MATH 598

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:

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

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*.

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

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

where

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

is the *indicator function*.

That is,

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

where V is a *degenerate* random variable with

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

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:

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

Some basic probability calculus

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

- fixing $Z = \mathsf{z}$ independently of X,
- computing for each fixed x

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

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

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

Some basic probability calculus

Regression: we might propose

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

or

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

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

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

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

$$
\begin{aligned}
\Pr[Y = y | Z = z] &= \sum_{x=0}^{1} \Pr[Y = y | X = x, Z = z] \Pr[X = x | Z = z] \\
&= \Pr[Y = y | X = 0, Z = z] \Pr[X = 0 | Z = z] \\
&\qquad + \Pr[Y = y | X = 1, Z = z] \Pr[X = 1 | Z = z] \\
&\neq \Pr[Y = y | X = 0, Z = z] \Pr[X = 0] \\
&\qquad + \Pr[Y = y | X = 1, Z = z] \Pr[X = 1]
\end{aligned}
$$

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:

$X = 0$

		Y = 0	Y = 1
Z	0	36	234
	1	6	81

$X = 1$

		Y = 0	Y = 1
Z	0	25	55
	1	71	192

Collapsing over X:

		Y = 0	Y = 1
Z	0	61	289
	1	77	273

Some basic probability calculus

Estimated cure rates for the two treatment groups:

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

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

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

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

In the pooled data:

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

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

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

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

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

and

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

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

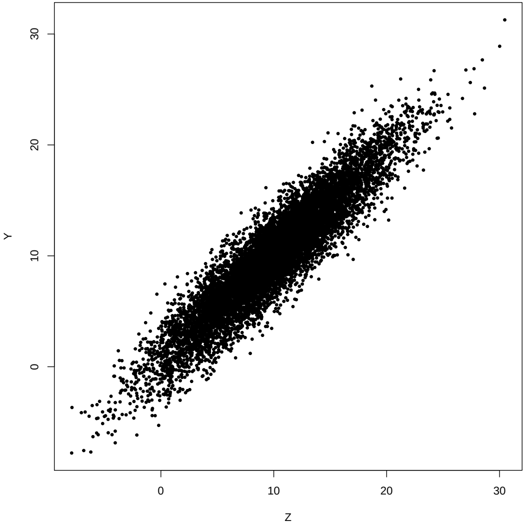
$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim Normal_3(\mu, \Sigma)$$

constructed as follows:

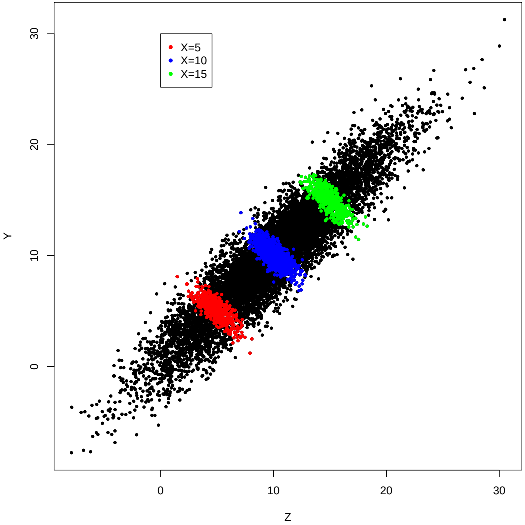
- *Marginal* for X: $X \sim Normal(\mu_X, \sigma_X^2)$
- *Conditional* for (Y, Z) given $X = x$:

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

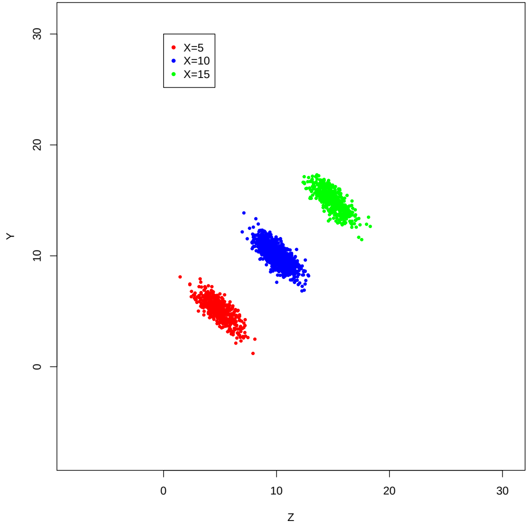
Some basic probability calculus



Some basic probability calculus



Some basic probability calculus



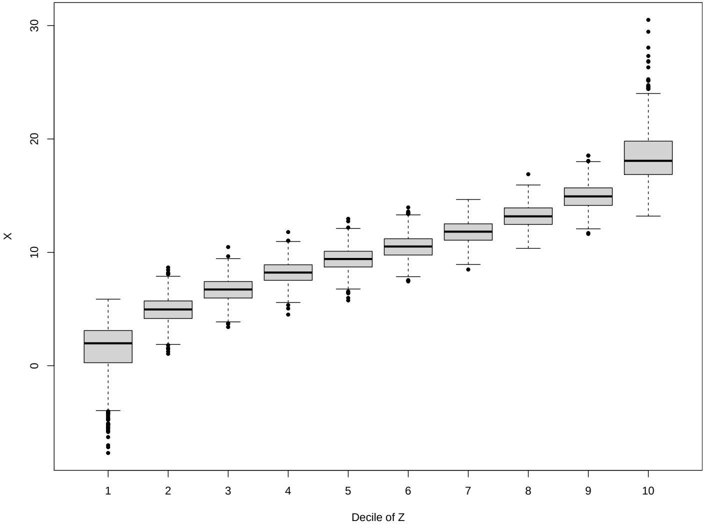
Some basic probability calculus

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

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

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

$$\begin{aligned} f_{X|Z}(x|z) &\propto f_{Z|X}(z|x) f_X(x) \\ &\equiv Normal\left(\frac{z + \mu_X/\sigma_X^2}{1 + 1/\sigma_X^2}, \frac{1}{1 + 1/\sigma_X^2}\right) \end{aligned}$$

Some basic probability calculus

Therefore

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

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

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

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

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

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

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

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

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

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

Regression models

Most commonly for continuous-valued Y

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

and

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

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

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

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

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

Regression models

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*.

- 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.

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

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

- 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,

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

- *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,

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

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 '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

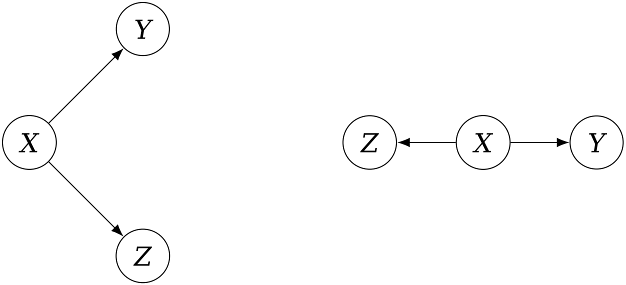
$$Z \perp\!\!\!\perp Y|X$$

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*;

$$X \longrightarrow Y$$

corresponds to

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

Causal Graphs

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

$$Z \longleftarrow X \longrightarrow Y$$

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

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

  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:

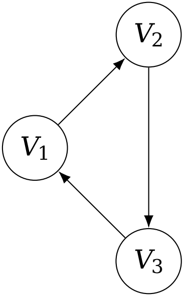
- A set of *nodes* or *vertices*, $V_1, V_2, \ldots$, representing variables.
- A set of *edges*, $E_1, E_2, \ldots$, representing dependencies.
- Two nodes are *adjacent* if there is an edge between them.
- Edges can be *directed*, denoted using arrows, or *undirected*; if all edges are directed, the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

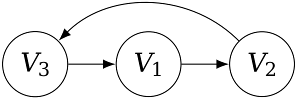
Causal Graphs

- 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.

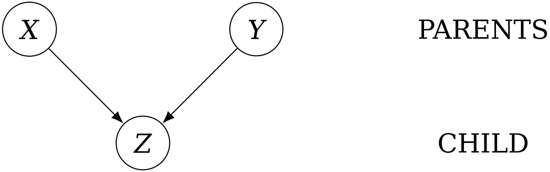




A directed graph that has no cycles is termed a *directed acyclic graph* (DAG).

Causal Graphs

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



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

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

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

as

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

Causal Graphs

However, conditioning on $Z = z$

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

in general.

That is,

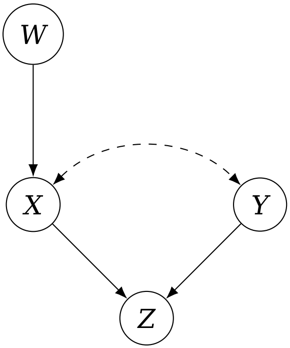
$$X \perp\!\!\!\perp Y$$

but

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

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

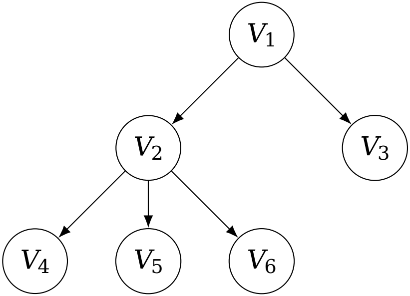
Causal Graphs



- 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.

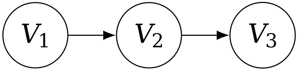
Causal Graphs

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*.



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

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

such that for $X_k \notin \mathrm{P}_{\mathrm{A}_j}$,

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

and that no proper subset of $\mathrm{P}_{\mathrm{A}_j}$ yields the conditional independence. That is, $\mathrm{P}_{\mathrm{A}_j}$ is the *smallest* set of variables for which the conditional independence statement holds.

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

$$X_1 \qquad X_2$$

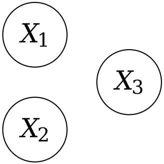
$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2)$$

or

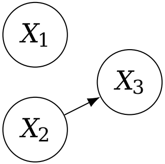
$$X_1 \longrightarrow X_2$$

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

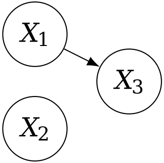
Add in X_3: independence case



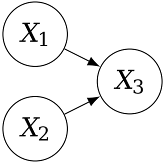
$X_3 \perp\!\!\!\perp (X_1, X_2)$



$X_3 \perp\!\!\!\perp X_1 \mid X_2$



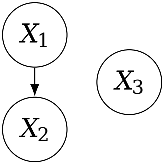
$X_3 \perp\!\!\!\perp X_2 \mid X_1$



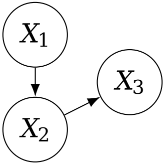
$X_3 \not\perp\!\!\!\perp X_1, X_3 \not\perp\!\!\!\perp X_2$

Causal Graphs

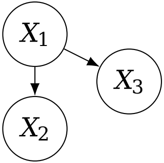
Add in X_3: dependence case



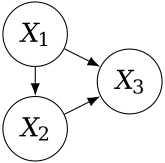
$X_3 \perp\!\!\!\perp (X_1, X_2)$



$X_3 \perp\!\!\!\perp X_1 \mid X_2$



$X_3 \perp\!\!\!\perp X_2 \mid X_1$



$X_3 \not\perp\!\!\!\perp X_1, X_3 \not\perp\!\!\!\perp X_2$

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

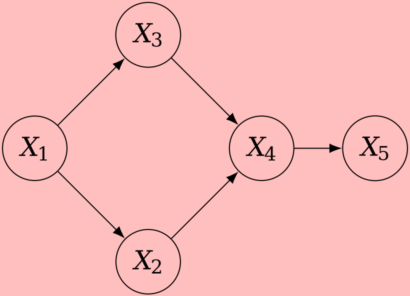
- note the G does not define P, merely the chain rule factorization that P admits;
- termed '*Markov compatibility*'; P is *Markov with respect to G*, that is, we may deduce from G that

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

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

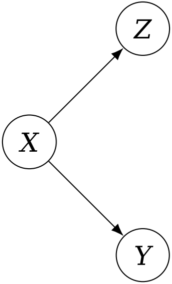
$$f_{X_3|X_1,X_2}(x_3|x_1,x_2)$$

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 ?

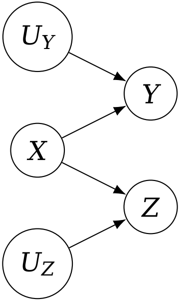
Structural models

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

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

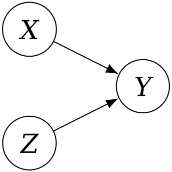
$$Z = g_2(X, U_Z)$$

For example

$$Y = X + U_Y$$

$$Z = X + U_Z$$

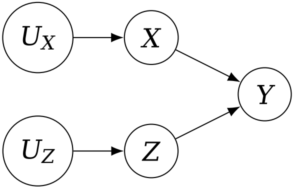
Structural models



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

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

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

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

$$X \rightarrow Z \leftarrow Y$$

Z is a collider on this path.

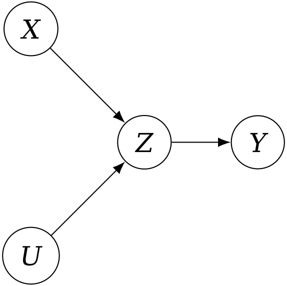
A *directed path* from one node to another cannot contain a collider; all parts must be

forks $\quad X \leftarrow Z \rightarrow Y$

chains $\quad X \rightarrow Z \rightarrow Y$

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.

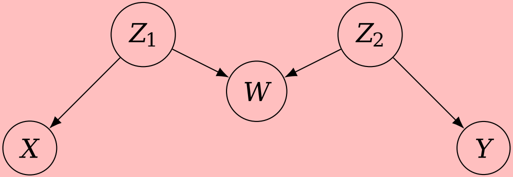
Unconditional d-separation: A path is *open* (or *unblocked*, or *active*) unconditionally if there is no collider on the path; if there is a collider, the path is *closed* (*blocked*, *inactive*)

Two 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.

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



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.

d-separation

Conditional d-separation: we can consider similar statements obtained after *conditioning* on a variable.

For a non-collider Z: consider conditioning on Z

$X \leftarrow Z \rightarrow Y$ $\qquad\qquad$ $X \perp\!\!\!\perp Y \mid Z$

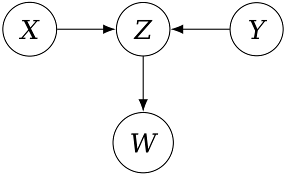
$X \rightarrow Z \rightarrow Y$ $\qquad\qquad$ $X \perp\!\!\!\perp Y \mid Z$

For a collider Z: consider conditioning on Z

$X \rightarrow Z \leftarrow Y$ $\qquad\qquad$ $X \not\!\perp\!\!\!\perp Y \mid Z$

d-separation

However, consider



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

We have that X and Y are independent.

d-separation

But

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

Therefore we have that

$$X \longrightarrow W \longleftarrow Y \qquad\qquad X \not\perp\!\!\!\perp Y \mid W$$

and W is a collider.

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;

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*.

d-separation

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

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

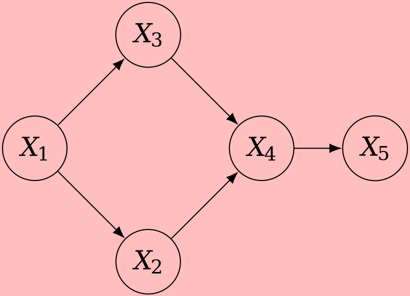
- If S does not d-separate X and Y, then X and Y may be dependent, and

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

  cannot be made independent of y in general.

d-separation

Example:

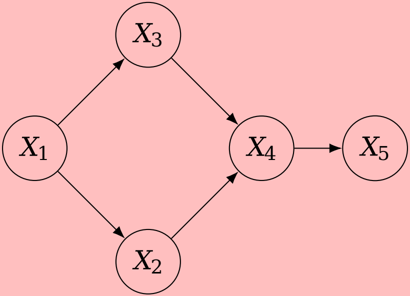


$\{X_2\}$ and $\{X_3\}$ are d-separated by $\{X_1\}$, and $X_2 \perp\!\!\!\perp X_3 \mid X_1$.

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

d-separation

Example:

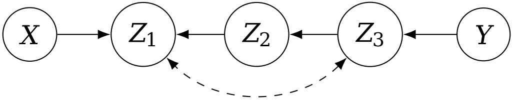


$\{X_2\}$ and $\{X_3\}$ are *not* d-separated by $\{X_1, X_5\}$:

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

Selection bias: Conditioning on the common effect of two causes renders the two causes dependent; this is known as

- *selection bias* or *Berkson bias*

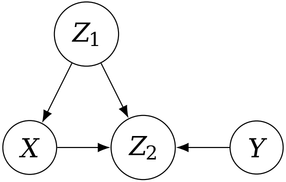


Here $X \perp\!\!\!\perp Y$: there are two paths between X and Y

- $XZ_1Z_2Z_3Y$ is blocked by the collider Z_1.
- XZ_1Z_3Y is blocked by the colliders Z_1 and Z_3.

Therefore $X \not\perp\!\!\!\perp Y \mid \{Z_1, Z_3\}$.

d-separation



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

- XZ_1Z_2Y
- XZ_2Y

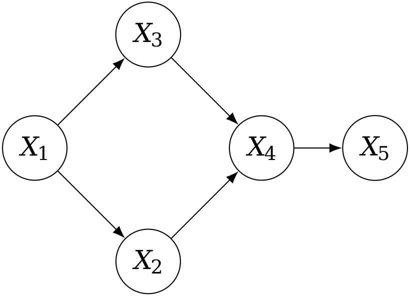
both blocked by collider Z_2. Therefore $X \not\perp\!\!\!\perp Y \mid \{Z_2\}$.

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

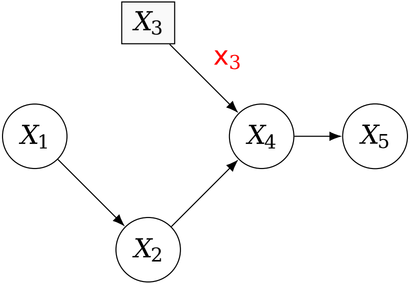
Consider the DAG



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

Interventions

Suppose we *intervene* to set $X_3 = \mathsf{x}_3$. The relevant DAG is now



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

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

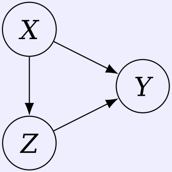
Intervening on a variable X to set the level to x has the effect of

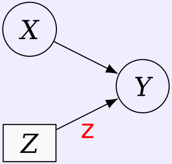
- removing all *incoming* arrows to X
- switching the marginal for X to the *degenerate distribution* $f_X^*(.)$

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

Note

In the earlier example





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

Graphical representation of bias

We aim to understand the effect of Z on Y, say

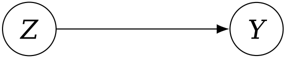
- An *open* undirected path between Z and Y allows for the *association* between Z and Y to be modified by the presence of other variables.

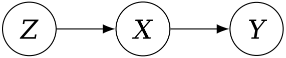
  This is known as a *biasing* path.

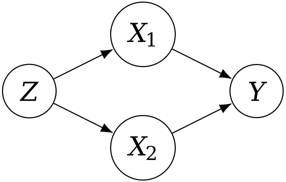
- by *association*, we typically mean some form of *correlation* (or *partial correlation*).

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

For variables S, S is *sufficient* to control bias in the association between Z and Y if, conditional on S, the *open* paths between Z and Y are precisely the *directed* paths between Z and Y.

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

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

Graphical representation of bias

Conditioning on descendants of Z: such conditioning

(i) *blocks* directed paths

$$Z \longrightarrow X \longrightarrow Y$$

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

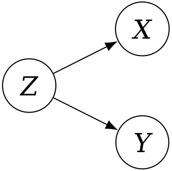
Graphical representation of bias

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

- collider case
- selection bias case.

Graphical representation of bias

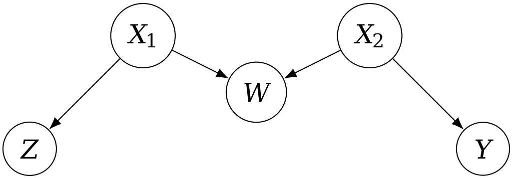
(iii) may be unnecessary in statistical terms: for example



Conditioning on X will not affect bias.

Graphical representation of bias

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



The only path from Z to Y is a backdoor path.

Graphical representation of bias

Before conditioning

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

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

Graphical representation of bias

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

(i) S contains no descendant of Z, and

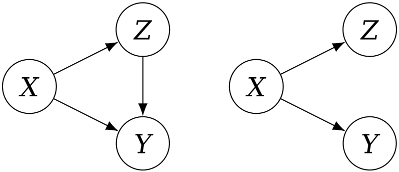
(ii) there is no open backdoor path from Z to Y after conditioning on S.

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

Graphical representation of bias

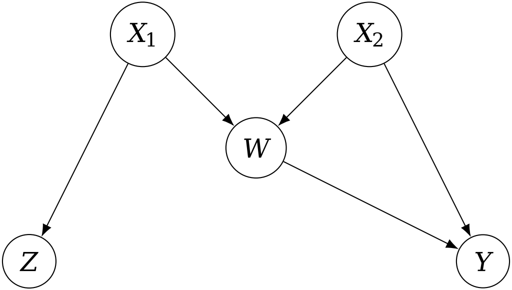
Confounding: A *confounding path* between Z and Y is a biasing path (that is, an undirected open path) that ends with an arrow into Y.

Variables on a confounding path are termed *confounders*.



X is a confounder in both cases.

Graphical representation of bias



W is a collider on the path from Z to Y

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

and hence this path is blocked.

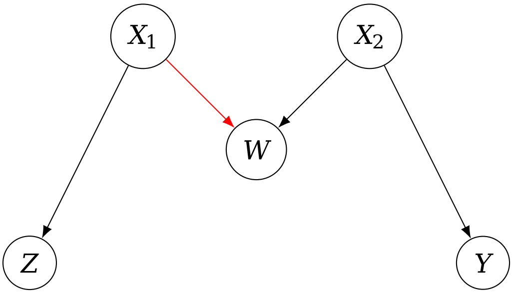
Graphical representation of bias

However unconditional on W, the effect of Z on Y is confounded by the backdoor path, Path 2: ZX_1WY.

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

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

Graphical representation of bias

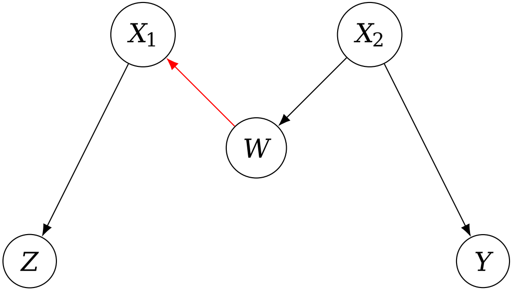


Conditioning on W *opens* the confounding path. Therefore $Z \perp\!\!\!\perp Y$ (as there is no open path between them), but

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

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

Graphical representation of bias



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

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

will block the path.

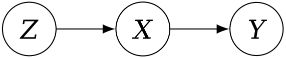
Direct and indirect effects

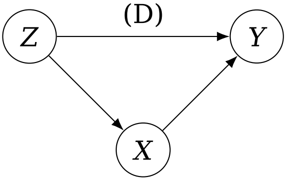
- *Direct effect:* A direct effect of Z on Y (relative to X) is the effect captured by a *directed* path from Z to Y that does not pass through X.
- *Indirect effect:* An indirect effect of X on Y that is captured by directed paths that pass through X.

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



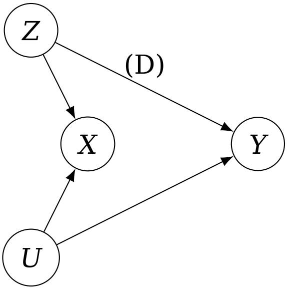


Indirect effect

Direct (D) & Indirect effect

Direct effect is confounded

Direct and indirect effects



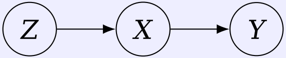
No indirect effect

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

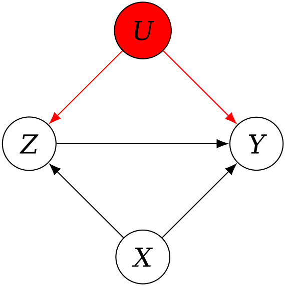


If we ignore X as a mediator:

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

Unmeasured confounding

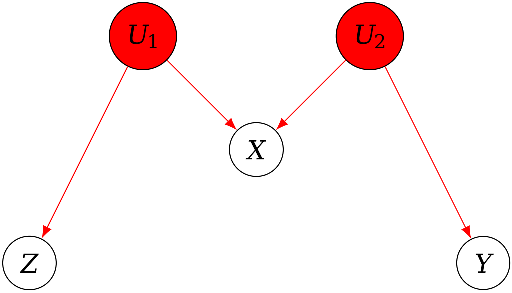
Suppose that in reality there is a further variable U that is a confounder, but is unmeasured in the observed data.



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

Unmeasured confounding

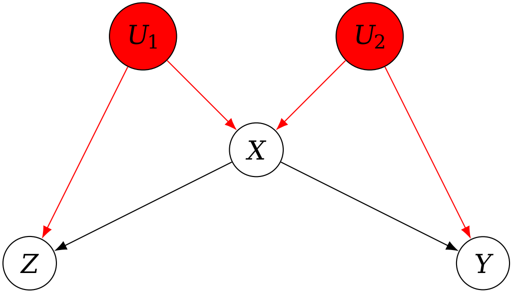
With two unmeasured confounders:



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

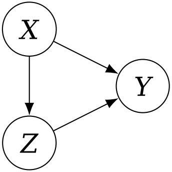
of Y.

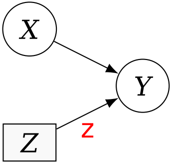
The intervention changes Z from z_0 to z_1, say; in the earlier notation, we consider the intervention model

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

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

Potential outcome notation

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

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

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

$$Y(0): \text{ outcome if intervention sets } \mathsf{z} = 1$$

$$Y(1): \text{ outcome if intervention sets } \mathsf{z} = 0$$

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

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

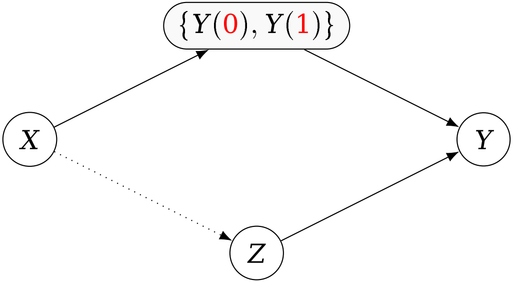
(with probability 1). That is, the observed outcome is identical to the potential outcome arising from the counterfactual treatment that matches the observed treatment.

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

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

According to earlier assumptions, we should have that

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

that is, conditional on X, the effect on Y of the intervention to set $Z = \mathsf{z}$ 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.

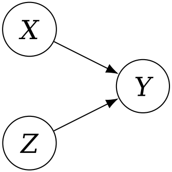
A simple form of experimental study proceeds as follows: $2n$ individuals are sampled from a homogeneous population.

- 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.

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



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

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

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

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

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

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

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).

In terms of the above distributions, we may write

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

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

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

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

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

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

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

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 $Z = \mathsf{z}$;
- denominator is merely the number of individuals who received treatment $Z = \mathsf{z}$.

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

For the binary case, the formulae simplify

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

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

Note that we know that

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

so we can consider the alternative estimator

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

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

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

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

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

Note that

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

and

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

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.

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

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

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.

It transpires that

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

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.

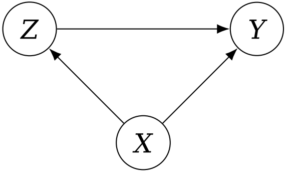
We denote the data generating mechanism

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

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



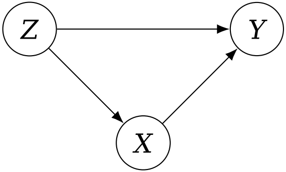
$$f^{\mathcal{O}}_X(x) f^{\mathcal{O}}_{Z|X}(z|x) f^{\mathcal{O}}_{Y|X,Z}(y|x,z)$$

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

To get at the causal effect of Z on Y, we must block the back-door 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 $Z\ X\ Y$, 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.

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

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

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

$$Z \longrightarrow X \qquad\qquad X \longrightarrow Z$$

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

Observational studies

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

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

This can be contrasted with the earlier formula for the APO

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

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

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

and so evidently

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

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;
- 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

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

where

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

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

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

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

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

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 the observations $Z = 1$ 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

$$\mu(x, z)$$

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

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

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.