \rangle = 0 \qquad \forall \mathbf{B} \in \mathbb{R}^{q \times r}$$

that is,

$$\mathbb{E}[(h - \mathbf{B}_0 v)^\top \mathbf{B} v] = 0$$

Now, this equation can be re-written elementwise as

$$\sum_{i=1}^{q}\sum_{j=1}^{r} B_{ij}\mathbb{E}\left[(h-\mathbf{B}_0 v)_i v_j\right] = 0$$

where subscripting identifies elements of the vector/matrix.

As this must hold for all $\mathbf{B}$, it needs to hold for such matrices that have 1 in position (i,j) and zero elsewhere, which implies

$$\mathbb{E}\left[(h-\mathbf{B}_0 v)_i v_j\right] = 0 \qquad \forall (i,j)$$

and hence that

$$\mathbb{E}\left[(h-\mathbf{B}_0 v)v^\top\right] = \mathbf{0}.$$

Thus we may deduce that

$$\mathbf{B}_0 = \mathbb{E}[hv^\top]\left\{\mathbb{E}[vv^\top]\right\}^{-1} \qquad (\blacksquare)$$

so that

$$\mathbf{B}_0 v = \mathbb{E}[hv^\top]\left\{\mathbb{E}[vv^\top]\right\}^{-1} v. \qquad (\blacklozenge)$$

- $\mathbb{E}[hv^\top] \equiv \mathrm{Cov}_Z[h(Z), v(Z)]$ is $q \times r$;
- $\mathbb{E}[vv^\top] \equiv \mathrm{Var}_Z[v(Z)]$ is $r \times r$;
- it is assumed that $\mathbb{E}[vv^\top]$ is non-singular.

Parametric Inference

Suppose $Z_1, \ldots, Z_n$ are i.i.d. with pdf $f_Z(z;\theta)$ with $\theta \in \mathbb{R}^p$, and

$$\theta = \begin{bmatrix} \psi \\ \\ \beta \end{bmatrix} \qquad \begin{matrix} q \times 1 \\ \\ r \times 1 \end{matrix}$$

with true value θ_0 comprised of ψ_0 and β_0.

- ψ – parameter of interest;
- β – nuisance parameter.

The data generating model is thus $f_Z(z;\theta_0)$.

Step 1: Asymptotically Linear Estimators

An estimator $\widehat{\psi}_n$ of parameter ψ_0 is *asymptotically linear* if there exists a q-dimensional function, $\varphi(Z)$, with

$$\mathbb{E}[\varphi(Z)] = \mathbf{0}_q \qquad \mathbb{E}[\varphi(Z)\{\varphi(Z)\}^\top] < \infty, \text{ nonsingular}$$

such that

$$\sqrt{n}(\widehat{\psi}_n - \psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Z_i) + \mathrm{o}_p(1).$$

Step 1: Asymptotically Linear Estimators

Then as

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi(Z_i) \xrightarrow{d} \text{Normal}\left(\mathbf{0}_q, \mathbb{E}[\varphi(Z)\{\varphi(Z)\}^\top]\right)$$

$\sqrt{n}(\widehat{\psi}_n - \psi_0)$ also has this asymptotic distribution.

If such a $\varphi(.)$ exists, it is unique; it is termed the *influence function* for the estimator.

Step 1: Asymptotically Linear Estimators

Example: Sample mean

If $Z_1, \ldots, Z_n$ are a random sample from a population with finite mean θ and variance σ^2, then the estimator

$$\widehat{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i$$

yields the representation

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \theta_0)$$

so we can deduce the influence function

$$\varphi(Z) \equiv \varphi(Z, \theta_0) = Z - \theta_0.$$

Step 1: Asymptotically Linear Estimators

Example: Sample mean

By the Central Limit Theorem, we have that as $n \longrightarrow \infty$

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} Normal(0, \sigma^2)$$

where

$$\sigma^2 \equiv \mathbb{E}[(Z - \theta_0)^2] \equiv \mathbb{E}\left[\{\varphi(Z)\}^2\right].$$

Step 1: Asymptotically Linear Estimators

Recall that in *likelihood estimation*, we have the log-likelihood

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f_Z(z_i;\theta) = \sum_{i=1}^{n} \ell(z_i,\theta).$$

say. Under regularity conditions, by the *mean-value theorem*,

$$\dot{\ell}_n(\theta) = \dot{\ell}_n(\theta_0) + \ddot{\ell}_n(\theta')(\theta - \theta_0)$$

where $\|\theta' - \theta_0\| < \|\theta - \theta_0\|$, and where

$$\dot{\ell}_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} \qquad \ddot{\ell}_n(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta^\top}.$$

are $(p \times 1)$ and $(p \times p)$ respectively.

Step 1: Asymptotically Linear Estimators

Evaluating at $\theta = \widehat{\theta}_n$, and noting that $\dot{\ell}_n(\widehat{\theta}_n) = \mathbf{0}_p$, we have on rearrangement and multiplying through by $1/\sqrt{n}$ that

$$\left\{ -\frac{1}{n}\ddot{\ell}_n(\theta') \right\} \sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0)$$

where

$$\|\theta' - \theta_0\| < \|\widehat{\theta}_n - \theta_0\|$$

Step 1: Asymptotically Linear Estimators

As $n \longrightarrow \infty$, we have for the random variable version

$$\left\{ \frac{1}{n} \ddot{\ell}_n(\theta') \right\} \xrightarrow{p} \mathbb{E}\left[\left. \frac{\partial^2 \log f_Z(Z;\theta)}{\partial\theta\partial\theta^\top} \right|_{\theta=\theta_0} \right] = \mathbb{E}\left[\ddot{\ell}(Z,\theta_0) \right]$$

say, as

$$\widehat{\theta}_n \xrightarrow{p} \theta_0.$$

Step 1: Asymptotically Linear Estimators

Therefore

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \left\{-\mathbb{E}\left[\ddot{\ell}(Z, \theta_0)\right]\right\}^{-1} \frac{1}{\sqrt{n}} \dot{\ell}_n(\theta_0) + \mathrm{o}_p(1)$$

which yields the influence function

$$\begin{aligned}\varphi(Z) \equiv \varphi(Z, \theta_0) &= \left\{-\mathbb{E}\left[\ddot{\ell}(Z, \theta_0)\right]\right\}^{-1} \dot{\ell}(Z, \theta_0) \\ &= \mathcal{J}^{-1}(\theta_0)\dot{\ell}(Z, \theta_0)\end{aligned}$$

say.

Step 1: Asymptotically Linear Estimators

To obtain the asymptotic variance, we compute

$$\text{Var}[\varphi(Z)] = \mathcal{J}^{-1}(\theta_0)\text{Var}[\dot{\ell}(Z,\theta_0)]\mathcal{J}^{-1}(\theta_0)$$

where, under standard likelihood theory,

$$\mathcal{I}(\theta_0) = \text{Var}[\dot{\ell}(Z,\theta_0)] = \mathbb{E}\left[\dot{\ell}(Z,\theta_0)\{\dot{\ell}(Z,\theta_0)\}^\top\right] \equiv -\mathbb{E}\left[\ddot{\ell}(Z,\theta_0)\right]$$

that is $\mathcal{I}(\theta_0) = \mathcal{J}(\theta_0)$ so that

$$\text{Var}[\varphi(Z)] = \left\{\mathbb{E}\left[\dot{\ell}(Z,\theta_0)\{\dot{\ell}(Z,\theta_0)\}^\top\right]\right\}^{-1} = \mathcal{I}^{-1}(\theta_0)$$

which is the asymptotic variance of $\widehat{\theta}_n$.

Step 2: Regular Estimators

Consider the *local data generating process* (LDGP)

$$Z_{1n}, \ldots, Z_{nn} \sim f_Z(z; \theta_n) \qquad \text{i.i.d.}$$

with

$$\sqrt{n}(\theta_n - \theta^*) \longrightarrow \text{constant}$$

as $n \longrightarrow \infty$.

Estimator $\widehat{\psi}_n$ of ψ_n is *regular* if its limiting distribution does not depend on θ_n.

Step 2: Regular Estimators

That is, in the Normal case, $\widehat{\psi}_n$ is *regular* if

$$\sqrt{n}(\widehat{\psi}_n - \psi^*) \xrightarrow{d} Normal(0, \Sigma^*)$$

under $f_Z(z; \theta^*)$ implies that

$$\sqrt{n}(\widehat{\psi}_n - \psi_n) \xrightarrow{d} Normal(0, \Sigma^*)$$

under $f_Z(z; \theta_n)$.

Step 3: Score Functions

The *score function* $S_\theta(z, \theta_0)$ is defined by

$$S_\theta(z, \theta_0) = \frac{\partial}{\partial \theta} \{\log f_Z(z; \theta)\}_{\theta=\theta_0} = \begin{bmatrix} S_\psi(z, \theta_0) \\ S_\beta(z, \theta_0) \end{bmatrix} \begin{matrix} q \times 1 \\ r \times 1 \end{matrix}$$

with the subscript denoting the variable with respect to which the derivative is being taken.

Step 4: Key Theorem for RAL Estimators

Consider a q-dimensional parameter of interest $\Psi(\theta)$, and let

$$\Gamma(\theta) = \frac{\partial \Psi(\theta)}{\partial \theta^\top} \qquad q \times p.$$

Suppose that $\widehat{\Psi}_n$ is *regular and asymptotically linear* (RAL) with influence function $\varphi(.)$ such that $\mathbb{E}[\{\varphi(Z)\}^\top \varphi(Z)] < \infty$.

Then

$$\mathbb{E}[\varphi(Z)\{S_\theta(Z, \theta_0)\}^\top] = \Gamma(\theta_0).$$

Step 4: Key Theorem for RAL Estimators

Special case: If

$$\Psi(\theta) \equiv \psi$$

then

$$\Gamma(\theta_0) = \begin{bmatrix} \mathbf{I}_{q\times q} & \mathbf{0}_{q\times r} \end{bmatrix}$$

and so

(i) $\mathbb{E}[\varphi(Z)\{S_\psi(Z,\theta_0)\}^\top] = \mathbf{I}_{q\times q}$;

(ii) $\mathbb{E}[\varphi(Z)\{S_\beta(Z,\theta_0)\}^\top] = \mathbf{0}_{q\times r}$.

That is, $\varphi(Z)$ is *orthogonal* to $S_\beta(Z,\theta_0)$.

Step 5: m-estimation

In m-estimation, we replace the *score equation* for θ

$$\sum_{i=1}^{n} S_\theta(Z_i, \theta) \equiv \sum_{i=1}^{n} \dot{\ell}(Z_i, \theta) = \mathbf{0}_p$$

by the more general form

$$\sum_{i=1}^{n} m(Z_i, \theta) = \mathbf{0}_p$$

for function $m(.,.)$

Step 5: m-estimation

We must have

$$\text{(i)} \quad \mathbb{E}[m(Z,\theta)] = \mathbf{0}_p$$

$$\text{(ii)} \quad \mathbb{E}[\{m(Z,\theta)\}^\top m(Z,\theta)] < \infty$$

$$\text{(iii)} \quad \mathbb{E}[m(Z,\theta)\{m(Z,\theta)\}^\top] \text{ nonsingular}$$

for all possible data generating θ.

Step 5: m-estimation

Using the same logic as in the likelihood case, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_m(Z_i) + \mathrm{o}_p(1)$$

for the influence function associated with m

$$\varphi_m(Z) = \{-\mathbb{E}\left[\dot{m}(Z, \theta_0)\right]\}^{-1} m(Z, \theta_0).$$

Step 5: m-estimation

We have by the CLT and elementary results for the Normal distribution that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \longrightarrow Normal(\mathbf{0}_p, \mathcal{J}^{-1}\mathcal{I}\mathcal{J}^{-\top})$$

where

$$\mathcal{I} \equiv \mathcal{I}(\theta_0) = \mathbb{E}[m(Z, \theta_0)\{m(Z, \theta_0)\}^\top]$$

and

$$\mathcal{J} \equiv \mathcal{J}(\theta_0) = -\mathbb{E}[\dot{m}(Z, \theta_0)].$$

Step 5: m-estimation

These $p \times p$ matrices are typically estimated by

$$\widehat{I}_n = \frac{1}{n} \sum_{i=1}^{n} m(Z_i, \theta_0)\{m(Z_i, \theta_0)\}^\top \qquad \widehat{J}_n = -\frac{1}{n} \sum_{i=1}^{n} \dot{m}(Z_i, \theta_0)$$

with θ_0 replaced by estimator $\widehat{\theta}_n$.

Step 6: The Geometry of Influence Functions

Consider probability model $f_Z(z;\theta)$, $\theta^\top = (\psi^\top, \beta^\top)$.

- $\mathcal{H}_q$: Hilbert space of q-dimensional zero-mean functions with finite variance, with the covariance inner product.
- $\mathcal{T}$: linear subspace of $\mathcal{H}_q$ – the *tangent space*

$$\mathcal{T} \equiv \{\mathbf{B}S_\theta(z,\theta) : \mathbf{B} \text{ a } q \times p \text{ matrix}\}$$

  that is, the space spanned by the score S_θ.
- Λ: linear subspace of $\mathcal{H}_q$ – the *nuisance tangent space*

$$\Lambda \equiv \{\mathbf{B}S_\beta(z,\theta) : \mathbf{B} \text{ a } q \times r \text{ matrix}\}$$

  that is, the space spanned by the nuisance score S_β.

Step 6: The Geometry of Influence Functions

By the previous theorem, if $\widehat{\psi}_n$ is an RAL estimator of ψ with influence function $\varphi(Z) \in \mathcal{H}_q$, then we must have

$$\mathbb{E}\left[\varphi(Z)\left\{S_\beta(Z, \theta_0)\right\}^\top\right] = \mathbf{0}_{q \times r}$$

that is, $\varphi(Z)$ is *orthogonal to the nuisance tangent space*.

We write

$$\varphi(Z) \in \Lambda^\perp.$$

where

$$\Lambda^\perp \equiv \quad \text{“space orthogonal to } \Lambda\text{.”}$$

Step 6: The Geometry of Influence Functions

We have that

$$\mathcal{H}_q = \Lambda \oplus \Lambda^{\perp}$$

that is, any $h \in \mathcal{H}_q$ can be written

$$h = s_1 + s_2$$

for $s_1 \in \Lambda, s_2 \in \Lambda^{\perp}$, with s_1 orthogonal to s_2.

Step 6: The Geometry of Influence Functions

For any $h \in \mathcal{H}$, write s_0 for the projection of h onto Λ

$$\Pi(h|\Lambda) = s_0$$

that is, s_0 is the *unique* point in Λ that is at the "foot" of the perpendicular dropped from h onto Λ.

Then as $h = s_0 + (h - s_0)$, we can deduce that

$$\Pi(h|\Lambda^{\perp}) = h - s_0$$

by the projection theorem.

Step 6: The Geometry of Influence Functions

By the results on page 399, we know how to compute s_0 explicitly; we have that for any $h \in \mathcal{H}_q$

$$\Pi(h|\Lambda) = \mathbb{E}[hS_\beta^\top]\left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta$$

so therefore all elements in the space $\Lambda^\perp$ can be written

$$h - \Pi(h|\Lambda) = h - \mathbb{E}[hS_\beta^\top]\left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta.$$

Step 6: The Geometry of Influence Functions

Finally, as

$$\mathbf{B} S_\theta = \mathbf{B} \begin{bmatrix} S_\psi \\ S_\beta \end{bmatrix}$$

we can further decompose

$$\mathcal{T} = \mathcal{T}_\psi \oplus \Lambda$$

where

$$\mathcal{T}_\psi \equiv \{\mathbf{B}_1 S_\psi(Z, \theta_0) : \mathbf{B}_1 \text{ a } q \times q \text{ matrix}\}.$$

Step 7: Constructing an Estimating Function

Let influence function $\varphi(Z)$ satisfy the Theorem on page 413. Set

$$m(Z, \psi, \beta) = \varphi(Z) - \mathbb{E}[\varphi(Z)]$$

where the expectation is taken with respect to $f_Z(.; \psi, \beta)$. The function $m(Z, \psi, \beta)$ has mean zero and finite variance

Then $\widehat{\psi}_n$ satisfying

$$\sum_{i=1}^{n} m(Z_i, \widehat{\psi}_n, \widehat{\beta}_n(\widehat{\psi}_n)) = \mathbf{0}_p$$

is RAL with influence function $\varphi(Z)$.

Step 8: Efficient Influence Function

The *efficient influence function*, $\varphi^{\text{eff}}(Z)$, is the influence function with the smallest variance.

For an arbitrary influence function $\varphi(Z)$, using the projection theorem, it is evident that

$$\varphi^{\text{eff}}(Z) = \varphi(Z) - \Pi(\varphi(Z)|\mathcal{T}^{\perp}) \equiv \Pi(\varphi(Z)|\mathcal{T}).$$

Step 8: Efficient Influence Function

If we denote

$$\mathcal{I}(\theta_0) = \mathbb{E}\left[S_\theta(Z, \theta_0)\left\{S_\theta(Z, \theta_0)\right\}^\top\right]$$

we have explicitly that to estimate $\Psi(\theta_0)$

$$\varphi^{\text{eff}}(Z) = \Gamma(\theta_0)\left\{\mathcal{I}(\theta_0)\right\}^{-1} S_\theta(Z, \theta_0)$$

where, recall,

$$\Gamma(\theta) = \frac{\partial \Psi(\theta)}{\partial \theta^\top} \qquad q \times p.$$

Step 8: Efficient Influence Function

The *efficient score function* for ψ is obtained by projecting the score $S_\psi(Z, \theta_0)$ onto the nuisance tangent space Λ, and taking the residual.

By the result on page 399

$$\Pi(S_\psi|\Lambda) = \mathbb{E}[S_\psi S_\beta^\top] \left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta(Z, \theta_0)$$

so therefore the efficient score function for ψ is

$$S_\psi^{\text{eff}}(Z, \theta_0) = S_\psi(Z, \theta_0) - \mathbb{E}[S_\psi S_\beta^\top] \left\{\mathbb{E}[S_\beta S_\beta^\top]\right\}^{-1} S_\beta(Z, \theta_0)$$

Step 8: Efficient Influence Function

The *efficient influence function* for ψ is

$$\varphi_{\psi}^{\text{eff}}(Z) = \left\{ \mathbb{E}\left[S_{\psi}^{\text{eff}} \left\{ S_{\psi}^{\text{eff}} \right\}^{\top} \right] \right\}^{-1} S_{\psi}^{\text{eff}}(Z, \theta_0)$$

which has variance

$$\left\{ \mathbb{E}\left[S_{\psi}^{\text{eff}} \left\{ S_{\psi}^{\text{eff}} \right\}^{\top} \right] \right\}^{-1}$$

This result is the generalization of the earlier results.

Step 9: Semiparametric modelling

Consider the class of *semiparametric* models

$$\mathcal{P} \equiv \{f_Z(z;\psi,\beta(.)) : \psi \text{ is } q \times 1, \beta(.) \text{ is } \textit{infinite dimensional}\}$$

with true model $f_0(z) = f_Z(z;\psi_0,\beta_0(.))$.

$\beta(.)$ represents an unknown density function, say

Step 9: Semiparametric modelling

Consider the *parametric submodel*

$$\mathcal{P}_{\psi,\gamma} \equiv \{f_Z(z;\psi,\gamma) : \psi \text{ is } q \times 1, \gamma \text{ is } r \times 1\}$$

where

(i) $\mathcal{P}_{\psi,\gamma} \subset \mathcal{P}$;

(ii) $f_0(z) \in \mathcal{P}_{\psi,\gamma}$; that is

$$f_0(z) \equiv f_Z(z;\psi_0,\gamma_0).$$

The parametric submodel is *identical* to the *true model* for one setting of the parameters (ψ,γ).

Step 9: Semiparametric modelling

Note

The parametric submodel is simply a way to allow us to compute score functions explicitly.

Note that the parametric submodel is in general specified in terms of the true model.

Step 9: Semiparametric modelling

Example: Restricted Moment Model

The *restricted moment model*

$$Y_i = \mu(X_i; \psi) + \varepsilon_i$$

where $\beta(.)$ specifies the density of ε_i admits the parametric submodel where

$$\varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

that is, $\gamma \equiv \sigma^2$.

This parametric model is only a parametric submodel if we believe the data are normally distributed.

Step 9: Semiparametric modelling

Example: Proportional Hazards Model

In the *proportional hazards model* with hazard function

$$\lambda(t|X;\psi,\beta) = \beta(t)\exp\{\mathbf{x}\psi\}$$

with $\beta(.)$ nonparametrically specified, with true values of the parameters ψ_0 and $\beta_0(.)$.

Step 9: Semiparametric modelling

Example: Proportional Hazards Model

A parametric submodel takes the form

$$\lambda(t|X;\beta,\gamma) = \beta_0(t)\exp\{\gamma_1 g_1(t) + \cdots + \gamma_r g_r(t)\}\exp\{\mathbf{x}\psi\}$$

for specified functions $g_1(t),\ldots,g_r(t)$.

This parametric model is specified in terms of the *true model* $\beta_0(t)$ which is not known. However, if we specify $\psi = \psi_0$, and

$$\gamma_1 = \cdots = \gamma_r = 0$$

we recover the true model.

Step 10: Semiparametric inference

Consider the Hilbert space $\mathcal{H}_q$, and the parametric submodel.

(i) *nuisance tangent space*

$$\Lambda_\gamma \equiv \{\mathbf{B}S_\gamma(Z, \psi_0, \gamma_0) : \mathbf{B} \text{ a } q \times r \text{ matrix}\}$$

where

$$S_\gamma(Z, \psi_0, \gamma_0)$$

is the score function component corresponding to γ from the parametric submodel.

Step 10: Semiparametric inference

(ii) *efficient influence function*

$$\varphi^{\text{eff}}_{\psi,\gamma}(Z) = \{\mathcal{I}(\psi_0, \gamma_0)\}^{-1} \, S^{\text{eff}}_{\psi,\gamma}(Z, \psi_0, \gamma_0)$$

where

$$\mathcal{I}(\psi_0, \gamma_0) = \mathbb{E}\left[S^{\text{eff}}_{\psi,\gamma}(Z, \psi_0, \gamma_0) \left\{ S^{\text{eff}}_{\psi,\gamma}(Z, \psi_0, \gamma_0) \right\}^{\top} \right]$$

Step 10: Semiparametric inference

(iii) *efficient score function* for ψ is

$$S_\psi^{\text{eff}}(Z, \psi_0, \gamma_0) = S_\psi(Z, \psi_0, \gamma_0) - \Pi(S_\psi(Z, \psi_0, \gamma_0)|\Lambda_\gamma).$$

As ever, the RHS is defined in terms of (ψ_0, γ_0), which is identical to $(\psi_0, \beta_0(.))$ by assumption.

Step 10: Semiparametric inference

(iv) *smallest asymptotic variance* amongst RAL estimators for ψ is

$$\left\{ \mathbb{E}\left[S_{\psi}^{\text{eff}}(Z, \psi_0, \gamma_0) \left\{ S_{\psi}^{\text{eff}}(Z, \psi_0, \gamma_0) \right\}^{\top} \right] \right\}^{-1}$$

All expectations are taken with respect to the *true* model

$$f_Z(z; \psi_0, \gamma_0) \equiv f_Z(z; \psi_0, \beta_0).$$

Step 10: Semiparametric inference

- An estimator for ψ is RAL for the semiparametric model if it is RAL for *every* parametric submodel;
- Any influence function of an RAL estimator in the semiparametric model *must* be an influence function of an RAL estimator within a parametric submodel;
- Any influence function of an RAL estimator in the semiparametric model must be *orthogonal* to *all* parametric submodel nuisance tangent spaces;
- The variance of any RAL semiparametric influence function must be no smaller than the variance on page 440 computed for any such submodel.

Step 11: The Restricted Mean Model

Suppose scalar response Y follows the model

$$Y = \mu(X; \psi) + \varepsilon$$

with ψ a $q \times 1$ vector. Suppose $\mathcal{P} \equiv \{f_Z(z; \psi, \beta()), z = (y, x)\}$ with

$$f_{Y,X}(y, x) \equiv f_{\varepsilon,X}(y - \mu(x; \psi), x)$$

with the requirement $\mathbb{E}[\varepsilon \mid X] = 0$.

Step 11: The Restricted Mean Model

Write

$$f_{\varepsilon,X}(\varepsilon, \mathbf{x}) = \beta_1(\varepsilon, \mathbf{x})\beta_2(\mathbf{x})$$

where

- Conditional model:

$$\beta_1(\varepsilon, \mathbf{x}) \equiv f_{\varepsilon|X}(\varepsilon|\mathbf{x})$$

- Marginal model;

$$\beta_2(\mathbf{x}) \equiv f_X(\mathbf{x}).$$

Suppose the true (data generating) functions are

$$\beta_{10}(\varepsilon, \mathbf{x}) \qquad \beta_{20}(\mathbf{x}).$$

Step 11: The Restricted Mean Model

We require for all x

$$\int \beta_1(\varepsilon, \mathbf{x})\ d\varepsilon = 1 \qquad \int \varepsilon\, \beta_1(\varepsilon, \mathbf{x})\ d\varepsilon = 0$$

and that $\beta_2(x)$ is non-negative and satisfies

$$\int \beta_2(x)\ dx = 1.$$

With no further restrictions, we have a semiparametric specification with these infinite dimensional nuisance parameters.

Step 11: The Restricted Mean Model

A parametric submodel is

$$\mathcal{P}_{\psi,\gamma} \equiv \{f_Z(z;\psi,\gamma) = f_{\varepsilon|X}(y-\mu(x;\psi)|x;\gamma_1)f_X(x;\gamma_2)\}$$

where γ_1 is $r_1 \times 1$ and γ_2 is $r_2 \times 1$, with $r = r_1 + r_2$.

Denote the true model

$$f_0(z) = f_{\varepsilon|X}(y-\mu(x;\psi)|x;\gamma_{10})f_X(x;\gamma_{20}).$$

Step 12: Efficient Inference

We have in the parametric submodel

$$\log f_Z(z;\psi,\gamma) \equiv \log f_{\varepsilon|X}(\varepsilon|x;\gamma_1) + \log f_X(x;\gamma_2).$$

Therefore

$$S_{\gamma_1}(\varepsilon,\mathbf{x};\psi_0,\gamma_0) = \frac{\partial}{\partial\gamma_1}\left\{\log f_{\varepsilon|X}(\varepsilon|\mathbf{x};\gamma_1)\right\}_{\gamma_1=\gamma_{10}}$$

$$S_{\gamma_2}(\varepsilon,\mathbf{x};\psi_0,\gamma_0) = \frac{\partial}{\partial\gamma_2}\left\{\log f_X(\mathbf{x};\gamma_2)\right\}_{\gamma_2=\gamma_{20}}$$

We will suppress the dependence on (ψ_0,γ_0).

Step 12: Efficient Inference

A typical element in the parametric submodel nuisance tangent space is given by

$$\mathbf{B}S_\gamma(\varepsilon, X) = \mathbf{B}_1 S_{\gamma_1}(\varepsilon, X) + \mathbf{B}_2 S_{\gamma_2}(X)$$

where

- $\mathbf{B}_1$ is $q \times r_1$,
- $\mathbf{B}_2$ is $q \times r_2$.

We define the spaces Λ_γ, Λ_{γ_1} and Λ_{γ_2} by

$$\mathbf{B}S_\gamma(\varepsilon, X) \in \Lambda_\gamma \qquad \mathbf{B}_1 S_{\gamma_1}(\varepsilon, X) \in \Lambda_{\gamma_1} \qquad \mathbf{B}_2 S_{\gamma_2}(X) \in \Lambda_{\gamma_2}.$$

That is, in terms of the corresponding spaces

$$\Lambda_\gamma = \Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}$$

with Λ_{γ_1} and Λ_{γ_2} orthogonal as

$$\mathbb{E}\big[S_{\gamma_1}(\varepsilon, X)\{S_{\gamma_2}(X)\}^\top\big] = \mathbf{0}_{r_1 \times r_2}$$

by iterated expectation.

Step 12: Efficient Inference

Let

- Λ be the *mean-square closure* of *all* parametric submodel nuisance tangent spaces

$$\Lambda \equiv \{\text{mean-square closure of all } \Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}\};$$

- $\Lambda_{1s} = \{\text{mean-square closure of all } \Lambda_{\gamma_1}\};$
- $\Lambda_{2s} = \{\text{mean-square closure of all } \Lambda_{\gamma_2}\}$

For Λ, the mean-square closure is the set of functions that can be represented as the limit of sequences of score functions arising from the parametric submodels.

Step 12: Efficient Inference

- Λ_{2s} is a subspace of $\mathcal{H}_q$ arising from the unknown $f_X(x)$ (which is *essentially unrestricted*)
  - comprising elements that are eligible *score functions*;
  - where every bounded element $\alpha(x) \in \Lambda_{2s}$ is the score for some parametric submodel, say

$$f_X(x; \gamma_2) = f_0(x)(1 + \gamma_2^\top \alpha(x))$$

    for γ_2 small enough (the submodel density must be *non-negative* and integrate to one).

See Tsiatis, pp 78–79.

Step 12: Efficient Inference

- Λ_{1s} is a subspace of $\mathcal{H}_q$ comprising functions $a(\varepsilon, X)$
  - which satisfy
    (i) $\mathbb{E}[a(\varepsilon, X)|X] = \mathbf{0}_q$;
    This says that $a(\varepsilon, x)$ must be a *score function*.
    (ii) $\mathbb{E}[a(\varepsilon, X)\varepsilon|X] = \mathbf{0}_q$;
    This says that $a(\varepsilon, x)$ must be a *uncorrelated* with ε for all x, and enforces the requirement

$$\mathbb{E}[\epsilon|X] = 0 \qquad \text{w.p. } 1.$$

  - which arise from some parametric submodel, say

$$f_{\varepsilon|X}(\varepsilon|x;\gamma_1) = f_0(\varepsilon|x)(1 + \gamma_1^\top a(\varepsilon, x))$$

See Tsiatis, pp 80–81.

Step 12: Efficient Inference

Let

$$\alpha(X) \in \Lambda_{2s} \qquad \text{and} \qquad a(\varepsilon, X) \in \Lambda_{1s}.$$

Then

$$\mathbb{E}_{X,\varepsilon}[\alpha(X)^\top a(\varepsilon, X)] = \mathbb{E}_X\left[\alpha(X)^\top \mathbb{E}_{\varepsilon|X}[a(\varepsilon, X) \mid X]\right] = 0$$

by iterated expectation, as

$$\mathbb{E}_{\varepsilon|X}[a(\varepsilon, X) \mid X] = \mathbf{0}_q.$$

Therefore Λ_{1s} and Λ_{2s} are *orthogonal*, and we have precisely characterized the nuisance tangent space as

$$\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}.$$

Step 12: Efficient Inference

The space Λ_{1s} is a space of q-dimensional random functions, $a(\varepsilon, X)$, say, defined by two conditions:

$$\mathbb{E}[a(\varepsilon, X)|X] = \mathbf{0}_q \tag{C1}$$

$$\mathbb{E}[a(\varepsilon, X)\varepsilon|X] = \mathbf{0}_q \tag{C2}$$

To see this, consider the restrictions

$$\int f_{\varepsilon|X}(t|x;\gamma_1)\, dt = 1 \qquad \int t f_{\varepsilon|X}(t|x;\gamma_1)\, dt = 0$$

and then differentiate wrt γ_1 under the integral.

Step 12: Efficient Inference

Suppose $\alpha(X) \in \Lambda_{2s}$. Then, for any $a(\varepsilon, X)$ satisfying (C1), we have that

$$\mathbb{E}_{X,\varepsilon}[\alpha(X)^\top a(\varepsilon, X)] = 0$$

by iterated expectation. That is, $\alpha(X)$ is *orthogonal* to $a(\varepsilon, X)$.

Similarly, as

$$\mathbb{E}_{\varepsilon|X}[\alpha(X)\varepsilon|X] = \mathbf{0}_q$$

$\alpha(X)$ *also* satisfies (C2).

Step 12: Efficient Inference

It follows that the *nuisance tangent space* Λ comprises *precisely* those functions that satisfy (C2),

That is, Λ comprises functions $h \equiv h(\varepsilon, X) \in \mathcal{H}_q$ such that

$$\mathbb{E}[h(\varepsilon, X)\varepsilon|X] = \mathbf{0}_q.$$

See Tsiatis, pp 82–83.

Step 12: Efficient Inference

For the efficient score for ψ we need to *project* the ordinary score onto Λ, and take the '*residual*'

$$S_\psi^{\text{eff}}(\varepsilon, X) = S_\psi(\varepsilon, X) - \Pi(S_\psi(\varepsilon, X)|\Lambda).$$

By this construction, we see that

$$S_\psi^{\text{eff}}(\varepsilon, X) \in \Lambda^\perp$$

as

$$\begin{aligned} S_\psi(\varepsilon, X) &= \Pi(S_\psi(\varepsilon, X)|\Lambda) + \{S_\psi(\varepsilon, X) - \Pi(S_\psi(\varepsilon, X)|\Lambda)\} \\ &= \Pi(S_\psi(\varepsilon, X)|\Lambda) + S_\psi^{\text{eff}}(\varepsilon, X). \end{aligned}$$

Step 12: Efficient Inference

In the restricted mean model, the space *orthogonal to the nuisance tangent space* is seen to be

$$\Lambda^{\perp} = \{A(X)\varepsilon : \text{where } A \text{ is } q \times 1\}$$

To see this, note that for all $a(\varepsilon, X)$ satisfying condition (C2)

$$\mathbb{E}[\{a(\varepsilon, X)\}^{\top} A(X)\varepsilon] = 0$$

using iterated expectation.

Step 12: Efficient Inference

The projection of an arbitrary h onto $\Lambda^{\perp}$ is

$$\mathbb{E}[h\varepsilon|X]\left\{\mathbb{E}[\varepsilon^2|X]\right\}^{-1}\varepsilon.$$

We can then characterize elements of $\Lambda^{\perp}$ as taking the form

$$h(\varepsilon, X) - \Pi(h(\varepsilon, X)|\Lambda) = g(X)\varepsilon$$

for arbitrary $h \in \mathcal{H}$ and for $g(X)$ the q-dimensional vector

$$g(X) = \mathbb{E}[h\varepsilon|X]\left\{\mathbb{E}[\varepsilon^2|X]\right\}^{-1}$$

Taking $h(\varepsilon, X) \equiv S_{\psi}(\varepsilon, X)$, we obtain the efficient score for ψ.

Step 12: Efficient Inference

The efficient score is then

$$\begin{aligned} S_\psi^{\text{eff}}(\varepsilon, X) &= S_\psi(\varepsilon, X) - \Pi(S_\psi(\varepsilon, X)|\Lambda) \\ &\equiv \mathbb{E}[S_\psi(\varepsilon, X)\varepsilon|X] \left\{\mathbb{E}[\varepsilon^2|X]\right\}^{-1} \varepsilon \\ &= D(X)^\top \left\{V(X)\right\}^{-1} \varepsilon \end{aligned}$$

say, where

$$D(X) \equiv D(X; \psi_0) = \left.\frac{\partial \mu(X; \psi)}{\partial \psi^\top}\right|_{\psi=\psi_0}$$

is a $1 \times q$ vector.

Step 12: Efficient Inference

Hence we must (in principle) solve the estimating equation

$$\sum_{i=1}^{n} D(X_i;\psi)^\top \{V(X_i)\}^{-1} (Y_i - \mu(X_i;\psi)) = \mathbf{0}_q.$$

This requires knowledge of the true model $f_0(.)$, as the formula depends on

$$V(X) \equiv V(X;\psi_0,\beta_0) = \mathbb{E}[\varepsilon^2|X]$$

In practice we cannot implement the estimation procedure without further modelling.

Step 12: Efficient Inference

However, we can implement estimation based on

$$\sum_{i=1}^{n} A(X_i)(Y_i - \mu(X_i;\psi)) = \mathbf{0}_q.$$

where $A(.)$ is some pre-specified $q \times 1$ function of X.

If

$$A(X) = D(X;\psi_0)^\top \{V(X)\}^{-1}$$

we obtain optimal inference.

G-estimation

Consider the simple structural (causal) specification given by

$$\mathbb{E}[Y(z) - Y(0)|X] \equiv \mathbb{E}[Y(z) - Y(0)] = z\psi_0.$$

This states that

- compared to the 'baseline' case of $z = 0$, exposure at level $z = z$ leads to a change $z\psi_0$ in the expected (potential) outcome, and
- that this quantity does not depend on confounders X.

We proceed assuming Z is binary.

G-estimation

The model can be considered as one that specifies that

$$Y(z) = \mu_0(X) + z\psi_0 + \varepsilon$$

where $\mu_0(X) = \mathbb{E}[Y(0)|X]$ is a nuisance component.

Suppose a proposed parametric submodel is

$$\mu_0(X) = \mu_0(X; \beta_0).$$

G-estimation

We consider the probability model based on the semiparametric regression formulation

$$f_{X,Y,Z}(x,y,z) = f_{\varepsilon|X,Z}(y - \mu_0(x;\beta) - z\psi|x,z) f_{X,Z}(x,z)$$

and for simplicity focus on the parametric submodel for the first component

$$f_{\varepsilon|X,Z}(\varepsilon|x,z) \equiv f_{\varepsilon}(\varepsilon;\sigma_0) \equiv \text{Normal}(0,\sigma_0^2).$$

G-estimation

Then the score function for this submodel is

$$\begin{pmatrix} S_\beta \\ S_\psi \end{pmatrix} = \frac{1}{\sigma_0^2} \begin{pmatrix} \dot{\mu}_{0\beta}(X;\beta_0) \\ Z \end{pmatrix} (Y - \mu_0(X;\beta_0) - Z\psi_0)$$

$$= \frac{1}{\sigma_0^2} \begin{pmatrix} \dot{\mu}_{0\beta}(X;\beta_0) \\ Z \end{pmatrix} \varepsilon$$

say.

G-estimation

The nuisance tangent space corresponds to the S_β component, and can be seen to take the form

$$\{A(X)\varepsilon : A(X) \text{ arbitrary}\}$$

where $A(X)$ has the same dimension as ψ: in the previous notation we have that $A(X) = \mathbf{B}\dot{\mu}_{0\beta}(X;\beta_0)$ with $\mathbf{B}$ arbitrary.

We then need to project S_ψ onto this tangent space: the projection is the quantity $A_0(X)\varepsilon$, where we must have

$$\mathbb{E}[(Z\varepsilon - A_0(X)\varepsilon)A(X)\varepsilon] = 0 \qquad \forall A(X)$$

G-estimation

By iterated expectation, if

$$e(X) \equiv \mathbb{E}[Z|X]$$

we therefore must have

$$\mathbb{E}[(e(X)\varepsilon - A_0(X)\varepsilon)A(X)\varepsilon] = 0 \qquad \forall A(X)$$

implying that

$$A_0(X) \equiv e(X).$$

Therefore the projection of the score S_ψ onto the nuisance tangent space yields the efficient score function

$$S_\psi^{\text{eff}} = S_\psi - \Pi(S_\psi|\Lambda) = (Z - e(X))\varepsilon.$$

G-estimation

Therefore efficient estimation is achieved by solving

$$\sum_{i=1}^{n} \begin{pmatrix} \dot{\mu}_{0\beta}(\mathbf{x}_i;\beta) \\ z_i - e(\mathbf{x}_i) \end{pmatrix} (y_i - \mu_0(\mathbf{x}_i;\beta) - z_i\psi) = \mathbf{0}$$

which corresponds to G-estimation.

Inverse Probability Weighting

Weighting estimators can be justified using semiparametric theory by adopting a *missing* (or *coarsened*) data strategy.

Recall that for the ATE in the binary case, using the potential outcome notation

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

Inverse Probability Weighting

If such data were available, we would use the estimator

$$\widehat{\mu}(z) = \frac{1}{n} \sum_{i=1}^{n} Y_i(z).$$

based on the *complete data* estimating equation

$$\sum_{i=1}^{n} (Y_i(z) - \mu(z)) = 0.$$

However, we do not observe the counterfactual outcomes.

Inverse Probability Weighting

We need to utilize the *observed data* estimating equation:

$$\sum_{i=1}^{n} \left(\frac{\mathbb{1}_{\{z\}}(Z_i) Y_i}{f_{Z|X}(Z_i|X_i)} - \mu(z) \right) = 0.$$

With $z = 1$, this becomes

$$\sum_{i=1}^{n} \left(\frac{Z_i Y_i}{e(X_i)} - \mu(1) \right) = 0.$$

Inverse Probability Weighting

To construct the efficient (observed data) score and influence function, it can be shown first that all observed data influence functions can be written

$$h(X, Y, Z) + L(X, Y, Z)$$

where $h(X, Y, Z)$ satisfies

$$\mathbb{E}[h(X, Y, Z)|Y(0), Y(1), X] = Y(1) - \mu(1)$$

and $L(X, Y, Z)$ satisfies

$$\mathbb{E}[L(X, Y, Z)|Y(0), Y(1), X] = 0.$$

The function L is termed the *augmentation function*.

Inverse Probability Weighting

From above, one suitable h is

$$h(X, Y, Z) = \frac{ZY}{e(X)} - \mu(1)$$

as

$$\begin{aligned}
\mathbb{E}\left[\frac{ZY}{e(X)}\middle| Y(0), Y(1), X\right] &= \mathbb{E}\left[\frac{ZY(1)}{e(X)}\middle| Y(0), Y(1), X\right] \\
&= Y(1)\mathbb{E}\left[\frac{Z}{e(X)}\middle| Y(0), Y(1), X\right] \\
&= Y(1).
\end{aligned}$$

Inverse Probability Weighting

For $L(X, Y, Z)$ we can always write in the binary case

$$L(X, Y, Z) = (1 - Z)L_0(X, Y) + ZL_1(X, Y)$$

Therefore, taking expectations

$$\mathbb{E}[L(X, Y, Z)|Y(0), Y(1), X] = (1 - e(X))L_0(X, Y(0)) + e(X)L_1(X, Y(1))$$

Inverse Probability Weighting

Equating this to zero implies that

$$L_0(X, Y(0)) = -\frac{e(X)}{(1 - e(X))} L_1(X, Y(1)).$$

provided $0 < e(X) < 1$.

Further, as the left hand side is a function of $Y(0)$ and the right hand side is a function of $Y(1)$, this equation can only hold in general if L_0 and L_1 do not depend on Y at all.

Inverse Probability Weighting

Thus we can simplify

$$L_0(X) = -\frac{e(X)}{(1-e(X))} L_1(X).$$

so that

$$\begin{aligned} L(X,Y,Z) \equiv L(X,Z) &= \left(-(1-Z)\frac{e(X)}{(1-e(X))} + Z \right) L_1(X) \\ &= \left(\frac{Z-e(X)}{1-e(X)} \right) L_1(X). \end{aligned}$$

Inverse Probability Weighting

Combining terms in X, we see that the space of augmentation functions takes the form

$$\Lambda = \{(Z - e(X))g(X) : g(X) \text{ arbitrary}\}.$$

We find the optimal influence function by projecting

$$h(X, Y, Z) = \frac{ZY}{e(X)} - \mu(1)$$

onto Λ; this identifies a specific element of Λ

$$(Z - e(X))g_0(X)$$

say that obeys the usual orthogonality results.

Inverse Probability Weighting

That is, for arbitrary g, we need to solve for g_0 the condition

$$\mathbb{E}\left[\left(\frac{ZY}{e(X)}-\mu(1)-(Z-e(X))g_0(X)\right)(Z-e(X))g(X)\right]=0.$$

The expectation can be rewritten

$$\mathbb{E}_{X,Z}\left[\left(\frac{Z\mu(X,Z)}{e(X)}-\mu(1)-(Z-e(X))g_0(X)\right)(Z-e(X))g(X)\right]$$

Inverse Probability Weighting

Using iterated expectation, we have conditional on X that the interior expectation of Z is

$$(1-e(X))(-\mu(1)+e(X)g_0(X))(-e(X))g(X)$$

$$+e(X)\left(\frac{\mu(X,1)}{e(X)}-\mu(1)-(1-e(X))g_0(X)\right)(1-e(X))g(X).$$

To make this identically zero in expectation for any $g(X)$, we must have

$$e(X)(1-e(X))\left(\frac{\mu(X,1)}{e(X)}-g_0(X)\right)=0.$$

Inverse Probability Weighting

We find that

$$g_0(X) = \frac{\mu(X,1)}{e(X)} = \frac{1}{e(X)}\mathbb{E}[Y|X,Z=1]$$

so that

$$h(X,Y,Z)+L(X,Z) = \frac{ZY}{e(X)} - \frac{(Z-e(X))}{e(X)}\mathbb{E}[Y|X,Z=1] - \mu(1).$$

Inverse Probability Weighting

Therefore, the efficient influence function for $\mu(1)$ is

$$\varphi^{\text{eff}}(X,Y,Z) = \frac{ZY}{e(X)} - \frac{(Z-e(X))}{e(X)}\mu(X,1) - \mu(1)$$

or equivalently

$$\varphi^{\text{eff}}(X,Y,Z) = \frac{Z(Y-\mu(X,1))}{e(X)} + \mu(X,1) - \mu(1)$$

which is the basis of AIPW estimation.

Inverse Probability Weighting

The corresponding estimating equation is therefore

$$\sum_{i=1}^{n}\left(\frac{Z_i Y_i}{1-e(X_i)}-\frac{(Z_i-e(X_i))}{e(X_i)}\mu(X_i,1)-\mu(1)\right)=0.$$

Similarly for $\mu(0)$ we have

$$\sum_{i=1}^{n}\left(\frac{(1-Z_i)Y_i}{e(X_i)}+\frac{(Z_i-e(X_i))}{1-e(X_i)}\mu(X_i,0)-\mu(0)\right)=0.$$

as

$$(1-Z_i)-(1-e(X_i))=-(Z_i-e(X_i)).$$

Part 6
Extensions

Extensions

The development above has mainly focussed on

- binary treatment,
- linear models,
- average treatment effects,
- single time-point studies,

but extensions can be developed to handle each case.

Continuous treatments

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

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

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

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

We may also need to adjust using further balancing models, such as

$$b_2(X) = \mathbb{E}^{\mathcal{O}}_{Z|X}[Z^2|X]$$

depending on the treatment effect model.

Continuous treatments

Example: G-estimation

The 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 $\mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X]$

- justified by a focus on *orthogonality* under the *covariance* inner product;
- focus on conditional *uncorrelatedness* rather than conditional *independence*.

Continuous treatments

For IPW estimation, the construction

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

also works in the continuous case as

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

but in practice this estimator can be variable in finite sample.

Continuous treatments

The focus on $\mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X]$ emphasizes the role of *regression* approaches to constructing the treatment model: we denote

$$\xi(X) \equiv \xi(X;\alpha) = \mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X].$$

In most cases, some form of *generalized linear model* will suffice, with estimating equation such as

$$\sum_{i=1}^{n} \mathbf{x}_i^\top (z_i - \xi(\mathbf{x}_i\alpha)) = \mathbf{0}$$

for parameter α so that the fitted values

$$\xi(\mathbf{x}_i\widehat{\alpha}) \quad i = 1, \ldots, n$$

can be computed.

Continuous treatments

In any subsequent procedure, these fitted values are used in estimating the causal effects: for example, in *G-estimation*, we consider the *plug-in* estimator

$$\widehat{\delta}_{\text{G}} = \frac{\sum_{i=1}^{n} Y_i(Z_i - \xi(\mathbf{X}_i\widehat{\alpha}))}{\sum_{i=1}^{n} Z_i(Z_i - \xi(\mathbf{X}_i\widehat{\alpha}))}$$

In IPW estimation

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

The application of many of the causal adjustment methods rely on regression modelling for

- the outcome mean model

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

- the propensity or balancing score

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

Advanced modelling

In either case we can utilize

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

to construct fitted versions of each model.

Advanced modelling

The advanced methods can be effective, but there are potential pitfalls:

1. The quantification of *uncertainty*;
    - no ready analytic answers, typically relies on bootstrap;
    - large computational burden
2. *Positivity violations*.
    - prediction (of treatment mechanism) is not the fundamental goal;
    - can be overcome using methods that target balance/overlap.

Beyond linear models

If the outcome variable Y is discrete, or it is contended that the outcome mean model

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X = \mathbf{x}, Z = z] = \mu(\mathbf{x}, z; \beta, \psi)$$

is not linear (in the treatment or parameters), it is necessary to extend some of the previous concepts.

In the following,

- X is the random vector of confounders, and x its observed values,
- $\mathbf{x}$ is the vector of functions of x that form the linear predictor,
- $\mathbf{X}$ is the random variable version of $\mathbf{x}$.

Beyond linear models

Suppose that a *log-linear* model is deemed appropriate:

$$\log \mu(x, z; \beta, \psi) = \mathbf{x}_\beta \beta + z\mathbf{x}_\psi \psi$$

Regarding this as the structural model, we then have that

$$\mathbb{E}[Y(z)] \equiv \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \mathbb{E}^{\mathcal{E}}_{X}[\exp\{\mathbf{X}_\beta \beta + z\mathbf{X}_\psi \psi\}]$$

which would then invoke the estimator

$$\widetilde{\mu}(z) = \frac{1}{n} \sum_{i=1}^{n} \exp\{\mathbf{x}_{\beta i} \widehat{\beta} + z\mathbf{x}_{\psi i} \widehat{\psi}\}.$$

where $(\widehat{\beta}, \widehat{\psi})$ are estimated from a correctly specified model.

Beyond linear models

Then in the binary treatment case

$$\frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} = \frac{\mathbb{E}_X^{\mathcal{E}}[\exp\{\mathbf{X}_\beta\beta + \mathbf{X}_\psi\psi\}]}{\mathbb{E}_X^{\mathcal{E}}[\exp\{\mathbf{X}_\beta\beta\}]}$$

but notice that unlike in the linear case

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

depend on *both* β and ψ. Also

$$\frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} \neq \mathbb{E}\left[\frac{Y(1)}{Y(0)}\right]$$

in general.

Beyond linear models

The above model assumes that

$$\mu(\mathbf{x}, 1; \beta, \psi) = \mu(\mathbf{x}, 0; \beta, \psi) \exp\{\mathbf{x}_\psi \psi\}.$$

with the effect of Z represented *conditional* on X; *marginally*, ψ alone does not capture the effect of treatment.

We could formulate a model for the potential outcomes where

$$Y(1) = Y(0) \exp\{\mathbf{x}_\psi \psi\}$$

that is

$$\log Y(1) - \log Y(0) = \mathbf{x}_\psi \psi.$$

Beyond linear models

If we can estimate β and ψ consistently, then we can recover APO and ATE estimators from them. By standard regression arguments, we know that *correct specification* of the outcome model is needed.

For the log-linear model,

$$\log \mu(x, z; \beta, \psi) = \mathbf{x}_\beta \beta + z\mathbf{x}_\psi \psi$$

the standard estimating equation takes the form

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^\top \\ z_i \mathbf{x}_{\psi i}^\top \end{pmatrix} (y_i - \exp\{\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi\}) = \mathbf{0}.$$

Beyond linear models

We might try to make inference robust to mis-specification using a *G-estimation*-like strategy, and modify the estimating equation to be

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^{\top} \\ (z_i - e(x_i))\mathbf{x}_{\psi_i}^{\top} \end{pmatrix} (y_i - \exp\{\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi\}) = \mathbf{0}$$

where $e(x_i)$ is the propensity score; if necessary, this can be estimated using a further parametric model

$$\sum_{i=1}^{n} \mathbf{x}_i^{\top}(z_i - e(x_i;\alpha)) = \mathbf{0}.$$

Beyond linear models

In the *linear* case, this modification led to consistent estimation of ψ as the resulting estimating equation is *unbiased*, that is

$$\mathbb{E}_{X,Y,Z}[(Z - e(X))\mathbf{X}_\psi^\top(Y - \mathbf{X}_\beta\beta - Z\mathbf{X}_\psi\psi)] = \mathbf{0}$$

provided $\mathbf{x}_\psi\psi$ correctly captures the effect of treatment, *and either*

- propensity score $e(X)$, *or*
- treatment-free mean component $\mathbf{x}_\beta\beta$

is correctly specified.

Beyond linear models

In the *log-linear* case, the equivalent requirement would be

$$\mathbb{E}_{X,Y,Z}[(Z - e(X))\mathbf{X}_\psi^\top (Y - \exp\{\mathbf{X}_\beta \beta + Z\mathbf{x}_\psi \psi)\}] = \mathbf{0}.$$

However, this requirement is *not* met if the treatment-free mean component is mis-specified, even if the propensity score model is correctly specified.

Beyond linear models

Suppose in reality

$$\mathbb{E}_{Y|X,Z}[Y|X = \mathbf{x}, Z = z] = \mu(\mathbf{x}, z) = \mu_0(\mathbf{x}) \exp\{z\mathbf{x}_\psi\psi\}.$$

Then

$$\begin{aligned}\mathbb{E}_{Y|X,Z}[Y - \exp\{\mathbf{x}_\beta\beta + Z\mathbf{x}_\psi\psi)\}|X = \mathbf{x}, Z = z] \\ = \mu_0(\mathbf{x}) \exp\{z\mathbf{x}_\psi\psi\} - \exp\{\mathbf{x}_\beta\beta + z\mathbf{x}_\psi\psi)\} \\ = \exp\{z\mathbf{x}_\psi\psi\}(\mu_0(\mathbf{x}) - \exp\{\mathbf{x}_\beta\beta\})\end{aligned}$$

which cannot be made independent of z unless

$$\mu_0(\mathbf{x}) \equiv \exp\{\mathbf{x}_\beta\beta\}.$$

Beyond linear models

We must modify the estimating function to be

$$\varphi(X, Y, Z) = (Z - e(X))\mathbf{X}_\psi^\top \exp\{-Z\mathbf{X}_\psi\psi\}(Y - \exp\{\mathbf{X}_\beta\beta + Z\mathbf{X}_\psi\psi\})$$

so that

$$\begin{aligned}
&\mathbb{E}_{Y|X,Z}[\varphi(X, Y, Z)|X = x, Z = z] \\
&\quad = (z - e(x))\mathbf{x}_\psi^\top \exp\{-z\mathbf{x}_\psi\psi\} \exp\{z\mathbf{x}_\psi\psi\}(\mu_0(x) - \exp\{\mathbf{x}_\beta\beta\}) \\
&\quad = (z - e(x))\mathbf{x}_\psi^\top(\mu_0(x) - \exp\{\mathbf{x}_\beta\beta\})
\end{aligned}$$

It then follows that

$$\begin{aligned}&\mathbb{E}_{Z|X}[\mathbb{E}_{Y|X,Z}[\varphi(X,Y,Z)|X=\mathbf{x},Z=z]\mid X=\mathbf{x}]\\&\quad=\mathbf{x}_\psi^\top(\mu_0(\mathbf{x})-\exp\{\mathbf{x}_\beta\beta\})\mathbb{E}_{Z|X}[(Z-e(X))|X=\mathbf{x}]=\mathbf{0}\end{aligned}$$

if the propensity model is correctly specified. Thus for consistent estimation, we need to solve

$$\sum_{i=1}^{n}\begin{pmatrix}\mathbf{x}_{\beta i}^\top\\(z_i-e(\mathbf{x}_i))\mathbf{x}_{\psi_i}^\top\exp\{-z_i\mathbf{x}_{\psi i}\psi\}\end{pmatrix}(y_i-\exp\{\mathbf{x}_{\beta i}\beta+z_i\mathbf{x}_{\psi i}\psi\})=\mathbf{0}$$

Beyond linear models

If outcome Y is *binary*, we might attempt to use a *logistic* model and estimate using

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^{\top} \\ z_i \mathbf{x}_{\psi_i}^{\top} \end{pmatrix} (y_i - \text{expit}\{\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi\}) = \mathbf{0}.$$

where

$$\text{expit}(x) = \frac{e^x}{1 + e^x}.$$

This method is not robust to mis-specification of the mean model, and it cannot be rescued using the previous trick.

Beyond linear models

A solution is found by noting that

$$\exp\{\mathbf{x}_\psi \psi\} = \frac{\Pr[Y=1|Z=1,X=x]/\Pr[Y=0|Z=1,X=x]}{\Pr[Y=1|Z=0,X=x]/\Pr[Y=0|Z=0,X=x]}$$

$$\equiv \frac{\Pr[Z=1|Y=1,X=x]/\Pr[Z=0|Y=1,X=x]}{\Pr[Z=1|Y=0,X=x]/\Pr[Z=0|Y=0,X=x]}$$

that is, ψ parameterizes the conditional *log-odds ratio*.

Beyond linear models

Therefore we can construct a doubly robust estimator assuming that **either**

- the model for Y given X and Z, **or**
- the model for Z given X and Y

is correctly specified, given that the log-odds model is correctly specified.

The R package `drgee` uses the estimating procedure based on first estimating parameters in the two systems

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\gamma i}^{\top} \\ y_i \mathbf{x}_{\psi i}^{\top} \end{pmatrix} \left(z_i - \text{expit}(\mathbf{x}_{\gamma i}\gamma + y_i \mathbf{x}_{\psi i}\psi^{\dagger}) \right) = \mathbf{0}$$

and

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^{\top} \\ z_i \mathbf{x}_{\psi i}^{\top} \end{pmatrix} \left(y_i - \text{expit}(\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi^{\ddagger}) \right) = \mathbf{0}$$

utilizing two additional nuisance parameters $\psi^{\dagger}$ and $\psi^{\ddagger}$. This yields estimates of β and γ.

Then ψ is estimated using

$$\sum_{i=1}^{n}(z_i - e^*(\mathrm{x}_i;\psi,\widehat{\beta},\widehat{\gamma}))\mathbf{x}_{\psi i}^\top \left(y_i - \text{expit}(\mathbf{x}_{\beta i}\widehat{\beta} + z_i\mathbf{x}_{\psi i}\psi)\right) = \mathbf{0}$$

where

$$e^*(\mathrm{x};\psi,\widehat{\beta},\widehat{\gamma}) = \left[1 + \frac{1-\text{expit}(\mathbf{x}_\gamma\widehat{\gamma})}{\text{expit}(\mathbf{x}_\gamma\widehat{\gamma})}\frac{\text{expit}(\mathbf{x}_\beta\widehat{\beta})}{\text{expit}(\mathbf{x}_\beta\widehat{\beta} + \mathbf{x}_\psi\psi)}\right]^{-1}.$$

It can be shown that this is an unbiased estimating equation.

Average Treatment Effect on the Treated

In the binary case, we have mainly focussed on the estimation of the average treatment effect (ATE)

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y \mid Z = 1] - \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y \mid Z = 0]$$

The *average treatment effect on the treated* (ATT) is

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

using the potential outcome notation.

Average Treatment Effect on the Treated

That is, the ATT aims to identify the causal effect on intervening to change $Z = 0$ to $Z = 1$ but *only* in the subpopulation of individuals who are *observed* to receive treatment

Note that

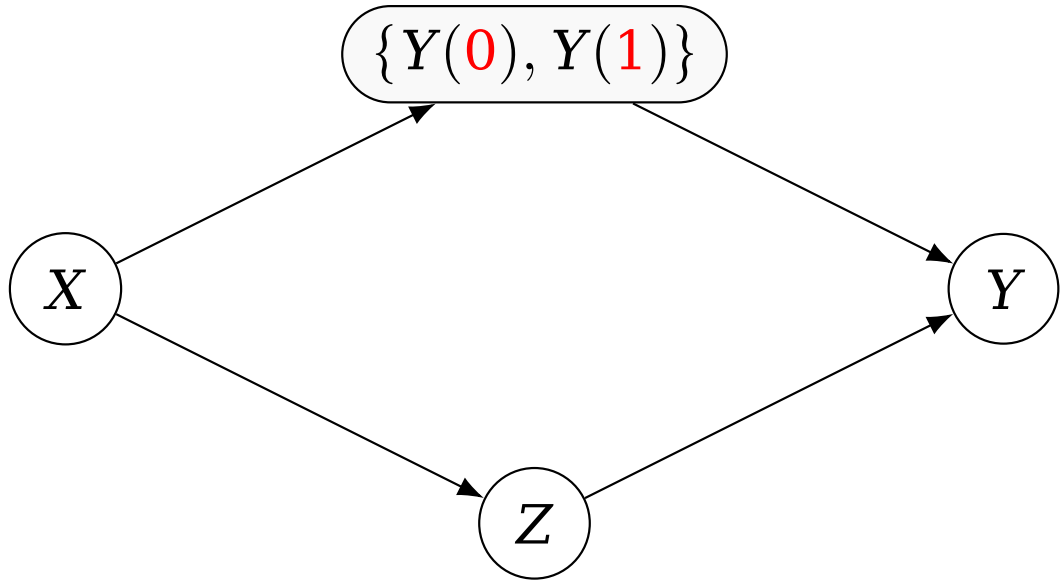
$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1) - Y(0) \mid Z = 0]\Pr[Z = 0]$$
$$+ \mathbb{E}[Y(1) - Y(0) \mid Z = 1]\Pr[Z = 1]$$

Average Treatment Effect on the Treated

In this calculation, we imagine

- the observational distribution $f^{\mathcal{O}}_{X,Y,Z}$ generating the observed data $\{(x_i, y_i, z_i), i = 1, \ldots, n\}$
- in the subgroup observed to have $Z = 1$, we then consider a second (hypothetical) experimental intervention to change Z to z which over-rides the original Z if $z = 0$,
- we then consider comparison of hypothetical outcomes between the two hypothetical subgroups indexed by z.

Average Treatment Effect on the Treated



$$f_X(x)\, f_{Z|X}(z|x)\, f_{Y(0),Y(1)|X}(y_0, y_1|x)\, f_{Y|Z,Y(0),Y(1)}(y|z, y_0, y_1)$$

defines the observational distribution, where

$$f_{Y|Z,Y(0),Y(1)}(y|z, y_0, y_1) = \begin{cases} 1 & y = (1-z)y_0 + zy_1 \\ 0 & \text{otherwise} \end{cases}.$$

Average Treatment Effect on the Treated

Then

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

$$= \int f_{Y|X,Z,Y(0),Y(1)}(y|x,z,y_0,y_1) f_{Y(0),Y(1)|X,Z}(y_0,y_1|x,z)\ dy_0\ dy_1$$

$$= \int f_{Y|Z,Y(0),Y(1)}(y|z,y_0,y_1) f_{Y(0),Y(1)|X}(y_0,y_1|x)\ dy_0\ dy_1$$

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

as the integral reduces to a evaluation at a single point where (y_0, y_1) satisfy

$$y = (1-z)y_0 + zy_1.$$

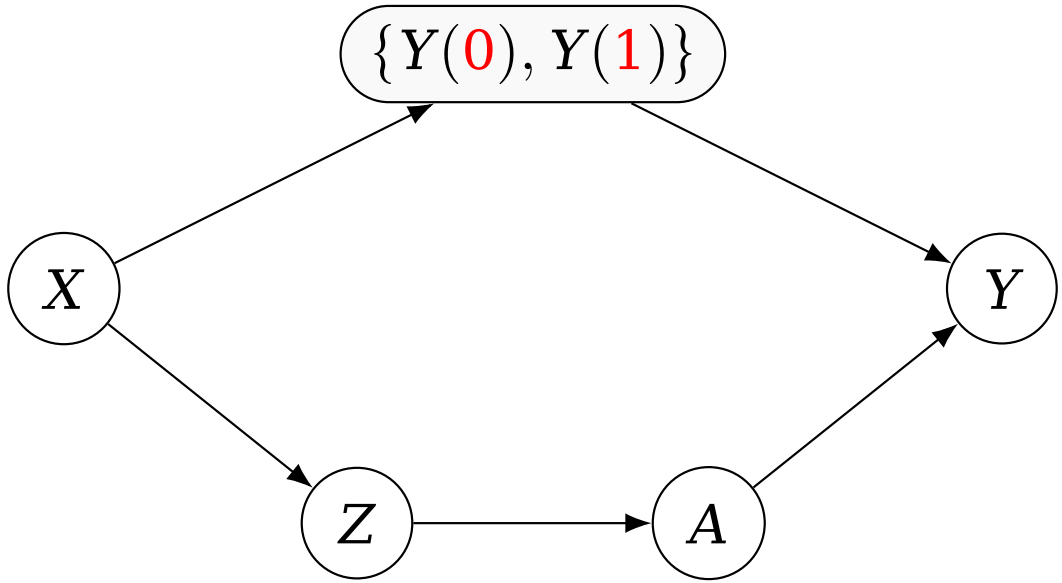
Average Treatment Effect on the Treated

For the ATT, we can represent the quantity of interest using a modified DAG that proposes a second hypothetical binary treatment, A.

We allow Z to cause A, and then allow Z to act as a selection mechanism, but ensure that

$$X \perp\!\!\!\perp A \mid Z.$$

Average Treatment Effect on the Treated



where the implied joint model is

$$f_X(x) f_{Z|X}(z|x) f_{Y(0),Y(1)|X}(y_0, y_1|x) f_{A|Z}(a|z) f_{Y|A,Y(0),Y(1)}(y|a, y_0, y_1)$$

This is the new 'experimental' distribution $\mathcal{E}$.

Average Treatment Effect on the Treated

Again

$$f^{\mathcal{E}}_{Y|A,Y(0),Y(1)}(y|a,y_0,y_1) = \begin{cases} 1 & y = (1-a)y_0 + ay_1 \\ 0 & \text{otherwise} \end{cases} .$$

As A is binary, the model $f_{A|Z}(a|z)$ must take the form

$$f^{\mathcal{E}}_{A|Z}(a|z) = p_z^a(1-p_z)^{1-a} \qquad a,z \in \{0,1\}$$

for $0 \leqslant p_z \leqslant 1$ for $z = 0, 1$.

We can then express the ATT via the new DAG as

$$\mathbb{E}^{\mathcal{E}}_{Y|A,Z}[Y|A=1,Z=1] - \mathbb{E}^{\mathcal{E}}_{Y|A,Z}[Y|A=0,Z=1]$$

Average Treatment Effect on the Treated

We have that from the DAG that

$$f^{\mathcal{E}}_{Y|A,X,Z}(y|a,x,z) \equiv f^{\mathcal{E}}_{Y|A,X}(y|a,x)$$

and hence as before

$$\begin{aligned} & f^{\mathcal{E}}_{Y|A,X}(y|a,x) \\ &= \int f^{\mathcal{E}}_{Y|A,Y(0),Y(1)}(y|a,y_0,y_1) f^{\mathcal{E}}_{Y(0),Y(1)|A,X}(y_0,y_1|a,x)\ dy_0\ dy_1 \\ &= f^{\mathcal{E}}_{Y(a)|X}(y|x). \end{aligned}$$

Also from the DAG, $X \perp\!\!\!\perp A \mid Z$, so for all a, x, z

$$f^{\mathcal{E}}_{X|A,Z}(x|a,z) \equiv f^{\mathcal{E}}_{X|Z}(x|z).$$

Average Treatment Effect on the Treated

For $\mathsf{a} = 0, 1$, we therefore have

$$\mathbb{E}^{\mathcal{E}}_{Y|A,Z}[Y \mid A = \mathsf{a}, Z = z]$$

$$= \iint y \, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x) f^{\mathcal{E}}_{X|Z}(x \mid z) \, dy \, dx.$$

Note that there is a potential *incompatibility* in the conditioning between

$$f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x) \qquad \text{and} \qquad f^{\mathcal{E}}_{X|Z}(x \mid z)$$

when we try to write the integral in terms of the data generating mechanism.

Average Treatment Effect on the Treated

As before, choosing the form of

$$f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x)$$

is to be avoided if possible.

We seek to resolve the incompatibility using the importance sampling trick, and write the expectation with respect to the observational model

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

Average Treatment Effect on the Treated

First note that

$$f^{\mathcal{E}}_{X|Z}(\mathrm{x} \mid z) = \frac{f^{\mathcal{E}}_{Z|X}(z \mid \mathrm{x}) f^{\mathcal{E}}_{X}(\mathrm{x})}{f^{\mathcal{E}}_{Z}(z)}$$

so the integral can be rewritten

$$\frac{1}{f^{\mathcal{E}}_{Z}(z)} \iint y \, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, \mathrm{x}) f^{\mathcal{E}}_{Z|X}(z \mid \mathrm{x}) f^{\mathcal{E}}_{X}(\mathrm{x}) \, dy \, dx.$$

Average Treatment Effect on the Treated

For $\mathsf{z} = 0, 1$, we can re-write the integrand using the importance sampling trick as

$$y\, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x) \frac{f^{\mathcal{E}}_{Z|X}(z \mid x)}{f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x)}\, f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x) f^{\mathcal{E}}_{X}(x)$$

which can be rearranged to

$$\left\{ y\, \frac{f^{\mathcal{E}}_{Z|X}(z \mid \mathrm{x})}{f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid \mathrm{x})} \right\}\, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, \mathrm{x})\, f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid \mathrm{x}) f^{\mathcal{E}}_{X}(\mathrm{x}),$$

Average Treatment Effect on the Treated

Comparing the observational and experimental DAGs, we see that for all x and z

$$f^{\mathcal{E}}_{Z|X}(z \mid x) \equiv f^{\mathcal{O}}_{Z|X}(z \mid x) \qquad f^{\mathcal{E}}_{X}(x) \equiv f^{\mathcal{O}}_{X}(x) \qquad f^{\mathcal{E}}_{Z}(z) \equiv f^{\mathcal{O}}_{Z}(z).$$

Also, we have for any t and y that

$$f^{\mathcal{E}}_{Y|A,X}(y \mid t, x) \equiv f^{\mathcal{O}}_{Y|X,Z}(y \mid x, t).$$

Therefore we have

$$f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x)\, f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x) f^{\mathcal{E}}_{X}(x) \equiv f^{\mathcal{O}}_{Y|X,Z}(y \mid x, \mathsf{a})\, f^{\mathcal{O}}_{Z|X}(\mathsf{a} \mid x) f^{\mathcal{O}}_{X}(x).$$

Average Treatment Effect on the Treated

Thus

$$\mathbb{E}^{\mathcal{E}}_{Y|A,Z}[Y \mid A = \mathsf{a}, Z = z]$$

$$= \frac{1}{f^{\mathcal{O}}_{Z}(z)} \iint \left\{ y \frac{f^{\mathcal{O}}_{Z|X}(z \mid x)}{f^{\mathcal{O}}_{Z|X}(\mathsf{a} \mid x)} \right\} f^{\mathcal{O}}_{X,Y,Z}(x, y, \mathsf{a})\, dy\, dx$$

$$= \frac{1}{f^{\mathcal{O}}_{Z}(z)} \iiint \left\{ \mathbb{1}_{\{\mathsf{a}\}}(t) y \frac{f^{\mathcal{O}}_{Z|X}(z \mid x)}{f^{\mathcal{O}}_{Z|X}(t \mid x)} \right\} f^{\mathcal{O}}_{X,Y,Z}(x, y, t)\, dy\, dx\, dt$$

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

Average Treatment Effect on the Treated

For the ATT, we are interested only in $z = 1$. The moment-based estimator is therefore

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|A,Z}[Y \mid A = \mathsf{a}, Z = 1] = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\mathsf{a}\}}(Z_i) w_1(X_i, Z_i) Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\{1\}}(Z_i)}$$

where

$$w_z(X_i, Z_i) = \frac{f^{\mathcal{O}}_{Z|X}(z \mid X_i)}{f^{\mathcal{O}}_{Z|X}(Z_i \mid X_i)}$$

Average Treatment Effect on the Treated

When $\mathsf{a} = 1$,

$$\mathbb{1}_{\{\mathsf{a}\}}(Z_i) w_1(X_i, Z_i) = \mathbb{1}_{\{1\}}(Z_i) = Z_i \quad \text{w.p. } 1$$

as the weight is identically 1, so therefore

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|A,Z}[Y \mid A = 1, Z = 1] = \frac{\sum_{i=1}^{n} Z_i Y_i}{\sum_{i=1}^{n} Z_i}$$

that is, the mean in the treated group.

Average Treatment Effect on the Treated

When $\mathsf{a} = 0$,

$$\mathbb{1}_{\{\mathsf{a}\}}(Z_i) w_1(X_i, Z_i) = \mathbb{1}_{\{0\}}(Z_i) \frac{f^{\mathcal{O}}_{Z|X}(1 \mid X_i)}{f^{\mathcal{O}}_{Z|X}(Z_i \mid X_i)}$$

$$= (1 - Z_i) \frac{f^{\mathcal{O}}_{Z|X}(1 \mid X_i)}{f^{\mathcal{O}}_{Z|X}(0 \mid X_i)}.$$

Average Treatment Effect on the Treated

Therefore

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|A,Z}[Y \mid A = 0, Z = 1] = \frac{\sum_{i=1}^{n} (1 - Z_i) w(X_i) Y_i}{\sum_{i=1}^{n} Z_i}$$

where

$$w(X_i) = \frac{f^{\mathcal{O}}_{Z|X}(1 \mid X_i)}{f^{\mathcal{O}}_{Z|X}(0 \mid X_i)} = \frac{e(X_i)}{1 - e(X_i)}$$

That is, this estimator is a weighted sum of contributions from the *untreated* individuals.

Average Treatment Effect on the Treated

Thus the estimator for the ATT is

$$\frac{\sum_{i=1}^{n} \left(Z_i - (1 - Z_i) w(X_i)\right) Y_i}{\sum_{i=1}^{n} Z_i}.$$

Under the standard assumptions, this estimator is consistent for the ATT and asymptotically normally distributed if $e(x)$ is correctly specified; that is, it is *singly robust*.

Average Treatment Effect on the Treated

To achieve *double robustness*, we proceed in the usual fashion and augment the estimand for $\mathsf{a} = 0$ as follows:

$$\mathbb{E}[Y(0) \mid Z = 1] = \mathbb{E}[Y(0) - \mu(X, 0)|Z = 1] + \mathbb{E}[\mu(X, 0)|Z = 1]$$

where

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

is the modelled conditional mean for Y.

Average Treatment Effect on the Treated

To estimate the first term, we use

$$\widehat{\mathbb{E}}[Y(0) - \mu(X, 0) \mid Z = 1] = \frac{\sum_{i=1}^{n} (1 - Z_i) w(X_i)(Y_i - \mu(X_i, 0))}{\sum_{i=1}^{n} Z_i}$$

as in the singly robust case. For the second term, we have

$$\widehat{\mathbb{E}}[\mu(X, 0) \mid Z = 1] = \frac{\sum_{i=1}^{n} Z_i \mu(X_i, 0)}{\sum_{i=1}^{n} Z_i}$$

Average Treatment Effect on the Treated

Therefore

$$\widehat{\mathbb{E}}[Y(0) \mid Z = 1]$$

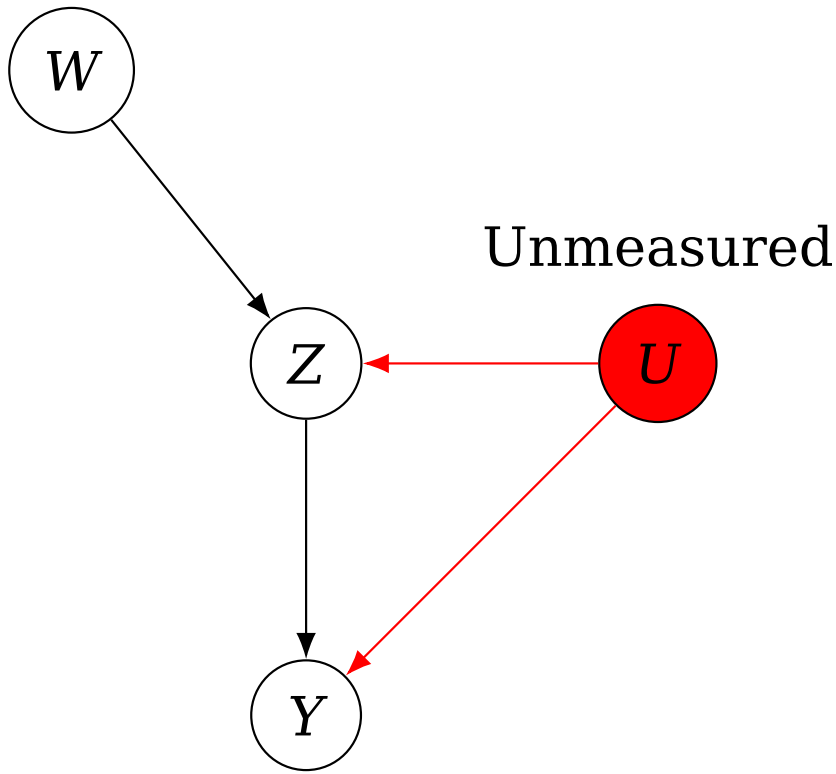
$$= \frac{\sum_{i=1}^{n} (1 - Z_i) w(X_i)(Y_i - \mu(X_i, 0)) + Z_i \mu(X_i, 0)}{\sum_{i=1}^{n} Z_i}$$

which yields the augmented ATT estimator

$$\frac{\sum_{i=1}^{n} (Z_i - (1 - Z_i) w(X_i))(Y_i - \mu(X_i, 0))}{\sum_{i=1}^{n} Z_i}.$$

Instrumental variables

Suppose we have the following system:



where $U \perp\!\!\!\perp W$, and $Y \perp\!\!\!\perp W \mid Z$.

Instrumental variables

We aim to find out the causal effect of Z on Y (both scalars).

- There is an open biasing path ZUY
- We cannot condition on U to block this path as it is unmeasured.

Instrumental variables

Suppose the structural model underlying this system is linear

$$Y_i = \beta_0 + Z_i\psi_0 + \mathbf{u}_i\zeta + \varepsilon_i$$

where

$$\mathbb{E}[\varepsilon_i|\mathbf{u}_i] = 0$$

and for all i, $Z_i \perp\!\!\!\perp \varepsilon_i$.

Clearly, regressing Y on Z alone will lead to inconsistent estimation of ψ_0.

Instrumental variables

Suppose now that W (also scalar) is observed and we in fact have the two-equation structural system

$$Y_i = \beta_0 + Z_i\psi_0 + \mathbf{u}_i\zeta + \varepsilon_i$$

$$Z_i = \gamma_0 + W_i\gamma_1 + \epsilon_i$$

with ϵ_i uncorrelated with W_i and ε_i. Then, by the usual trick, we have

$$W_iY_i = W_i\beta_0 + W_iZ_i\psi_0 + W_i\mathbf{u}_i\zeta + W_i\varepsilon_i$$

$$W_iZ_i = W_i\gamma_0 + W_i^2\gamma_1 + W_i\epsilon_i$$

Instrumental variables

Taking expectations and standardizing, using the usual arguments from linear regression, we have

$$\text{Cov}[W, Y] = \psi_0 \text{Cov}[W, Z]$$

$$\text{Cov}[W, Z] = \gamma_1 \text{Var}[W]$$

as W_i and U_i are *independent* by assumption.

Hence on rearrangement we have that

$$\psi_0 = \frac{\text{Cov}[W, Y]}{\text{Cov}[W, Z]} = \frac{\text{Cov}[W, Y]/\text{Var}[W]}{\text{Cov}[W, Z]/\text{Var}[W]} = \frac{\text{Cov}[W, Y]/\text{Var}[W]}{\gamma_1}.$$

Instrumental variables

This suggests a procedure to estimate ψ_0:

(i) regress Z on W to obtain estimate $\widehat{\gamma}_1$

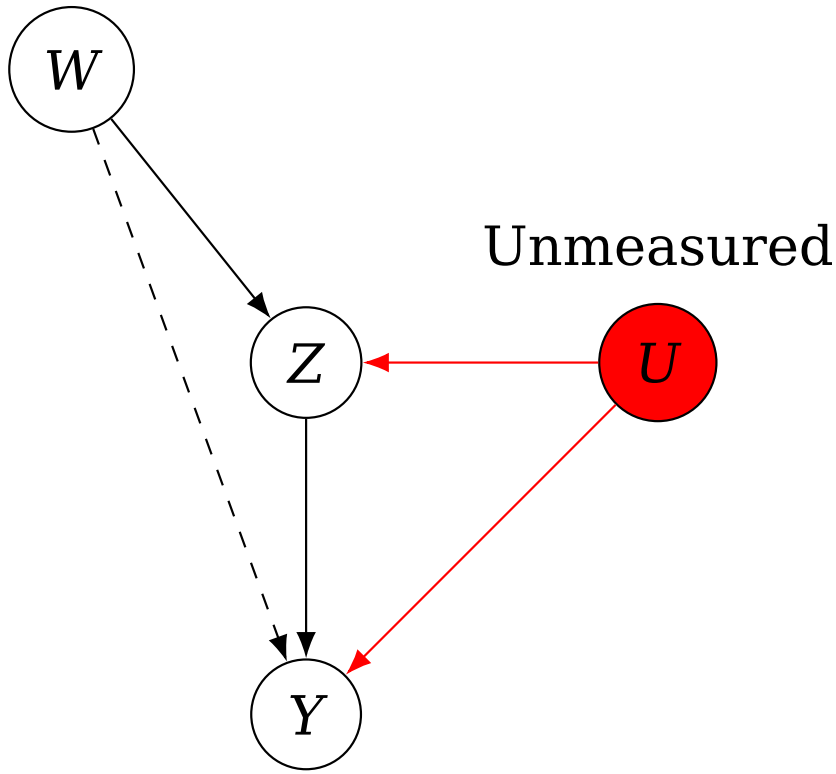
(ii) regress Y on W to obtain estimate $\widehat{\lambda}_1$

(iii) form the estimate of ψ_0 as

$$\widehat{\psi}_0 = \frac{\widehat{\lambda}_1}{\widehat{\gamma}_1}$$

This is a an *instrumental variable ratio estimator* using the *instrument* W.

Instrumental variables



Instrumental variables

Note

- For this method to work, we cannot have an *interaction* between Z and U in the structural model
- If there is only *weak association* between W and Z, so that γ_1 is close to zero, this estimator can have undesirable properties.
- Statistical properties of $\widehat{\psi}_0$ not trivial to derive.

Instrumental variables

A common situation in which the linear model is not quite appropriate is when W and Z are both *binary*. For example, we might have in a randomized study

- W is the randomized *allocated* treatment indicator; eg

$$W = 1 \quad \Longrightarrow \quad \text{“allocated to treatment arm”}$$

- Z is the actual treatment *received* indicator; eg

$$Z = 1 \quad \Longrightarrow \quad \textit{“took treatment”}$$

- Could have $W = 1, Z = 0$ (*non-compliance*);
- In some situations, could have $W = 0$, $Z = 1$.

Instrumental variables

Then we have by iterated expectation, for $w = 0, 1$

$$\begin{aligned}\mathbb{E}[Y|W=w] =& \mathbb{E}[Y|W=w, Z=0]\Pr[Z=0|W=w] \\ &+ \mathbb{E}[Y|W=w, Z=1]\Pr[Z=1|W=w] \\ =& \mathbb{E}[Y|Z=0]\Pr[Z=0|W=w] \\ &+ \mathbb{E}[Y|Z=1]\Pr[Z=1|W=w]\end{aligned}$$

as $Y \perp\!\!\!\perp W \mid Z$.

Instrumental variables

Therefore

$$\mathbb{E}[Y|W=1] = \mathbb{E}[Y|Z=0]\Pr[Z=0|W=1] + \mathbb{E}[Y|Z=1]\Pr[Z=1|W=1]$$

$$\mathbb{E}[Y|W=0] = \mathbb{E}[Y|Z=0]\Pr[Z=0|W=0] + \mathbb{E}[Y|Z=1]\Pr[Z=1|W=0]$$

Instrumental variables

Hence subtracting the second equation from the first

$$\begin{aligned}
&\mathbb{E}[Y|W=1] - \mathbb{E}[Y|W=0] \\
&\quad = \mathbb{E}[Y|Z=0](\Pr[Z=0|W=1] - \Pr[Z=0|W=0]) \\
&\quad + \mathbb{E}[Y|Z=1](\Pr[Z=1|W=1] - \Pr[Z=1|W=0])
\end{aligned}$$

Instrumental variables

But

$$\Pr[Z = 0|W = 1] - \Pr[Z = 0|W = 0]$$

$$= \Pr[Z = 1|W = 0] - \Pr[Z = 1|W = 1]$$

Therefore

$$\mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]$$

$$= (\Pr[Z = 1|W = 1] - \Pr[Z = 1|W = 0])$$

$$\times (\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0])$$

Instrumental variables

That is, provided $\Pr[Z = 1|W = 1] \neq \Pr[Z = 1|W = 0]$

$$\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]$$

$$= \frac{\mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]}{\Pr[Z = 1|W = 1] - \Pr[Z = 1|W = 0]}$$

or equivalently

$$\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0] = \frac{\mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]}{\mathbb{E}[Z|W = 1] - \mathbb{E}[Z|W = 0]}.$$

Instrumental variables

As W is observed, we can deploy the moment-based estimator

$$\frac{\dfrac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i} - \dfrac{\sum_{i=1}^{n} (1 - W_i) Y_i}{\sum_{i=1}^{n} (1 - W_i)}}{\dfrac{\sum_{i=1}^{n} W_i Z_i}{\sum_{i=1}^{n} W_i} - \dfrac{\sum_{i=1}^{n} (1 - W_i) Z_i}{\sum_{i=1}^{n} (1 - W_i)}}$$

This is the binary analogue of the earlier regression-based ratio estimator.

Instrumental variables

In terms of potential outcomes, we can consider potential outcomes for treatment received

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

corresponding to two settings $\{0, 1\}$ of W, and potential outcomes for outcome

$$\{Y(\mathsf{w}, \mathsf{z}), \mathsf{w} = 0, 1, \mathsf{z} = 0, 1\}$$

corresponding to the four possible combinations of W and Z.

Instrumental variables

We have the "consistency" requirement relating the potential outcomes to the observed data:

$$Z = (1 - W)Z(0) + W\ Z(1)$$

and

$$Y = \sum_{z=0}^{1} \sum_{w=0}^{1} \mathbb{1}_{\{w\}}(W)\mathbb{1}_{\{z\}}(Z)Y(w, z).$$

Instrumental variables

We must have that for $z = 0, 1$

$$Y(1, z) = Y(0, z)$$

as $Y \perp\!\!\!\perp W | Z$ by assumption. We have *ignorability*

$$\{Y(w, z), Z(w) : \forall\ w, z\} \perp\!\!\!\perp Z.$$

Consequently

$$Y(0, 0) = Y(1, 0) \equiv Y(0)$$

$$Y(0, 1) = Y(1, 1) \equiv Y(1)$$

which simplifies the causal quantity of interest.

Instrumental variables

We are interested in the contrast

$$\mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(W, 1) - Y(W, 0)].$$

We have that

$$\begin{aligned} Y &= (1 - Z)Y(0) + ZY(1) \\ &= Y(0) + (Y(1) - Y(0))Z \end{aligned}$$

which suggests the linear model for the observed data

$$Y_i = \beta_0 + \psi_0 Z_i + \varepsilon_i$$

Instrumental variables

We also need that $W \not\perp\!\!\!\perp Z$, that is

$$\mathbb{E}[Z(1)] \neq \mathbb{E}[Z(0)] \qquad \text{or} \qquad \Pr[Z(1) = 1] \neq \Pr[Z(0) = 1]$$

- Manipulation of W *does* change Z;
- From the assumptions and the DAG, we can state

$$\mathbb{E}[Z(1) - Z(0)] \equiv \mathbb{E}[Z|W = 1] - \mathbb{E}[Z|W = 0].$$

Instrumental variables

Recall the earlier formula

$$\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]=\frac{\mathbb{E}[Y|W=1]-\mathbb{E}[Y|W=0]}{\mathbb{E}[Z|W=1]-\mathbb{E}[Z|W=0]}.$$

We now have that the denominator on the right hand side is non-zero.

For the numerator, we have

$$\begin{aligned}\mathbb{E}[Y|W=1] &= \mathbb{E}[Y(0)+(Y(1)-Y(0))Z|W=1]\\ &= \mathbb{E}[Y(0)+(Y(1)-Y(0))Z(1)]\end{aligned}$$

$$\mathbb{E}[Y|W=0] = \mathbb{E}[Y(0)+(Y(1)-Y(0))Z(0)]$$

Instrumental variables

Thus, taking the difference we have

$$\mathbb{E}[Y|W=1]-\mathbb{E}[Y|W=0]=\mathbb{E}[(Y(1)-Y(0))(Z(1)-Z(0))]$$

so therefore

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

We now try to understand the causal implications of this identity.

Instrumental variables

Stratum	$Z(1)$	$Z(0)$	Description
Compliers	1	0	*Follows* treatment protocol, *takes assigned treatment* w.
Defiers	0	1	*Defies* treatment protocol, takes *other* treatment $1 - w$.
Always Takers	1	1	*Ignores* treatment protocol, takes treatment 1.
Never Takers	0	0	*Ignores* treatment protocol, takes treatment 0.

Instrumental variables

In the *observed* data:

W	Z	Possible strata
1	0	Defier, Never Taker
0	1	Defier, Always Taker
1	1	Complier, Always Taker
0	0	Complier, Never Taker

We can never conclude with certainty to which stratum an individual belongs from the observed data.

Instrumental variables

It is common (and reasonable in most cases) to assume that

$$\Pr[Z(1) \geqslant Z(0)] = \Pr[Z(1) - Z(0) \geqslant 0] = 1$$

that is, being assigned $W = 1$ *cannot decrease* the value of $Z(\mathsf{w})$ compared with being assigned $W = 0$.

i.e. being allocated treatment increases or leaves unchanged the received treatment status variable.

This *monotonicity assumption* rules out the possibility of the *Defier* stratum.

Instrumental variables

Assuming $\Pr[Z(1) - Z(0) \geqslant 0] = 1$

W	Z	Possible strata
1	0	~~Defier~~, Never Taker
0	1	~~Defier~~, Always Taker
1	1	Complier, Always Taker
0	0	Complier, Never Taker

Instrumental variables

To get a further simplification in the expression

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

we use iterated expectation using the partitioning event

$$Z(1) > Z(0)$$

which is an event occurs if and only if

$$Z(1) = 1 \qquad \textit{and} \qquad Z(0) = 0.$$

Instrumental variables

We have in the numerator

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

$$= \mathbb{E}[(Y(1) - Y(0))(Z(1) - Z(0))|Z(1) > Z(0)]\Pr[Z(1) > Z(0)]$$

$$+ \mathbb{E}[(Y(1) - Y(0))(Z(1) - Z(0))|Z(1) \leqslant Z(0)]\Pr[Z(1) \leqslant Z(0)]$$

and similarly in the denominator

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

$$= \mathbb{E}[(Z(1) - Z(0))|Z(1) > Z(0)]\Pr[Z(1) > Z(0)]$$

$$+ \mathbb{E}[(Z(1) - Z(0))|Z(1) \leqslant Z(0)]\Pr[Z(1) \leqslant Z(0)]$$

Instrumental variables

Stratum	$Z(1)$	$Z(0)$	$Z(1) - Z(0)$
Compliers	1	0	1
Defiers	0	1	-1
Always Takers	1	1	0
Never Takers	0	0	0

Instrumental variables

We therefore have in the numerator:

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

$$= \mathbb{E}[(Y(1) - Y(0))|Z(1) > Z(0)]\Pr[Z(1) > Z(0)]$$

as only the contribution $Z(1) = 1$ and $Z(0) = 0$ remains.

Similarly in the denominator:

$$\mathbb{E}[(Z(1) - Z(0))] = \Pr[Z(1) > Z(0)]$$

Instrumental variables

Therefore

$$\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]=\frac{\mathbb{E}[(Y(1)-Y(0))(Z(1)-Z(0))]}{\mathbb{E}[Z(1)-Z(0)]}$$

$$\equiv \mathbb{E}[(Y(1)-Y(0))|Z(1)>Z(0)].$$

It is therefore evident that the IV estimation procedure estimates the causal contrast

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

Instrumental variables

As

$$Z(1) > Z(0) \quad \Longleftrightarrow \quad Z(1) = 1, Z(0) = 0$$

the quantity

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

is termed the

- *complier average treatment effect* (CATE), or
- *complier average causal effect* (CACE), or
- *local average treatment effect* (LATE),

as $Z(1) > Z(0)$ identifies the *Complier* stratum.

Note

- If we observe $Z = 1$, the subject could be a Complier or Always Taker; we do not know who the Compliers are.
- The Compliers may be different depending on the instrument W chosen.
- In general the three quantities

$$\begin{aligned} \text{ATE} &: \mathbb{E}[(Y(1) - Y(0))] \\ \text{ATT} &: \mathbb{E}[(Y(1) - Y(0))|Z = 1] \\ \text{CATE} &: \mathbb{E}[(Y(1) - Y(0))|Z(1) > Z(0)] \end{aligned}$$

  are not equal.

Note

- It is often in practice realistic to make the assumption

$$\Pr[Z = 1|W = 0] = 0$$

which asserts that the Always Takers stratum is empty, and that all subjects for whom $Z = 1$ are Compliers.

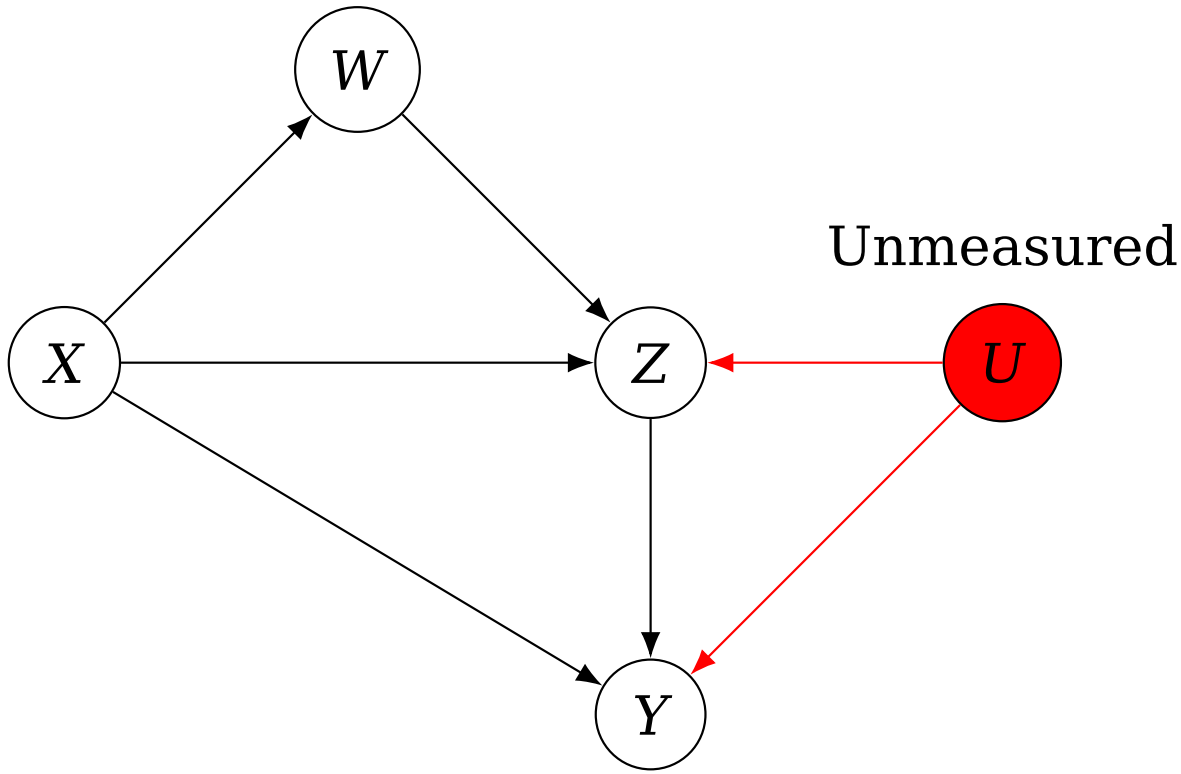
It follows that

$$\frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\Pr[Z = 1|W = 1]} = \mathbb{E}[(Y(1) - Y(0))|Z = 1]$$

which equates to the ATT.

Instrumental variables

Extension: with measured confounders



We have $U \perp\!\!\!\perp W$, and $Y \perp\!\!\!\perp W \mid Z, X$.

Instrumental variables

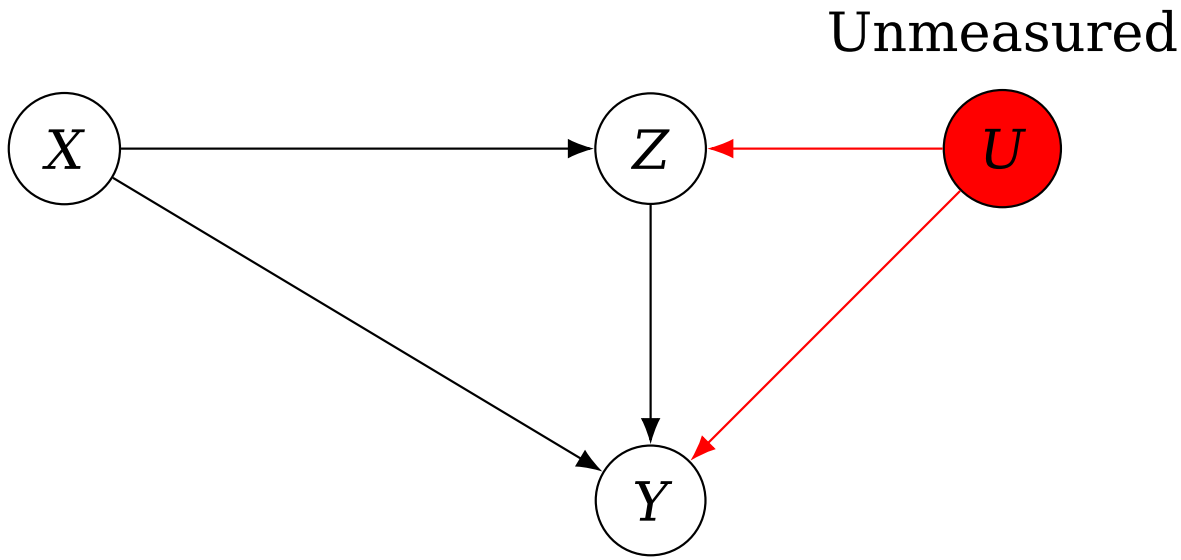
We can extend the previous logic by appealing to *conditional* ignorability

$$\{Y(\mathsf{w}, \mathsf{z}), Z(\mathsf{w}) : \forall\ \mathsf{w}, \mathsf{z}\} \perp\!\!\!\perp Z|X$$

and develop similar methods.

Instrumental variables

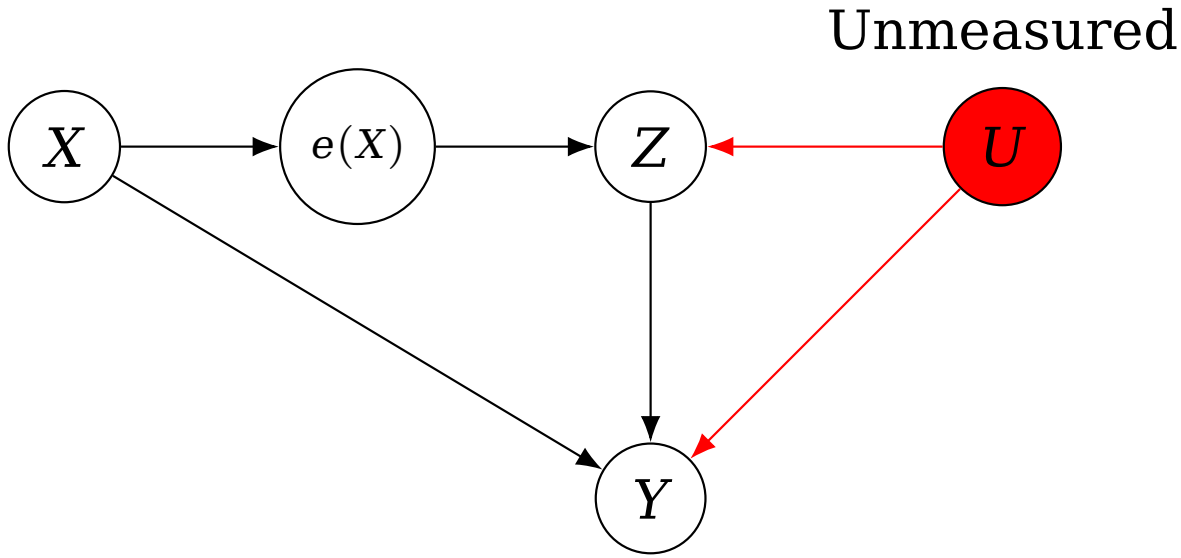
Suppose we have the following confounded system:



No adjustment method is feasible.

Instrumental variables

Including the propensity score, we have



This does not directly help with unmeasured confounding.

Instrumental variables

However, consider a variable W, with $\mathbb{E}[W] = 0$, such that

$$Z = e(X) + W$$

so that

$$W = Z - e(X) = Z - \mathbb{E}[Z|X]$$

We have

$$\mathbb{E}[W\ e(X)] = \mathbb{E}[(Z - e(X))e(X)] = \mathbb{E}[(e(X) - e(X))e(X)] = 0$$

by iterated expectation. Therefore W and $e(X)$ must be uncorrelated.

Instrumental variables

We also have that

$$\mathbb{E}[WZ] = \mathbb{E}[(Z - e(X))Z] = \mathbb{E}[Z - Ze(X)] = \mathbb{E}[e(X)(1 - e(X))]$$

as $Z^2 = Z$ (w.p. 1). Therefore as $\mathbb{E}[W] = 0$,

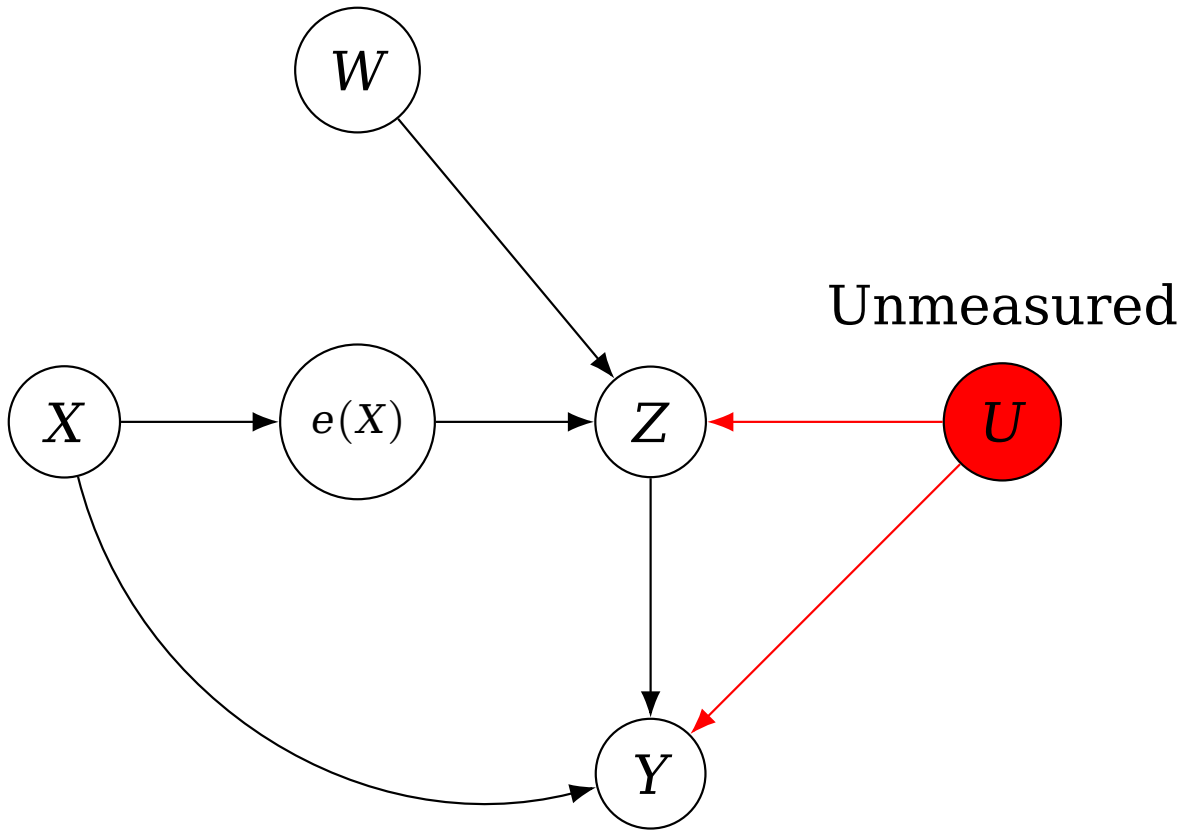
$$\text{Cov}[W, Z] = \mathbb{E}[e(X)(1 - e(X))] > 0$$

under the usual positivity assumption $0 < e(x) < 1$.

Hence $W \not\perp\!\!\!\perp Z$.

Instrumental variables

If we consider W a manipulable variable, we could propose the following DAG:



It is evident that $Y \perp\!\!\!\perp W \mid X, Z$, and also that $U \perp\!\!\!\perp W$.

Instrumental variables

Thus

$$W = Z - e(X) = Z - \mathbb{E}[Z|X].$$

is an instrumental variable. We can compute the ratio estimator based on the earlier identity

$$\psi_0 = \frac{\text{Cov}[W, Y]}{\text{Cov}[W, Z]}$$

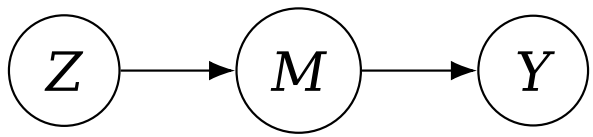
that is

$$\widehat{\psi}_0 = \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i Z_i} = \frac{\sum_{i=1}^{n} (Z_i - e(X_i)) Y_i}{\sum_{i=1}^{n} (Z_i - e(X_i)) Z_i}$$

which is identical to the *singly robust G-estimator*.

Mediation

Statistical methods for causal *mediation* analysis can also be formulated using potential outcomes:



where z and m will represent potential levels of treatment Z and mediator M respectively, with *potential outcome*

$$Y(\mathsf{z}, \mathsf{m})$$

and *potential mediator* value in light of $Z = \mathsf{z}$

$$M(\mathsf{z}).$$

Mediation

In general, both Z and M can take arbitrary values, but we most commonly consider the case when

- Z is binary $\{0, 1\}$
- M is
    - binary,
    - discrete, or
    - continuous.

Mediation

Then following Imai et al. (*Statistical Science,* 2010)

- the *total causal effect* (TCE) of treatment is based on

$$Y(z, M(z))$$

  i.e. set $Z = z$ and observe the consequence.

  For example

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

Mediation

- the *causal mediation effect* (CME) is based on

$$Y(z, M(z^*))$$

i.e. set $Z = z$ and observe the consequence of changing the mediator as if $Z = z^*$

For example

$$\delta(z) = \mathbb{E}[Y(z, M(1)) - Y(z, M(0))] \qquad z = 0, 1.$$

Mediation

- the *direct effect* (DE) of treatment is based on

$$Y(z^*, M(z))$$

i.e. set the mediator as if $Z = z$, and observe the consequence of changing treatment according to z^*.

For example

$$\zeta(z) = \mathbb{E}[Y(1, M(z)) - Y(0, M(z))] \qquad z = 0, 1.$$

Mediation

Terminology:

		Imai et al.	Pearl	Robins
Total effect	τ	ATCE	Total	Total
Indirect effect	$\delta(z)$	ACME	Natural	Pure/Total
Direct effect	$\zeta(z)$	ADE	Natural	Pure/Total

with

$$\text{Pure} : z = 0 \qquad \text{Total} : z = 1.$$

Mediation

That is,

$$\begin{aligned}
&\mathbb{E}[Y(1, M(1)) - Y(0, M(0))] \\
&\qquad = \mathbb{E}[Y(1, M(1)) - Y(1, M(0))] \\
&\qquad\quad + \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]
\end{aligned}$$

so that

$$\tau = \delta(1) + \zeta(0) = \delta(0) + \zeta(1).$$

Mediation

We have in the observed data that

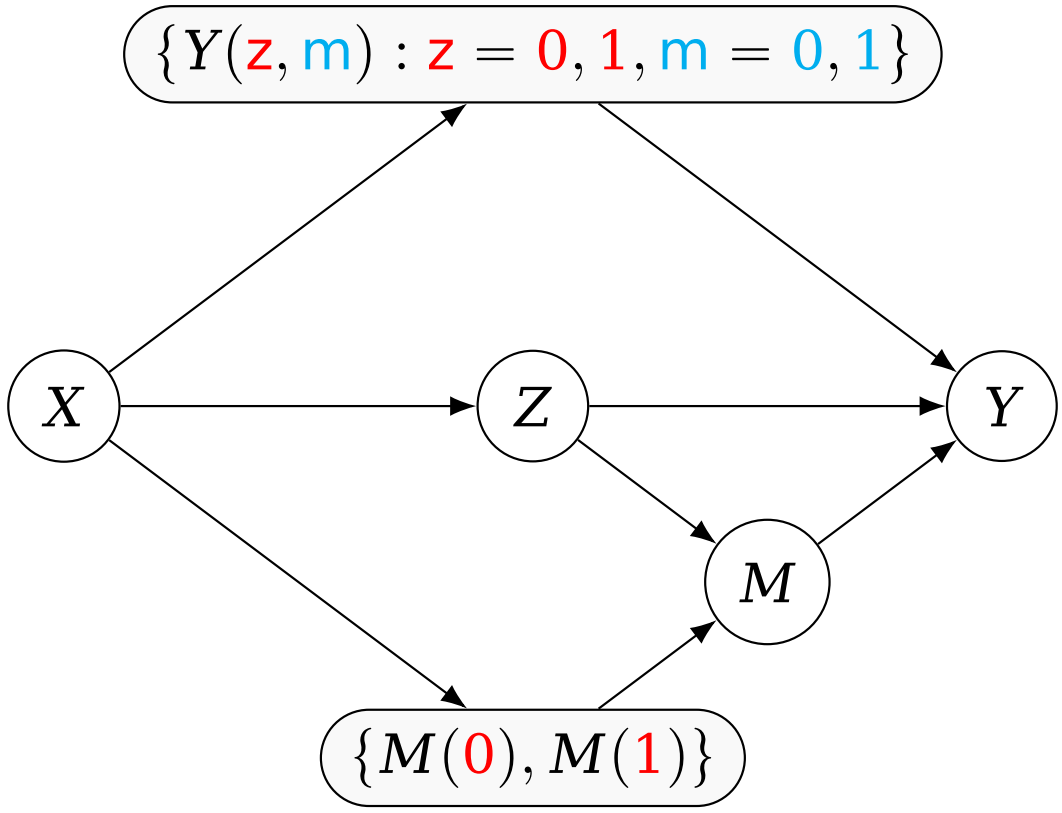
$$M = (1 - Z)M(0) + ZM(1)$$

and, if M is binary say,

$$\begin{aligned} Y = (1 - Z)(1 - M)\ Y(0, 0) \\ + (1 - Z)M\ Y(0, 1) \\ + Z(1 - M)\ Y(1, 0) \\ + Z\ M\ Y(1, 1) \end{aligned}$$

etc.

Mediation



Mediation

The DAG encompasses two ignorability assumptions

(i) $\{Y(\mathsf{z}^*, \mathsf{m}), M(\mathsf{z})\} \perp\!\!\!\perp Z \mid X = x;$

(ii) $Y(\mathsf{z}^*, \mathsf{m}) \perp\!\!\!\perp M(\mathsf{z}) \mid Z = \mathsf{z}, X = x;$

for all $\mathsf{z}, \mathsf{z}^*, \mathsf{m}$ and x, and relies upon two positivity assumptions

$$0 < \Pr[Z = z | X = x] < 1$$

and

$$0 < \Pr[M(z) = m | X = x, Z = z] < 1.$$

Mediation

Imai et al (2010) gives the following identities:

$$\delta(z) = \iiint y f_{Y|M,X,Z}(y|m,x,z) f_{M|X,Z}(m|x,1) f_X(x)\ dy\ dm\ dx$$
$$- \iiint y f_{Y|M,X,Z}(y|m,x,z) f_{M|X,Z}(m|x,0) f_X(x)\ dy\ dm\ dx$$

$$\zeta(z) = \iiint y f_{Y|M,X,Z}(y|m,x,1) f_{M|X,Z}(m|x,z) f_X(x)\ dy\ dm\ dx$$
$$- \iiint y f_{Y|M,X,Z}(y|m,x,0) f_{M|X,Z}(m|x,z) f_X(x)\ dy\ dm\ dx$$

Mediation

As for the ATT case, the challenge in estimating these quantities using moment-based methods is the mis-match in the terms

$$f_{Y|M,X,Z}(y|m,x,z)f_{M|X,Z}(m|x,z^*)$$

whenever $z \neq z^*$.

Mediation

Model-based estimation via the mean-model

$$\mu(m, x, z) = \int y \, f_{Y|M,X,Z}(y|m, x, z) \, dy$$

and the mediator model

$$f_{M|X,Z}(m|x, z).$$

is possible.

Longitudinal Data

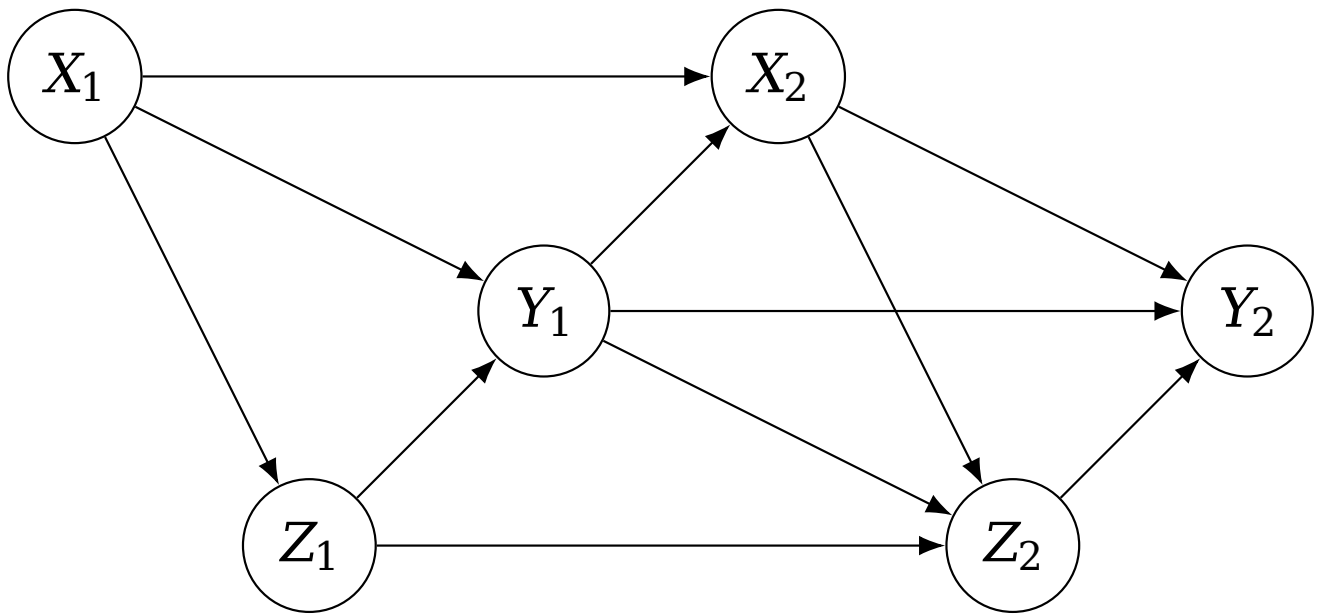
It is quite common to encounter longitudinal data with

- treatment
- confounders
- outcome

recorded over several time points for each individual.

Longitudinal Data

For example

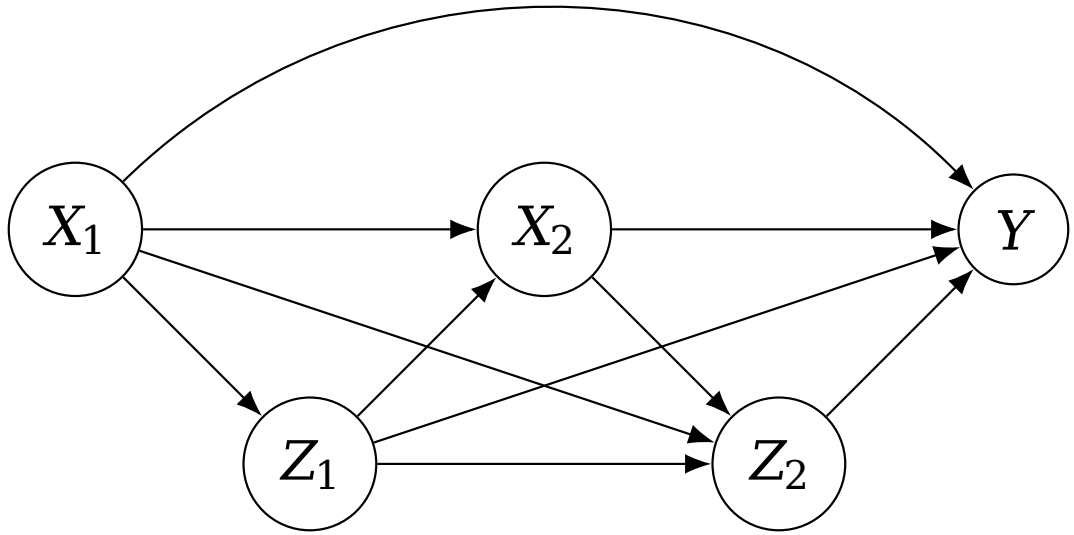


In such cases, and for the analysis of direct effects (within an interval) the methods already established can be utilized sequentially.

For *cumulative* (across interval) effects of treatment on a single terminal outcome, adjustment methods need careful application.

Longitudinal Data

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.

Longitudinal Data

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.

Longitudinal Data

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

- Z_1 has a *direct* effect on Y, but also has 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.

Therefore, 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

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

where

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

is modelled could be used. 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,

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

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.

Marginal structural models

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.

Optimal Dynamic Treatment Regimes

For treatment *sequence* (z_1, z_2), we seek the *optimal* treatment sequence $(z_1^{\text{opt}}, z_2^{\text{opt}})$ to *maximize* expected response.

- The objective is to establish a sequence of decision rules to be applied at each treatment interval;
- For interval $k = 1, 2$, the decision rule for interval k must utilize the information available only up to and including (the start of) that interval;
- The decision rule should be *hyperopic*, that is, prioritize *long-term* outcomes over short-term (*myopic*) outcomes;
- The rules will be learned from observational data, where optimal treatment is not necessarily guaranteed.

Optimal Dynamic Treatment Regimes

To solve this problem, a recursive approach is adopted

- optimal treatment sequence will be *personalized*, that is, dependent on an individual's characteristics;
- the computation will use *reverse* (or *Bellman*) optimization;
- first optimize Stage 2 treatment, and then optimize Stage 1 treatment *assuming optimal treatment at Stage 2*.

Optimal Dynamic Treatment Regimes

Decompose the expected (counterfactual) response as

$$\begin{aligned}\mathbb{E}[Y(z_1, z_2)] = \mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}})] \\ - \left\{\mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}}) - Y(z_1, z_2^{\text{opt}})]\right\} \\ - \left\{\mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(z_1, z_2)]\right\}\end{aligned}$$

This is the basis of the *Structural Nested Mean Model* (SNMM).

Optimal Dynamic Treatment Regimes

Further

$$\mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}}) - Y(z_1, z_2^{\text{opt}})] = \mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}}) - Y(0, z_2^{\text{opt}})] - \mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(0, z_2^{\text{opt}})]$$

$$\mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(z_1, z_2)] = \mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(z_1, 0)] - \mathbb{E}[Y(z_1, z_2) - Y(z_1, 0)]$$

Optimal Dynamic Treatment Regimes

We specify models

$$\mathbb{E}[Y(z_1, z_2) - Y(z_1, 0)] = (\mathbf{X}_2\psi_2)z_2 \qquad \text{Stage 2}$$

$$\mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(0, z_2^{\text{opt}})] = (\mathbf{X}_1\psi_1)z_1 \qquad \text{Stage 1}$$

where

- $\mathbf{X}_2$ can depend on all data – including observed treatments – observed up to Stage 2.
- $\mathbf{X}_1$ can depend on all data observed up to Stage 1.

Optimal Dynamic Treatment Regimes

1. Use *G-estimation* on Y conditioning on $(\mathbf{x}_1, z_1, \mathbf{x}_2, z_2)$ at second stage to estimate ψ_2 using a proposed mean model

$$Y = \mathbf{X}_{21}\beta_2 + z_2\mathbf{X}_{22}\psi_2 + \varepsilon$$

2. For each individual infer the optimal Stage 2 treatment

$$z_2^{\text{opt}} = \mathbb{1}\{\mathbf{X}_{22}\widehat{\psi}_2 > 0\}$$

3. Form *pseudo-outcome* Y_1

$$Y_1 = Y - (\mathbf{X}_{22}\widehat{\psi}_2)(z_2^{\text{opt}} - z_2)$$

Optimal Dynamic Treatment Regimes

4. Use G-estimation on Y_1 conditioning on $(\mathbf{x}_1, z_1)$ at first stage to estimate ψ_1 using a proposed mean model

$$Y_1 = \mathbf{X}_{11}\beta_1 + z_1\mathbf{X}_{12}\psi_1 + \varepsilon$$

5. For each individual infer the optimal Stage 1 treatment

$$z_1^{\text{opt}} = \mathbb{1}\{\mathbf{X}_{11}\widehat{\psi}_1 > 0\}$$

The method is robust to mis-specification of the *nuisance mean model*, provided the treatment model is correctly specified.

Optimal Dynamic Treatment Regimes

Can infer optimized potential outcome

$$Y + (\mathbf{X}_{12}\widehat{\psi}_1)(z_1^{\text{opt}} - z_1) + (\mathbf{X}_{22}\widehat{\psi}_2)(z_2^{\text{opt}} - z_2)$$

which takes the observed outcome Y and adds

- the additional benefit of optimal treatment at stage 1

$$(\mathbf{X}_{12}\widehat{\psi}_1)(z_1^{\text{opt}} - z_1)$$

- the additional benefit of optimal treatment at stage 2

$$(\mathbf{X}_{22}\widehat{\psi}_2)(z_2^{\text{opt}} - z_2).$$

Causal discovery

Many of the adjustment methods rely upon the construction of estimators assuming knowledge of

- confounder structure
- time-ordering
- propensity score construction

all of which depend on knowledge of the data generating DAG.

In many applications, the DAG is *unknown*.

Causal discovery

There are several approaches to *causal discovery*, that is, learning the DAG from the observed data

- PC algorithm: `pcalg` in `R`
- SGS algorithm

Both rely on recursive identification of conditional independencies, identification of colliders and edge orientation procedures based on statistical tests.

Can be effective, but typically need large sample sizes.