

MATH598

INTRODUCTION TO CAUSAL INFERENCE METHODS

Objective : to uncover the 'causal'
(i.e. uncontaminated, unconfounded)
effect of one variable on another.

i.e. if we change the value of one variable
by intervention, what is the resulting impact on the
other.

Some elementary probability calculus:

Consider ~~two~~ variables X, Y, Z .

• joint density $f_{X,Y,Z}$.

→ marginals f_X, f_Y, f_Z .

→ conditionals $f_{Y|X}, f_{Z|X}, f_{Y|Z,X}$ etc.

we know

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

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

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

where

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

$$= \frac{f_{X,Y,Z}(x, y, z)}{\int f_{X,Y,Z}(x, u, z) du}.$$

ad

$$f_{z|x}(z|x) = \frac{f_{x,z}(x,z)}{f_x(x)} = \frac{\int f_{x,y,z}(x,u,z) du}{\iint f_{x,y,z}(x,u,v) du dv}.$$

$$f_{x|z}(x|z) = \frac{f_{x,z}(x,z)}{f_z(z)}$$

ad
$$f_{z|x}(z|x) = \frac{f_{x|z}(x|z) f_z(z)}{f_x(x)}.$$

etc.

We can also compute expectations

$$E[Y] = \int y f_Y(y) dy.$$

$$= \iiint y f_{X,Y,Z}(x,y,z) dx dy dz.$$

$$= \iint \left\{ \int y f_{Y|X,Z}(y|x,z) dy \right\} f_{X,Z}(x,z) dx dz.$$

$$= \iint E[Y|X=x, Z=z] f_{X,Z}(x,z) dx dz.$$

$$= E_{Z,X} \left[\underset{Y|X,Z}{E[Y|X,Z]} \right]$$

'iterated
expectation'.

$$E[Y|Z=z] = \int y f_{Y|Z}(y|z) dy.$$

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

$$= \iiint \mathbb{I}_{\{z\}}(v) y f_{Y|X,Z}(y|x,v) f_{X|Z}(x|v) dy dx dv$$

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

i.e

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

This holds in general. However, suppose we assume $X \perp Z$

$$\text{i.e. } f_{X,Z}(x,z) = f_X(x) f_Z(z) \quad \forall x, z.$$

$$\text{Then } f_{X|Z}(x|z) = f_X(x) \quad \forall x, z.$$

$$E[Y|Z=z] = \iint \left\{ \int y f_{Y|X,Z}(y|x,z) dy f_X(x) \right\} dx dz.$$

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

z only appears here.

$$= E_X \left[E_{Y|X,Z} [Y|X, Z=z] \right]$$

• i.e. we fix z (independence of X).

– consider $E[Y|X=x, Z=z]$

– average this over the distribution of X , $f_X(x)$.

$E[Y|X=x, Z=z]$ – regression!

• eg $E[Y|X=x, Z=z] = \beta_0 + \beta_1 x + \beta_2 z$.

or

$$E[Y|X=x, Z=z] = \beta_0 + \beta_1 x + \beta_2 z + \beta_{12} x \cdot z.$$

or

$$\begin{aligned} E[Y|X=x, Z=z] &= \mu(x, z) \\ &= \mu(x, z; \beta). \end{aligned}$$

Also: suppose X, Y, Z are binary variables.

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

$$P[Y=y | Z=z] \quad (y, z) \in \{0, 1\}^2.$$

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

$$\Rightarrow P[Y=y | Z=z] = P[Y=y | X=0, Z=z] P[X=0 | Z=z] \\ + P[Y=y | X=1, Z=z] P[X=1 | Z=z]$$

$$\neq P[Y=y | X=0, Z=z] P[X=0] \\ + P[Y=y | X=1, Z=z] P[X=1]$$

in general.

Simpson's Paradox

$$Z = \begin{cases} 0 & \text{Treatment A (No Drug)} \\ 1 & \text{Treatment B (Drug)} \end{cases}$$

X = $\begin{cases} 0 & \text{Group 0} \\ 1 & \text{Group 1} \end{cases}$ (Men)
(Women)

$$Y = \begin{cases} 0 & \text{Not cured} \\ 1 & \text{Cured} \end{cases} \quad \begin{cases} \text{(Not recovered)} \\ \text{(Recovered)} \end{cases}$$

Data

$X = 0$

Y

	0	1	
0	36	234	270
1	6	81	87
	42	315	357

$X = 1$

	Y	
	0	1
Z	0	25
	1	71

80
263
96 247 343

Collapsing over X

	0	Y	1	
7	0	61	289	350
	1	77	273	350
		138	562	700

~~Re-est~~

(Cured)
Prob. Recovery

Estimate .

$$X=0 : z = \phi$$
$$z = 0$$

$$\frac{81}{87} = 0.93$$
$$\frac{234}{270} = 0.87$$

$$X=1 \quad z = \phi$$
$$z = 0$$

$$\frac{192}{263} = 0.73$$

$$\frac{55}{80} = 0.69$$

$\therefore$ In both groups, prob of cure is greater
if $z=1$.

Overall

$$z = \phi$$
$$z = 0$$

$$\frac{273}{350} = 0.78$$

$$\frac{289}{350} = 0.83$$

on aggregate, prob of cure is lower if $z=1$.

Not surprising.

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

Here

$$\hat{P}[Y=1|Z=1, X=0] = 0.93 \quad (a)$$

$$\hat{P}[Y=1|Z=0, X=0] = 0.87 \quad (b)$$

$$\hat{P}[Y=1|Z=1, X=1] = 0.71 \quad (c)$$

$$\hat{P}[Y=1|Z=0, X=1] = 0.69 \quad (d)$$

$$\hat{P}[Y=1|Z=1] = 0.78 \quad \text{convex comb}^n \text{ of (a) and (c)}$$

$$\hat{P}[Y=1|Z=0] = 0.83 \quad \text{" " " (b) and (d)}$$

with weights

$$P[X=0|Z=1] \text{ and } P[X=0|Z=0]$$

respectively. These weights are critical.

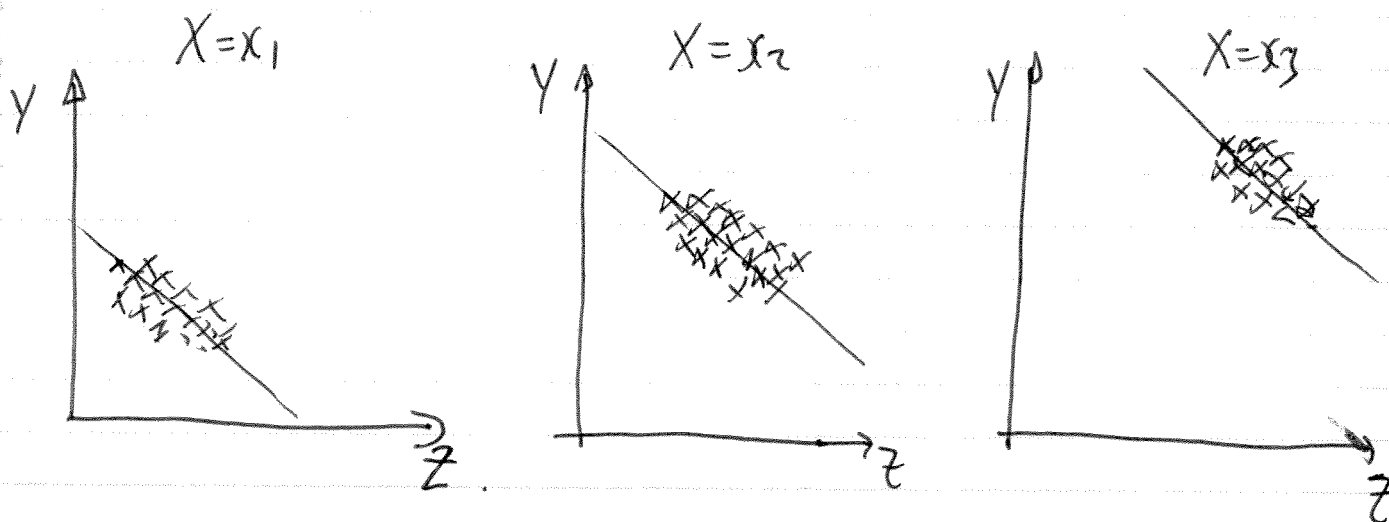
$P[X=0/Z=1]$ and $P[X=0/Z=0]$
are essentially unconstrained.

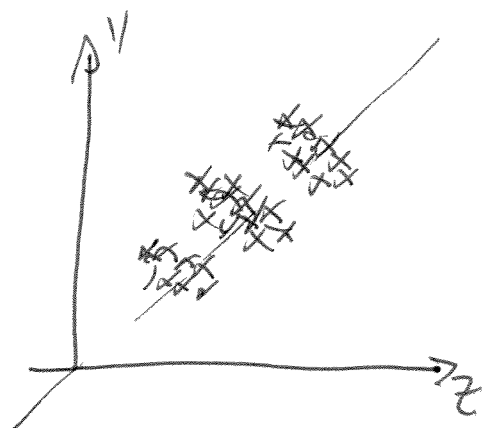
$\therefore$ No way to predict the relative
magnitudes of the group specific
and marginal probs.

If X, Y, Z continuous

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \text{Normal}_3(\mu, \Sigma)$$

- same kind of result.





Pooled over X .

Here we can write down the distribution -

$$Y|X=x, Z=z.$$

and

$$Y|Z=z.$$

using Normal theory.

Note In the data, we observe the critical distribution

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

which infers the result for $f_{Y|Z}(y|z)$.

However, for 'causal' questions, we want to answer the hypothetical question.

'what is the distribution of Y as the researcher manipulates Z
(not the subject) ?'

From the data, we cannot answer this question in general.

EXAMPLE

$X, Y, Z \sim \text{MVN}$

$$Y, Z \mid X=x \sim N\left(\begin{pmatrix} x \\ x \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

$$X \sim \text{Normal}(\mu_X, \sigma_X^2).$$

Now to compute

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

By standard bivariate theory

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

i.e

$$E_{Y|X,Z}[Y|X=x, Z=z] = x + \rho(z-x).$$

$\therefore$ For each fixed x , regressing Y on Z yields an intercept $(1-\rho)x$.

slope ρ .

However

$$f_{Y,Z}(y,z) = \int_{-\infty}^{\infty} f_{Y,Z|X}(y,z|x) f_X(x) dx.$$

$$f_{Y|Z}(y|z) = \int_{-\infty}^{\infty} f_{Y|X,Z}(y|x,z) f_{X|Z}(x|z) dx.$$

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

$$f_{Z|X}(z|x) \propto \exp\left\{-\frac{1}{2}(z-x)^2\right\}.$$

$$f_X(x) \propto \exp\left\{-\frac{1}{2\sigma_x^2}(x-\mu_x)^2\right\}.$$

$$\therefore f_{X|Z}(x|z) \propto \exp\left\{-\frac{1}{2}\left[(z-x)^2 + \frac{1}{\sigma_x^2}(x-\mu_x)^2\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(1 + \frac{1}{\sigma_x^2}\right)\left(x - \frac{z + \frac{\mu_x}{\sigma_x^2}}{1 + \frac{1}{\sigma_x^2}}\right)^2\right\}.$$

$$\text{i.e. } X|Z=z \sim N\left(\frac{z + \frac{\mu_x}{\sigma_x^2}}{1 + \frac{1}{\sigma_x^2}}, \frac{1}{1 + \frac{1}{\sigma_x^2}}\right).$$

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

$\therefore$ regressing Y on Z (ignoring X). yields
 intercept $\frac{(1-\rho)\mu_X}{\sigma_X^2 + 1}$.

slope $\rho + \frac{\sigma_X^2}{1 + \sigma_X^2}$

$\therefore$ can have $\rho < 0$ but $\rho + \frac{\sigma_X^2}{1 + \sigma_X^2} > 0$.

$$\begin{aligned}
\text{Var}_{Y|Z}[Y|Z=z] &= E_{X|Z}[\text{Var}_{Y|X,Z}[Y|X=x, Z=z]|Z=z] + \text{Var}_{X|Z}[E_{Y|X,Z}[Y|X=x, Z=z]|Z=z] \\
&= (1-\rho^2) + (1-\rho)^2 \sigma_X^2 / (1 + \sigma_X^2).
\end{aligned}$$

Note: To get at 'the effect of z ' on Y
we want to imagine changing z independently
of X .

- here, we could consider the conditional mode!

$$[E_{Y|X,z}[Y|X=x, z=z]] = \rho z + (1-\rho)x.$$

- 'effect' of z is ρ in the presence of $X=x$.

or we could consider this 'averaged' over
the observed distribution of X .

i.e

$$[E_X[E_{Y|X,z}[Y|X=x, z=z]]]$$

(i.e z fixed, independently
of X .)

i.e here

$$\begin{aligned} & \mathbb{E}_{X|Z} \left[\mathbb{E}_{Y|X,Z} [Y|X=x, Z=z] / Z=z \right] \\ &= \mathbb{E}_X \left[\mathbb{E}_{Y|X,Z} [Y|X=x, Z=z] / Z=z \right] \\ &= \mathbb{E}_X \left[\rho z + (1-\rho)X \right] \\ &= \rho z + (1-\rho)\mu_X. \end{aligned}$$

NB Here X IS NOT independent of Z

$$f_{X|Z}(x|z) \equiv N\left(\frac{z + \frac{\mu_X}{\sigma_X^2}}{1 + \frac{1}{\sigma_X^2}}, \frac{1}{(1 + \frac{1}{\sigma_X^2})}\right)$$

$$\neq N(\mu_X, \sigma_X^2) = f_X(x)$$

i.e. In the model

$$E_{Y|X,Z}[Y|X=x, Z=z]$$

is identical (unchanged) whatever we assume about the (X, Z) relationship.

. in the model X and Z are dependent

. we can retain $f_X(X)$ as unchanged.

. observed data $f_{Z|X}(z|x)$

. 'intervention' $f_{Z|X}^*(z|x) \equiv f_z^*(z)$

to set $z=z$

degenerate at ~~z~~
single point.

NOTE We specified

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \text{Normal}_3 \left(\begin{pmatrix} \mu_X \\ \mu_Y \\ \mu_Z \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \sigma_{XY} & \sigma_{XZ} \\ \sigma_{XY} & \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & \sigma_Z^2 \end{pmatrix} \right)$$

$$X \sim \text{Normal}(10, 5^2) \quad (\text{i.e. } \mu_X = 10, \sigma_X^2 = 5^2).$$

$$\begin{pmatrix} Y \\ Z \end{pmatrix} \Big| X=x \sim \text{Normal} \left(\begin{pmatrix} x \\ x \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

$$\text{i.e. } \mu_Y = \mu_Z \equiv \mu_X. \quad (\text{iterated expectation}).$$

But if

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \text{Normal} \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Then

$$X_2 \Big| X_1 = x_1 \sim \text{Normal} \left(\mu_2 + \Sigma_{21} \Sigma_{11}^{-1} (x_1 - \mu_1), \right. \\ \left. \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right).$$

$$\therefore \begin{pmatrix} Y \\ Z \end{pmatrix} \bigg| X=x.$$

$$\sim \text{Normal} \left(\begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix} + \frac{1}{\sigma_X^2} \begin{pmatrix} \sigma_{XY} \\ \sigma_{XZ} \end{pmatrix} (X - \mu_X), \right.$$

$$\left. \begin{pmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z^2 \end{pmatrix} - \frac{1}{\sigma_X^2} \begin{pmatrix} \sigma_{XY} \\ \sigma_{XZ} \end{pmatrix} (\sigma_{XY}, \sigma_{XZ}) \right).$$

$\therefore$ Must have

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

As $\mu_Y = \mu_X$, must have $\frac{\sigma_{XY}}{\sigma_X^2} = 1$.

$\mu_Z = \mu_X$, must have $\frac{\sigma_{XZ}}{\sigma_X^2} = 1$

i.e. $\sigma_{XY} = \sigma_{XZ} = \sigma_X^2$ ($= 10$).

and hence

$$\sigma_y^2 = 1 + \frac{\sigma_{xy}^2}{\sigma_x^2} = 1 + \sigma_x^2 \quad (= 26)$$

$$\sigma_z^2 = 1 + \sigma_x^2 \quad (= 26).$$

$$\sigma_{yz} = \rho + \frac{\sigma_{xy}\sigma_{xz}}{\sigma_x^2} = \rho + \sigma_x^2 \quad (= 24.1).$$

Here

$$\rho = \text{Corr}[Y, Z | X]$$

- 'partial correlation' - $\rho_{YZ.X}$

Could also work out the

partial correlations $\text{Corr}[X, Y | Z]$ $\rho_{XY.Z}$

$\text{Corr}[X, Z | Y]$ $\rho_{XZ.Y}$

These contrast with the 'ordinary'
correlations

$$\rho_{xy} \equiv \text{Cor}[X, Y]$$

$$\rho_{xz} \equiv \text{Cor}[X, Z]$$

$$\rho_{yz} \equiv \text{Cor}[Y, Z].$$

-in the simulation, we have

$$\rho_{yz|x} = -0.9.$$

but

$$\rho_{yz} = \frac{\sigma_{yz}}{\sqrt{\sigma_x^2 \sigma_z^2}} = \frac{\rho + \sigma_x^2}{(\sigma_x^2 + 1)} = \frac{24 \cdot 1}{26}.$$

$$\rho_{xz} = \rho_{xy} = \frac{\sigma_x^2}{\sqrt{\sigma_x^2(1 + \sigma_x^2)}} = \frac{25}{\sqrt{25 \times 26}}.$$

Regression models

Y - scalar

X - ~~scalar~~ $1 \times p$

β $p \times 1$

$$E[Y|X] = X\beta$$

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

Data $y_1, \dots, y_n$

$x_1, \dots, x_n \rightarrow X$ ($n \times p$).

$$Y = X\beta + \varepsilon$$

$\varepsilon|X \sim \text{zero mean}$

$$\text{Var}[\varepsilon|X] = \sigma^2 I_n$$

Typically, treat $x_1, \dots, x_n$ as fixed (non random) quantities.

—OLS etc.

However with X random

$$Y = X\beta + \varepsilon.$$

$$\therefore X^T Y = X^T X \beta + X^T \varepsilon.$$

$$\therefore E[X^T Y] = E[X^T X] \beta + E[X^T \varepsilon]$$

Usual assumption is

$$E[X^T \varepsilon] = 0$$

i.e

$$E[X^T (Y - X\beta)] = 0.$$

- here the expectation is over the joint distⁿ of (X, Y) .

Estimating equation. $\sum_{i=1}^n x_i^T (y_i - x_i \beta) = 0_p.$

OLS solution is

$$\hat{\beta}_n = (X^T X)^{-1} X^T y$$

i.e. $(X^T X) \hat{\beta} = X^T y$.

or $\left(\frac{1}{n} \sum_{i=1}^n x_i^T x_i \right) \hat{\beta}_n = \frac{1}{n} \sum_{i=1}^n x_i^T y_i$

By standard theory

$$\hat{\beta}_n \xrightarrow{p} \beta_0$$

β_0 - 'true value'.

as $n \rightarrow \infty$, and

~~and~~

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N_p(0, \sigma^2 V)$$

$$V = \lim_{n \rightarrow \infty} \left(\frac{1}{n} X^T X \right)$$

— holds under correct specification.

i.e. that

$$E[Y|X] = X\beta$$

$Y - X\beta$ is uncorrelated with (columns of) X .

i.e. we must have the correct explanatory model for these results to hold.

If $\hat{y}$ is the fitted value vector—

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy \text{ say}$$

$$\hat{y} = Hy \quad \text{in r.v. form.}$$

eg $p=3$.

$$E[Y_i | x_{i1}, x_{i2}] = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} \cdot x_{i2}.$$

- here

$$\begin{aligned} E[Y_i | x_{i1}, x_{i2}] - E[Y_i | 0, x_{i2}] \\ = \beta_1 + \beta_{12} x_{i2}. \end{aligned}$$

- depends on x_{i2} .

However, we can average this over x_{i2} to get an 'average' measure of the effect of changing x_{i1} from 0 to 1.

i.e

$$\frac{1}{n} \sum_{i=1}^n [E[Y_i | 1, x_{i2}] - E[Y_i | 0, x_{i2}]]$$

$$= \beta_1 + \beta_{12} \left(\frac{1}{n} \sum_{i=1}^n x_{i2} \right) = \beta_1 + \beta_{12} \bar{x}_2.$$

This requires that we correctly specify

$$E[Y_i | x_i]$$

otherwise the average will be incorrectly estimated.

In general for binary treatment z_i

$$\begin{aligned} E[Y_i | x_i] &= \mu(0, x_i) + z_i \mu(1, x_i) \\ &= (1 - z_i) \mu_0(x_i) + z_i \mu_1(x_i). \end{aligned}$$

$$\hat{y}_i = \sum_{j=1}^p x_{ij} \hat{\beta}_j = x_i \hat{\beta}$$

$$\therefore \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \left(\frac{1}{n} \sum_{i=1}^n x_i \right) \hat{\beta}$$

~~~~~

'average prediction'

across  $x$  values.

NOTE This is a 'conditional' model

$\therefore$  consequently, parameters and parameter estimates are conditional on the model proposed/fitted.

i.e.  $\beta_j$  is the (modelled) change in the value of  $E[Y_i/x_i]$  when  $x_{ij}$  changes from 0 to 1 in value

assuming all other components of  $x_i$  do not change

eg  $p=2$

$$E[Y_i | X] = \beta_1 x_{i1} + \beta_2 x_{i2}.$$

i.e.  $E[Y_i | x_{i1}, x_{i2}] = \beta_1 x_{i1} + \beta_2 x_{i2}.$

$$E[Y_i | 0, x_{i2}] = \beta_2 x_{i2}.$$

$$E[Y_i | 1, x_{i2}] = \beta_1 + \beta_2 x_{i2}.$$

$$\therefore E[Y_i | 1, x_{i2}] - E[Y_i | 0, x_{i2}] = \beta_1.$$

NOTE This assumes we can change  $x_{i1}$  from 0 to 1 without changing  $x_{i2}$ .

- not possible in 'interaction' models.

In parags we have used the 'Monte Carlo principle'.

- we may 'estimate'

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

by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \int x d\hat{F}_n(x).$$

where  $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}(x)$  ~~where~~

- puts mass  $\frac{1}{n}$  at  $\{x_1, \dots, x_n\}$ .

i.e we 'approximate'  $f_X(x)$  by  $\hat{f}_n(x)$ ,  
the set of masses  $\frac{1}{n}$  at  $x_1, \dots, x_n$ .

In monte Carlo, we generate the  $x$ 's.

i.e to approximate

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

we • sample  $x_1, \dots, x_N$  from  $f_X(x)$ .

• estimate the integral by

$$\hat{I}_N(g) = \frac{1}{N} \sum_{i=1}^N g(x_i).$$

By standard theory, for the estimator.

$$\hat{I}_N(g) \xrightarrow{P} I(g)$$

If we 'cannot' sample from  $f_x$ , we can write

$$I(g) = \int g(x) f_x(x) dx$$

$$= \int g(x) \frac{f_x(x)}{f_0(x)} f_0(x) dx.$$

(provided  $f_x(x) > 0 \Rightarrow f_0(x) > 0 \quad \forall x$ ).

Then we can

- sample  $x_1, \dots, x_N$  from  $f_0$ .

- estimate the integral by

$$\hat{I}_N(g; f_0) = \frac{1}{N} \sum_{i=1}^N \frac{g(x_i) f_x(x_i)}{f_0(x_i)}.$$

'Importance sampling'.  $\hat{I}_N(g; f_0) \xrightarrow{p} I(g)$   
as  $N \rightarrow \infty$ .

Depending on  $g$  and  $f_X$ , Importance Sampling can be more efficient.

i.e. the variance of  $\hat{I}_N(g; f_0)$  can be lower than that of  $\hat{I}_N(g)$ .

What is the relevance of this?

1) "Statistical Inference"

$\equiv$  'Monte Carlo'.

i.e. we use samples to approximate/estimate population quantities.

NOTE  $x, y, z \in \{0, 1\}$ .

$$f_{x,y,z}(x, y, z) = f_{y|x,z}(y|x, z) f_{z|x}(z|x) f_x(x).$$

$$f_{x,y,z}^*(x, y, z) = f_{y|x,z}(y|x, z) f_z^*(z) f_x(x).$$

$$\text{i.e. in } f_{x,y,z}^* \quad X \perp Z$$

Suppose  $f_z^*(z)$  is degenerate at  $z_0 \in \{0, 1\}$ .

$$\text{i.e. } P_z^*[Z = z_0] = 1.$$

Consider

$$E_{y|z}^*[Y|Z=z_0] = P_{y|z}^*[Y=1|Z=z_0].$$

$$= P_{y|x,z}^*[Y=1|X=0, Z=z_0] P_{x|z}^*[X=0|Z=z_0]$$

$$+ P_{y|x,z}^*[Y=1|X=1, Z=z_0] P_{x|z}^*[X=1|Z=z_0]$$

$$= P_{Y|X,Z}[Y=1|X=0, Z=z_0] P_X[X=0] \\ + P_{Y|X,Z}[Y=1|X=1, Z=z_0] P[X=1]$$

$$= P_{Y|X,Z}[Y=1|X=0, Z=z_0] P_X[X=0] P_Z^*[Z=z_0] \\ + P_{Y|X,Z}[Y=1|X=1, Z=z_0] P_X[X=1] P_Z^*[Z=z_0].$$

$$= P_{Y|X,Z}[Y=1|X=0, Z=z_0] \frac{P_X[X=0] P_Z^*[Z=z_0]}{P_X[X=0] P_{Z|X}[Z=z_0|X=0]} \cdot P_X[X=0] P_{Z|X}[Z=z_0|X=0]$$

$$+ \frac{P_{Y|X,Z}[Y=1|X=1, Z=z_0] P_X[X=1] P_Z^*[Z=z_0]}{P_X[X=1] P_{Z|X}[Z=z_0|X=1]} \cdot P_X[X=1] P_{Z|X}[Z=z_0|X=1]$$

$$= E_{X|Z} \left[ E_{Y|X,Z}[Y|X,Z] \frac{P_Z^*[Z=z_0]}{P_{Z|X}[Z=z_0|X]} \mid Z=z_0 \right]$$

$$\frac{P_{Y|X,Z}[Y=1|X,Z] \cdot P_Z^*[Z=z_0]}{P_{Z|X}[Z=z_0|X]}$$

$$P_Z^*[Z=z_0] \equiv f_Z^*(z_0) \text{ etc}$$

That is, we may write

$$E_{Y|Z}^* [Y | Z=z_0] = E_{X|Z} \left[ E_{Y|X,Z} [\cancel{Y} | X, Z] \frac{p_z^* [Z=z_0]}{p_{Z|X} [Z=z_0 | X]} \right] \Big|_{Z=z_0}$$

provided

$$p_{Z|X} [Z=z_0 | X=0] > 0$$

$$p_{Z|X} [Z=z_0 | X=1] > 0.$$

Data generating model:  $f_{X,Y,Z}$ .

Model of interest:  $f_{X,Y,Z}^*$ .

i.e. we can use reweighting to convert one prob. to another.

This akin to 'standardization' in biostat/epi

-(N) DIRECT STANDARDIZATION

- reweight category specific probs (rates etc) by 'standardizing' probs/weights.

- here we are 'standardizing' the  $f_{x,y,z}$  (observed) conditional expectations to the target ( $f_{x,y,z}^*$ ) distribution.

# Chapter 1 CAUSAL GRAPHS.

A joint distribution <sup>structure</sup> is essentially specified by the set of conditional dist<sup>n</sup> probability statements that allow the general chain rule factorization to be made more specific.

eg  $f_{x,y,z}(x,y,z) = f_x(x) f_{y|x}(y|x) f_{z|x,y}(z|x,y)$

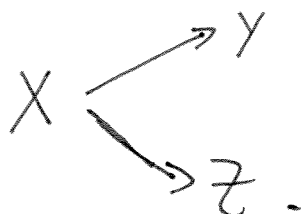
works in general, but perhaps we assume

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

conditional independence.

i.e

$$f_{x,y,z}(x,y,z) = f_x(x) f_{y|x}(y|x) f_{z|x}(z|x)$$



$$Y \leftarrow X \rightarrow Z.$$

Are  $Y$  and  $Z$  unconditionally independent?

- NO.

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

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

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

The simple graphical diagrams can be useful in describing (causal) dependencies between variables.

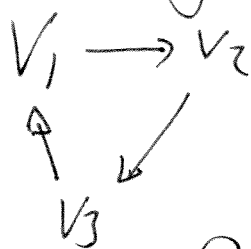
A graph  $G$  comprises a set of nodes or vertices  $V_1, V_2, \dots$  (representing variables) and a set of edges  $E_1, E_2, \dots$  (representing dependencies). Two nodes are adjacent if there is an edge between them.

The edges can be directed or undirected

If all edges are directed, the graph is directed; the graph with the directions removed is called the skeleton.

A path between two nodes  $V_i$  and  $V_j$  is a sequence of edges connecting  $V_i$  and  $V_j$ . (not necessarily following the direction of the directed edges); a directed path is a path where the direction of arrows are obeyed. Two nodes are connected if a path exists between them, disconnected otherwise.

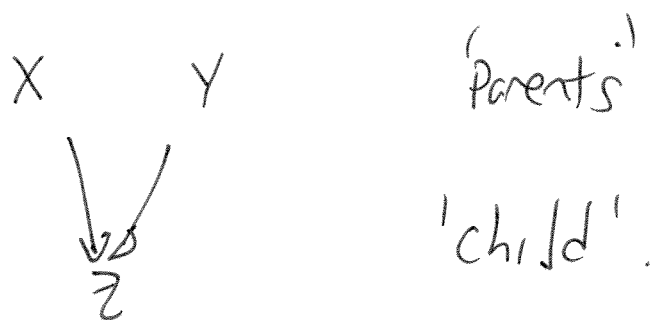
In general directed graphs may contain 'cycles'



where there is a path from a node to itself (with  $> 1$  nodes)

but a directed graph with no cycles is termed a directed acyclic graph (DAG).

The 'language of kinship' may be used to describe connections between edges



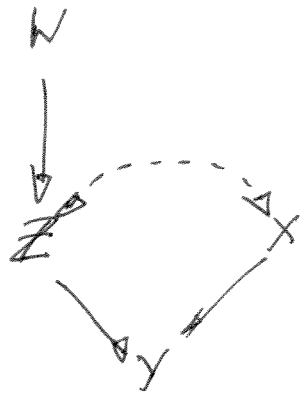
$$f_{X,Y|Z}(x,y,z) = f_X(x) f_Y(y) f_{Z|X,Y}(z|x,y).$$

$$X \perp Y.$$

$$\text{But } f_{X,Y|Z}(x,y,z) = \frac{f_{Z|X,Y}(z|x,y) f_X(x) f_Y(y)}{f_Z(z)}$$

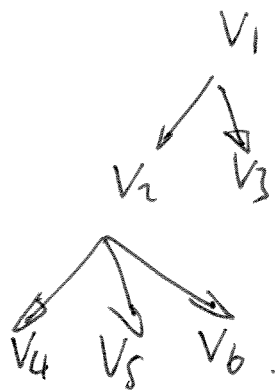
$$\neq f_X(x) f_Y(y)$$

in general. ie  $X \not\perp Y | Z$ .



$X$  and  $Z$  are parents of  $Y$   
 $W, X, Z$  are ancestors of  $Y$ .  
 $Z$  is a child of  $W$   
 $X$  and  $Z$  are spouses  
 $Y$  is a child of  $X$  and of  $Z$  etc.

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



A ~~chain~~ chain is a graph where each node has at most one child  $v_1 \rightarrow v_2 \rightarrow v_3 \dots$

Consider variables  $X_1, \dots, X_n$ .  
(with ordered indices). For each  $j$

$$\text{Let } P_j \equiv \{\text{set of parents of } X_j\} \\ \equiv \{X_1^{(j)}, \dots, X_{n_j}^{(j)}\} \text{ say}$$

Then

$$f_{X_j | X_1, \dots, X_{j-1}}(x_j | x_1, \dots, x_{j-1}) \\ \equiv f_{X_j | P_j}(x_j | x_1^{(j)}, \dots, x_{n_j}^{(j)})$$

That is  $X_j$  is conditionally independent of

$$X_k \in P_j' \text{ given } \{X_1^{(j)}, \dots, X_{n_j}^{(j)}\}$$

and no (proper) subset of  $P_j$   
has this property.

i.e.  $X_1^{(j)}, \dots, X_{n_j}^{(j)}$  is a minimal set of variables for which conditional independence holds.

In the general factorization.

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = f_{X_1}(x_1) \prod_{j=1}^n f_{X_{n_j}}(x_j | x_1, \dots, x_{j-1})$$

$$\equiv f_{X_1}(x_1) \prod_{j=1}^n f_{X_{n_j}}(x_j | x_1^{(j)}, \dots, x_{n_j}^{(j)})$$

Given the set of conditional independencies we can draw a DAG to represent the joint distribution structure.

$$(X_1, X_2) \quad X_1 \longrightarrow X_2 \quad \Leftrightarrow (X_1, X_2) \text{ dependent.}$$

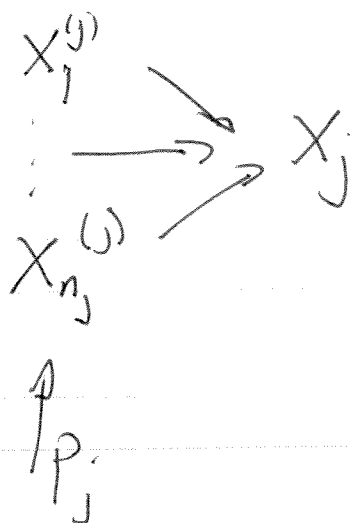
$$X_3 : X_3 \perp (X_1, X_2) \quad \text{no arrow}$$

$$\text{or} \quad \begin{array}{c} X_1 \\ X_2 \end{array} \longrightarrow X_3 \quad X_3 \perp X_1 \mid X_2$$

$$\text{or} \quad \begin{array}{c} X_1 \\ X_2 \end{array} \longrightarrow X_3 \quad X_3 \perp X_2 \mid X_1$$

$$\text{or} \quad \begin{array}{c} X_1 \\ X_2 \end{array} \longrightarrow X_3 \quad \text{general dependence.}$$

$X_j$



## 'Compatibility'

The probability distribution  $P_X$  is 'compatible' with graph  $G$  if  $P_X$  admits the factorization implied by  $G$ .

— note that  $G$  does not define  $P_X$ , merely the chain rule factorization that  $P_X$  admits.

— 'Markov compatibility' 'P is Markov wrt G'.

i.e. we may deduce from  $G$  that

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

but we do not know the numerical values of  $f_X(x)$ ,  $f_Y(y)$  etc.

NB Using the set of conditional probabilities that pertain is equivalent to defining  $G$ .

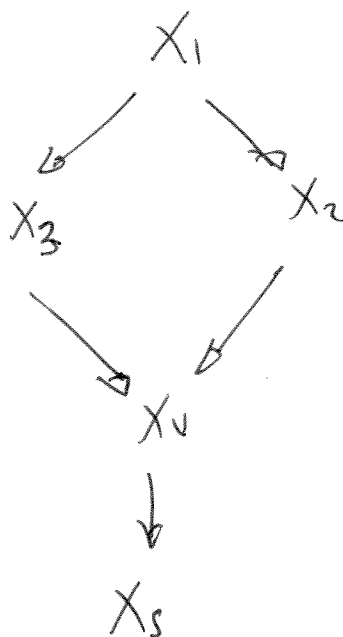
i.e  $G$  is a useful, but not strictly necessary tool for probabilistic specification.

Note need all conditional probabilities defined

→ guaranteed if

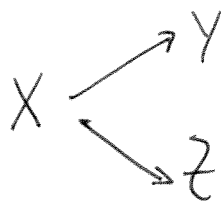
$$f_{X_1 \dots X_n}(x_1, \dots, x_n) > 0 \quad \forall (x_1, \dots, x_n) \in \mathbb{R}^n$$

eg



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

What does the symbol  $\rightarrow$  mean?



- generate  $X$  (independently)
- generate  $Y$  as a function of  $X$
- generate  $Z$  as a function of  $X$ .

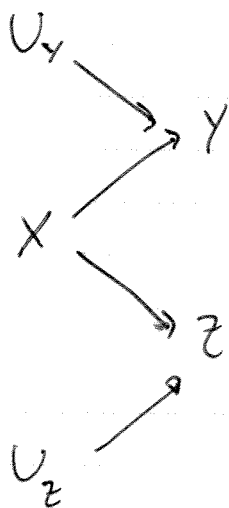
eg  $X$

$$Y = 3X$$

$$Z = 4X + 9.$$

$$Y = f_1(X)$$

$$Z = g_2(X).$$

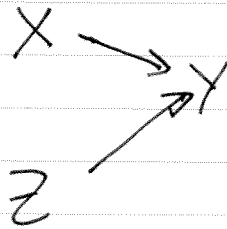


$X, U_X, U_Y$  independent

$$Y = g_1(X, U_Y) (= 3X + U_Y)$$

$$Z = g_2(X, U_Z) (= 4X + 9 + U_Z)$$

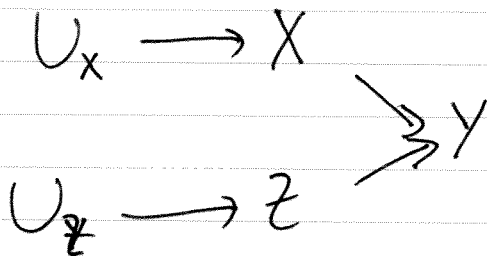
etc.



$X, Z$  - independent

$$Y = g(X, Z) (= 2X + 4Z)$$

- if we fixed  $X$  and  $Z$ , we would determine  $Y$  precisely.



$U_X, U_Z$  independent

$$X = g_1(U_X)$$

$$Y = g_2(U_Y)$$

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

NOTE

Fixing  $X$  and  $Z$  precisely determines  $Y$ .

i.e. if we know the values of  $X$  and  $Z$   
we do not need to know the values of  
 $U_X$  and  $U_Z$ .

$$\text{i.e. } Y \perp (U_X, U_Z) \mid X, Z$$

$$Y \perp U_X \mid X$$

$$Y \perp U_Z \mid Z$$

We can interpret 'causation' in terms of these functions

— 'X causes Y if it appears in the function  $g$  that assigns  $Y$ 's value' (i.e. if it is an ancestor of  $Y$ ), ~~if~~ if  $Y$  is a descendant of  $X$

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  $Y$ .

- 'structural (causal) models'
- 'graphical (causal) model'

Variables that have no 'causes' (ancestors) are termed 'exogenous'; the variables that have a cause are 'endogenous'.

It is clear that we can read off the collection of ~~compatible~~ conditional distributions from a DAG  $G$ .

D-separation  $\rightarrow$

Consider three sets of nodes  $X, Y, Z$  which are disjoint subsets of the nodes of a DAG  $G$ . To test whether

$$X \perp Y \mid Z$$

for any  $X \in X, Y \in Y, Z \in Z$  for a given distribution compatible with  $G$ .  
we must assess whether  $Z$  'blocks' paths for  $X$  to  $Y$ .

## D-separation and Conditional Independence

We have seen the importance of the 'collider' node in a path

$$X \rightarrow Z \leftarrow Y$$

- $Z$  is a 'collider' as it receives two arrows of different directions

- 'inverted fork'

A directed path from node  $V_i$  to node  $V_j$  cannot contain a collider; all parts of the path must be either

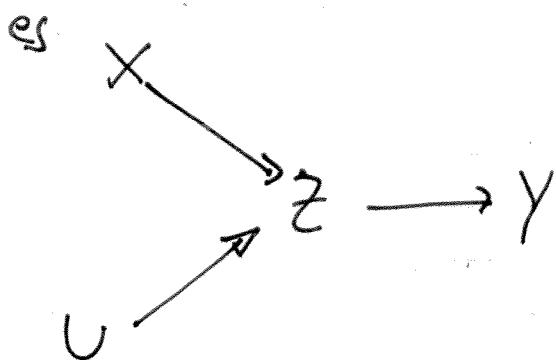
- forks

$$X \leftarrow Z \rightarrow Y$$

- chains

$$X \rightarrow Z \rightarrow Y$$

NOTE The notion of being a collider is path specific



Z is a collider on path ~~(X, Z, U)~~ (X, Z, U).

Z is not a collider on path (X, Z, Y).

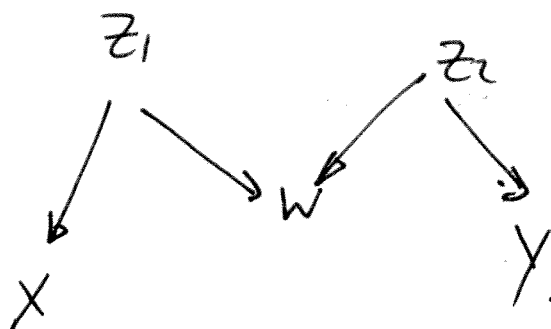
### Unconditional d-separation

A path is open (or unblocked, or active) unconditionally if there is no collider on the path; otherwise it is closed (or blocked, or inactive) — the collider blocks the path.

Two variables X and Y are d-separated

if there is no open path between them; otherwise they are d-connected.

EXAMPLE (Bhman et al p188).



$X$  and  $Y$  are d-separated. There is one path between them, but it is blocked at  $W$ .

$Z_1$  - family income

$Z_2$  - genetic risk

$W$  - parent diabetes

$X$  - low education

$Y$  - diabetes of subject

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

$\Rightarrow X$  and  $Y$  independent.

## Conditional d-separation

We consider conditional statements obtained after conditioning on a variable

For a non-collider:

$$X \rightarrow Z \rightarrow Y$$

$$X \perp Y / Z.$$

$$X \leftarrow Z \rightarrow Y$$

$$X \perp Y / Z.$$

~~z~~ z 'blocks' the path if we condition upon it

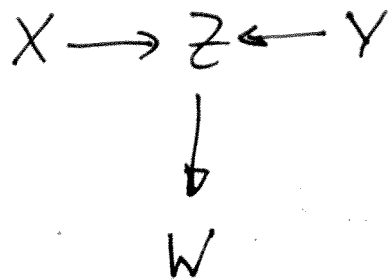
For a collider:

$$X \rightarrow Z \leftarrow Y$$

$$X \not\perp Y / Z.$$

-already established.

However, consider



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

$$X \perp Y | W ?$$

- no!

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

$$= f_X(x) f_Y(y) g(x,y,w) \quad \text{say.}$$

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

→ W is now a collider !!

Therefore

(i) conditioning on a noncollider  $Z$  blocks the path at  $Z$ .

(ii) Conditioning on a collider  $Z$  or a descendant  $W$  of a collider opens the path at  $Z$ ..

i.e conditioning reverses its status

- conditioning ~~closes~~ noncolliders  
opens colliders

Suppose  $S$  is a set of variables.

$S$  blocks a path from  $X$  to  $Y$  if, after conditioning on  $S$ , the path is closed; it 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.

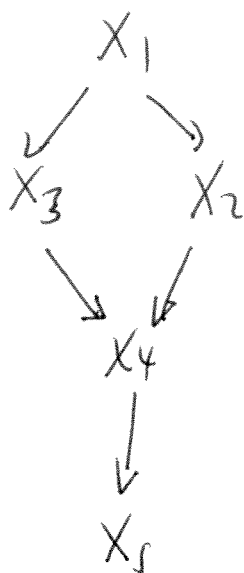
• if  $S$  d-separates  $X$  and  $Y$ , then

$$X \perp Y / S$$

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

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  $y$ , for some  $(x,y,s)$ .



$\{X_2\}$  and  $\{X_3\}$  are d-separated by  $\{X_1\}$ .

- there are two paths between  $X_2$  and  $X_3$

$$X_2 \leftarrow X_1 \rightarrow X_3 \quad \text{fork}$$

$$X_2 \rightarrow X_4 \leftarrow X_3 \quad \text{collider}$$

but  $X_4 \notin \{X_1\}$ .

$$\therefore X_2 \perp X_3 \mid X_1$$

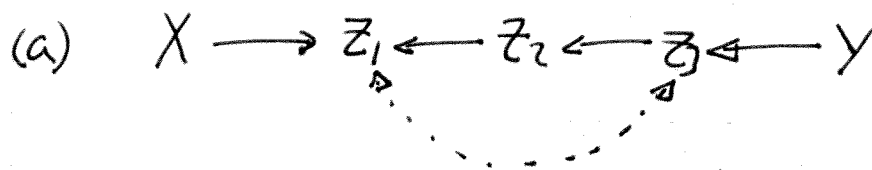
$\{X_2\}$  and  $\{X_3\}$  are not d-separated by  $\{X_1, X_5\}$

- $X_5$  is a descendant of collider  $X_4$ .

$$X_2 \not\perp X_3 \mid \{X_1, X_4\}$$

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

→ 'selection bias' 'Berkson bias'.



$$X \perp Y$$

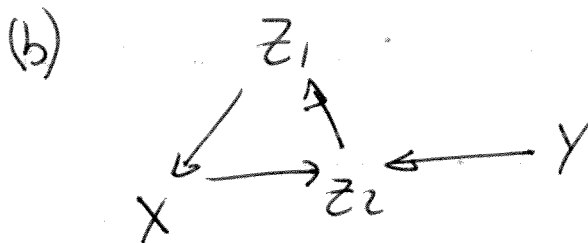
There are two paths that connect  $X$  to  $Y$ .

$$(X, Z_1, Z_2, Z_3, Y)$$

$$(X, Z_1, Z_3, Y)$$

-however both are blocked as  $Z_1$  and  $Z_3$  are colliders.

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



two paths connecting  $X$  to  $Y$

$(X, Z_1, Z_2, Y)$

$(X, Z_2, Y)$

but both involve collider  $Z_2$

—  $X \not\perp Y \mid Z_2$

If  $X$  and  $Y$  are separated by  $S$  ~~by~~ on  $G$ .

then  $X \perp Y \mid S$  for all distributions compatible

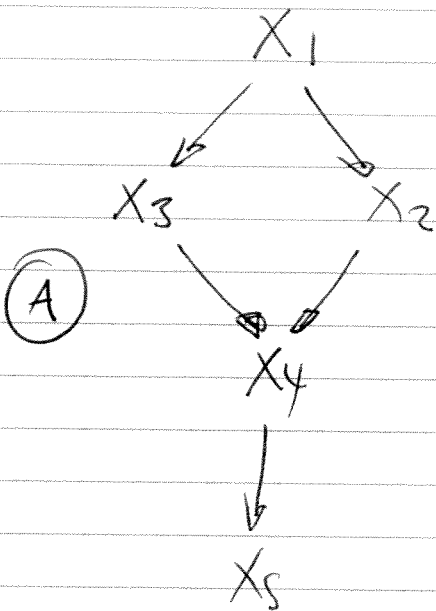
with  $G$ . Conversely, ~~if~~ if they are not

separated, then  $X$  and  $Y$  are dependent

given  $S$  for at least one distribution

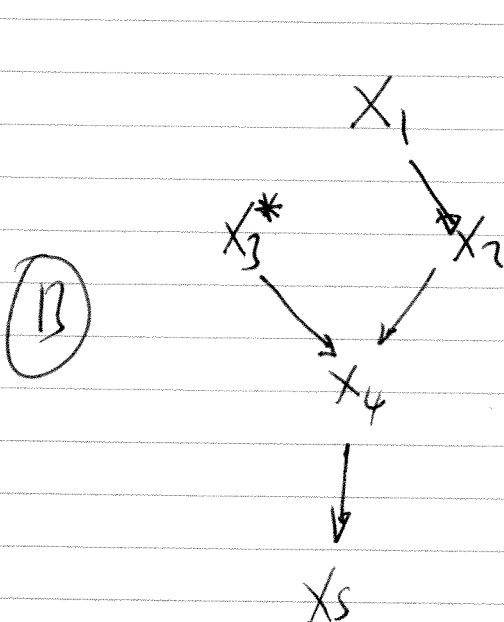
compatible with  $G$ .

# Intervention



$$\begin{aligned}
 & f_{X_1, X_2, X_3, X_4, X_5}(x_1, x_2, x_3, x_4, x_5) \\
 & f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_1}(x_3|x_1) \\
 & \times f_{X_4|X_2, X_3}(x_4|x_2, x_3) \\
 & \times f_{X_5|X_4}(x_5|x_4)
 \end{aligned}$$

How do we represent the 'intervention' to determine the value of  $X_3$  as  $x_3^*$ .



$$\begin{aligned}
 & f_{X_1, X_2, X_3, X_4, X_5}^*(x_1, x_2, x_3^*, x_4, x_5) \\
 & f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) \times f_{X_3}^*(x_3^*) \\
 & f_{X_4|X_2, X_3}(x_4|x_2, x_3^*) \quad \text{degenerate at } x_3^* \\
 & f_{X_5|X_4}(x_5|x_4)
 \end{aligned}$$

Note Pearl denotes this using the  
'do' operator.

i.e. distinguishes between

$$f_{X_1, X_2, X_4, X_5 | X_3}^*(x_1, x_2, x_4, x_5 | x_3^*) \quad (\text{regular condition},$$

and

$$f_{X_1, X_2, X_4, X_5 | X_3^*}^*(x_1, x_2, x_4, x_5 | x_3^*) \quad (\text{intervention}).$$

by writing

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

or

$$P[X_1, X_2, X_4, X_5 | \text{do}(X_3) = x_3]$$

etc.

That is, the notation

$$do(X_3) = X_3^*$$

means

- (i) switch from graph (A) to graph (B)
- (ii) set  $X_3^* = X_3$

NOTE The  $*$  is merely attached to  $X_3$  to denote it is no variable being intervened upon.  $X_3^*$  is merely  $X_3$ !

NOTE Recall the earlier discussion of

$$f_{X_1, X_2, Z}(x, y, z) = f_X(x) f_{Z|X}(z|x) f_{Y|X, Z}(y|x, z)$$

so

$$f_{X_1, X_2, Z}^*(x, y, z) = f_X(x) f_Z^*(z) f_{Y|X, Z}(y|x, z)$$

## GRAPHICAL REPRESENTATION OF DIAS

We aim to understand the effect of  $Z$  on  $Y$  say.

- an undirected path between  $Z$  and  $Y$  that is open allows for the association between  $Z$  and  $Y$  to be modified by the presence of other variables — a 'biasing' path.

- 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

$$Z \rightarrow Y$$

$$Z \rightarrow X \rightarrow Y$$

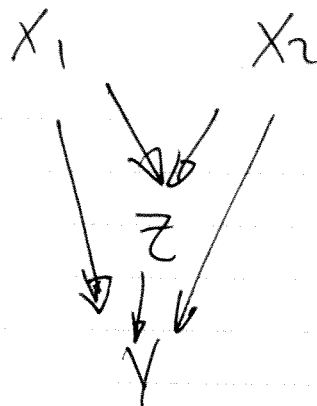
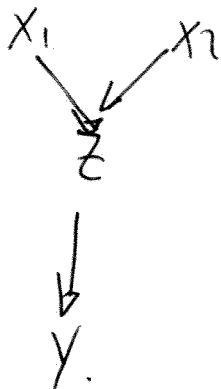
$$\begin{array}{ccccc} Z & \rightarrow & X_1 & \rightarrow & Y \\ & & \searrow & & \\ & & X_2 & \rightarrow & \end{array}$$

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  $Y$ .

—  $Z$  and  $Y$  are unbiased given  $S$ .

— 'minimally sufficient' if no proper subset of  $S$  is sufficient.

NOTE The set of parents of a node is always sufficient in an unconditional DAG. However this set may not be minimally sufficient.



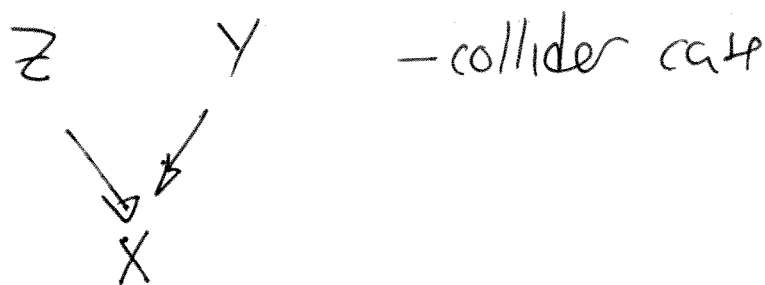
NOTE Conditioning on descendants of  $Z$

(i) blocks directed paths

$$Z \rightarrow X \rightarrow Y$$

$$- Z \perp Y / X \quad \text{but } Z \not\perp Y$$

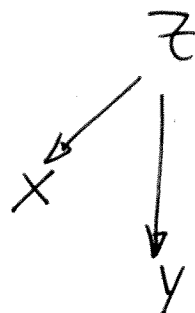
(ii) can unblock or create paths that can lead to biasing of the effect of  $Z$  on  $Y$ .



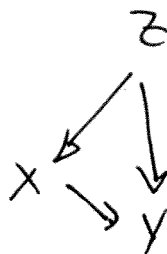
conditioning on a descendant of  $Z$  and  $Y$  creates 'Serkson' bias or 'selection' bias.

-  $Z$  and  $Y$  become related given  $X$ .

(iii) may be spurious in a statistical sense.



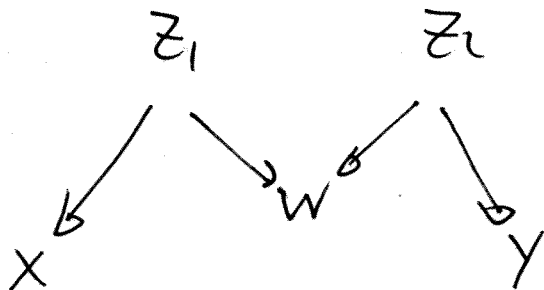
TRUE



MODELLLED.

Undirected paths from  $Z$  to  $Y$  are termed 'back door' paths relative to  $Z$  if they start with an arrow pointing into  $Z$ .

eg



only path ~~is~~ for  $X$  to  $Y$  is a back door path.

Before conditioning, all ~~backdoor~~<sup>biasing</sup> paths  
in a DAG are ~~backdoor~~<sup>biasing</sup> backdoor paths,  
and all open backdoor paths are biasing  
paths.

∴ to obtain the effect of  $Z$  on  $Y$   
all backdoor paths between  $Z$  and  $Y$   
must be blocked.

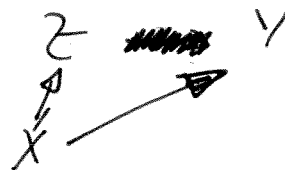
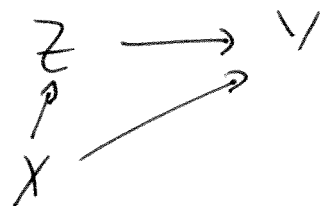
- a set  $S$  satisfies the backdoor  
criterion<sup>with  $Z$  and  $Y$</sup>  if  $S$  contains no descendent  
of  $Z$  and there is no open backdoor  
path from  $Z$  to  $Y$  after conditioning on  
 $S$ .

- conditioning on  $S$  allows identification  
of the (causal) effect of  $Z$  on  $Y$ .

A confounding path is a <sup>between Z and Y</sup> biasing path that ends with an arrow into Y.

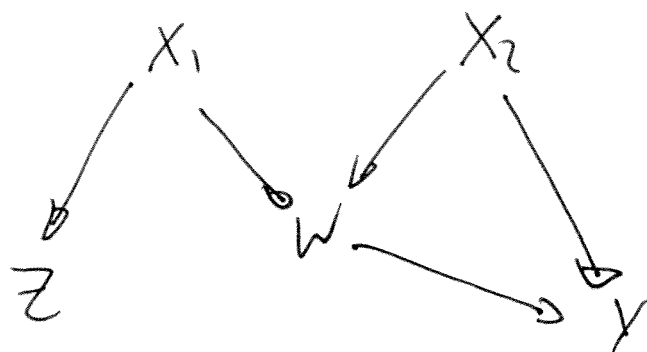
(biasing path - undirected open path).

- variables on a confounding path are confounders.



X is a confounder in both cases.

EXAMPLE



$W$  is a collider on one path.

However, unconditionally on  $W$   
the  $Z, Y$  relationship is confounded  
by the path (backdoor)

$Z - X_1 - W - Y$        $W$  is a confounder.

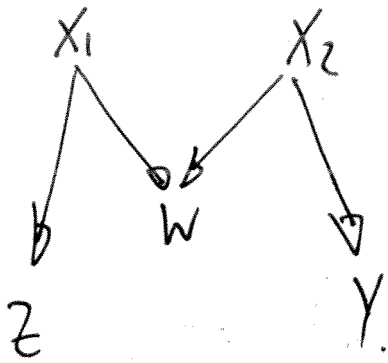
Conditioning on  $W$  opens the path

$Z - X_1 - W - X_2 - Y$

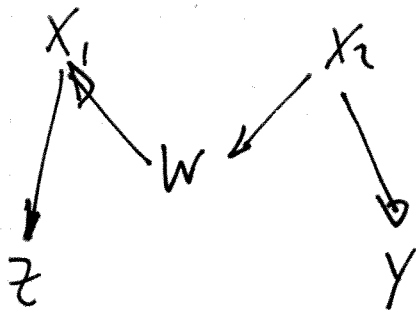
$\therefore$  need to condition upon  $X_2$  also.

otherwise conditioning on  $W$  will not  
remove bias.

More to follow later.



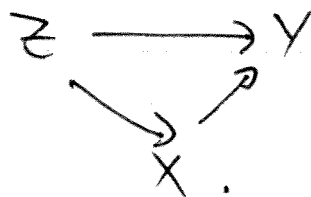
conditioning on  $W$   
opens the path from  $Z$  to  $Y$ .



conditioning on  $W$   
blocks the path.

$\Rightarrow \{X_1\}, \{W\}, \{X_2\}$   
would be sufficient to  
block.

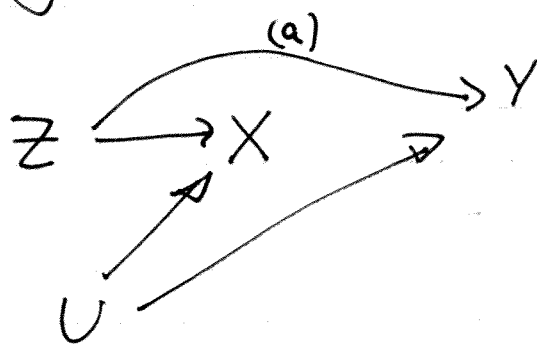
Direct and indirect effects



A direct effect of ~~Z~~ Z on Y

(relative to X) is the effect of Z on Y that is captured in a <sup>directed</sup> path from Z to Y that does not pass through the 'intermediate' or 'mediator' variable X.

An indirect effect of Z on Y is captured by directed paths that pass through X.



- unconfounded direct effect of Z on Y (path (a))

- no indirect effect of Z on Y

## NOTE

$$(i) \quad Z \longrightarrow X \longrightarrow Y$$

but  $X$  held constant at  $X=x$ .

- 'controlled' direct effect of  $Z$  on  $Y$ .

$$(ii) \quad Z \longrightarrow X \longrightarrow Y.$$

but  $Z$  held constant at  $Z=z$ ,

with  $X$  taking its possible values

- 'natural' direct effect of  $Z$  on  $Y$   
for each possible  $X=x$  value.

## Recap

For every joint distribution on  $n$  variables we can construct an  $n$ -node graph, defined by the collection of conditional distributions

$$f_{X_j | PA_j}(x_j | x_1^{(j)}, \dots, x_{n_j}^{(j)})$$

To assess dependence between  $X_i$  and  $X_j$  we consider paths between the variables in the graph.

- We may be able to manipulate dependence by conditioning

(i.e. by blocking backdoor paths).

## 2. CAUSAL EFFECTS

The causal effect of changing variable  $Z$  on variable  $Y$  is the (expected) change in (some function) of  $Y$  as  $Z$  is intervened upon to change  $Z$  from  $Z=z_0$  to  $Z=z_1$ , independently of all other variables,  $X$ .

In our earlier notation, we consider the distribution

$$f_{x,y,z}^*(x,y,z) = f_x(x) f_z^*(z) f_{y|x,z}(y|x,z).$$

and then imagine setting  $z=z_0, z=z_1$ .

### Counterfactual/Potential Outcome Notation

The counterfactual (or potential outcome) notation is used to formulate inference procedures for causal effects.

Let  $Y(z)$  denote the random variable recording the 'outcome'  $Y$  observed if we intervene to set  $Z=z$ .  $Y(z)$  is the potential outcome

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

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

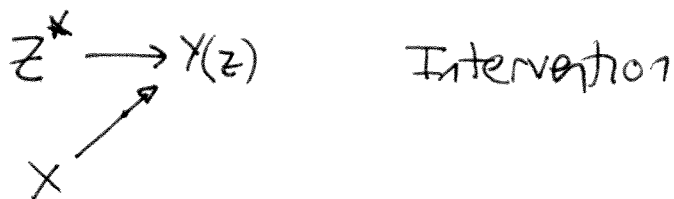
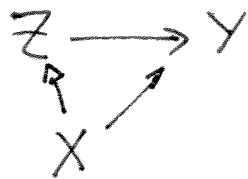
eg  $z \in \{0,1\}$  we have

$Y(0)$  - outcome if intervene to set  $z=0$

$Y(1)$  - outcome if intervention to set  $z=1$

If  $Y$  is the observed outcome, then

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



Note  $Y(0)$  and  $Y(1)$  are different hypothetical random variables.

We must consider, for example

$$\begin{aligned}
 & \cancel{f_{Y(1)|X,Z}(y|x,z)} \\
 & \cancel{f_{Y(0)|X,Z}(y|x,z)} \quad f_{Y(z)|X}(y|x)
 \end{aligned}$$

where

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

$$f_{Y(0)|X}(y|x) \equiv f_{Y(0)|X,Z}(y|x,z) \equiv f_{Y|X,Z}(y|x,0).$$

Then

$$\begin{aligned} f_{Y(1)}(y) &= \int f_{Y(1)|X}(y|x) f_X(x) dx \\ &\equiv \int f_{Y|X,Z}(y|x, 1) f_X(x) dx \end{aligned}$$

$$f_{Y(0)}(y) = \int f_{Y|X,Z}(y|x, 0) f_X(x) dx$$

In the original model

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

i.e. once  $X$  and  $Z$  are fixed we generate  $Y$

But the intervention fixes  $Z$ .

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

Could consider the causal effect of  $Z$  on  $Y$  to be some 'contrast' defined via the  $Y(z)$  values.

eg  $Y(1) - Y(0)$      $Y(z_1) - Y(z_0)$

$\log \frac{Y(1)}{Y(0)}$     etc.

etc. Such quantities are random variables.

The difficulty with this 'individual' contrast strategy is that we will almost never have the opportunity to collect data from such a data generating model.

i.e. we will never observe  $Y$  under both  $Z=1$  and  $Z=0$  in a data set.

## Experimental Studies

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

EXAMPLE  $2n$  individuals sampled from a homogeneous population.

$n$  - receive  $Z = 0$

$n$  - receive  $Z = 1$ .

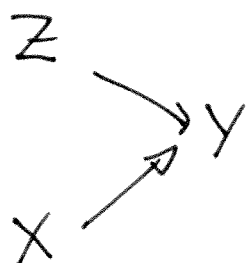
Irrespective of the other characteristics

that are thought relevant to the study  
i.e. the assignment of  $Z$  is independent of  
 $X$

Then the outcome  $Y$  is measured.

Here the data generating mechanism is given precisely by

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



although we need to consider the two halves carefully.

$$f_{Z_0}(z) = \begin{cases} 1 & z=0 \\ 0 & \text{otherwise} \end{cases} \quad \text{if } L \leq n$$

$$f_{Z_1}(z) = \begin{cases} 1 & z=1 \\ 0 & \text{otherwise} \end{cases} \quad \text{if } L > n$$

[No need for  $f_Z^*$  - the intervention is identical to the data generating mechanism]

We may also write

$$f_{x,y,z}^{\varepsilon}(x,y,z) = f_x^{\varepsilon}(x) f_z^{\varepsilon}(z) f_{y|x,z}^{\varepsilon}(y|x,z)$$

for this case.

For a randomized version of the experimental study, we consider

$$f_z^{(\varepsilon)}(z) = \begin{cases} p^z (1-p)^{1-z} & z \in \{0,1\} \quad \forall i \\ 0 & \text{otherwise} \end{cases}$$

here, a subject independently receives  $z=0$  w.p.  $1-p$ , and  $z=1$  w.p.  $p$ .

for some  $0 < p < 1$ .

Note  $N_1$  (the total number of individuals assigned  $Z=1$  after  $n$  individuals have been included)

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

— a random variable.

### A simple causal contrast

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

eg we have  $Y_{11}, \dots, Y_{1n}$  from the  $Z=1$  'arm'  
 $Y_{01}, \dots, Y_{0n}$  from the  $Z=0$  'arm'.

We presume an iid sample (or exchangeability) within each arm.

In ~~the~~ (randomized) study we know that  
we will never observe ~~an~~ an individual's  
response under both  $z=1$  and  $z=0$ .  
so the individual-level contrast is  
immediately replaced by the 'average contrast'

i.e. 
$$[E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$$

say. In our notation, this may  
be written

$$E_{Y|Z}^{\varepsilon}[Y|Z=1] - E_{Y|Z}^{\varepsilon}[Y|Z=0].$$

where the superscript  $\varepsilon$  indicates that  
the expectation is being computed under

$$f_{x,y,z}^{\varepsilon}(x,y,z).$$

$$E_{Y|Z}^{\varepsilon}[Y|Z=z] = \iint y f_{Y|X,Z}^{\varepsilon}(y|x,z) f_X^{\varepsilon}(x) dy dx$$

To estimate this from the <sup>not-random</sup> data, we use the ~~data~~ estimator

$$\frac{1}{n} \sum_{i=1}^n Y_{zi} \quad z=0,1$$

(sample mean each group).

For the randomized study, we have potentially different sample sizes for  $z=0, z=1$  cases

$$E_{Y|Z}^{\varepsilon}[Y|Z=1] = \iint y f_{Y|X,Z}^{\varepsilon}(y|x,1) f_X^{\varepsilon}(x) dy dx.$$

$$\equiv \sum_{z=0}^1 \iint \mathbb{I}_{\{z\}}(z) y f_{Y|X,Z}^{\varepsilon}(y|x,z) f_Z^{\varepsilon}(z) f_X^{\varepsilon}(x) dy dx$$


---


$$f_Z^{\varepsilon}(z)$$

$$= \frac{\sum_{z=0}^1 \iint \mathbb{I}_{\{1\}}(z) y f_{Y|X,Z}^{\varepsilon}(y|x,z) f_z^{\varepsilon}(z) f_x^{\varepsilon}(x) dx dy}{\sum_{z=0}^1 \mathbb{I}_{\{1\}}(z) f_z^{\varepsilon}(z)}$$

Numerator is ~~estimated~~

$$E_{X,Y,Z}^{\varepsilon} [\mathbb{I}_{\{1\}}(Z) Y]$$

Denominator is

$$E_Z^{\varepsilon} [\mathbb{I}_{\{1\}}(Z)]$$

Numerator is estimated by the sample average

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{1\}}(Z_i) Y_i$$

$$= \frac{1}{n} \sum_{i=1}^n z_i y_i$$

Denominator is estimated by

$$\frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{1\}}(z_i) = \frac{1}{n} \sum_{i=1}^n z_i$$

$\therefore$  An estimator of

$$\begin{aligned} E_{Y|Z}^{\mathcal{E}}[Y|Z=1] &= \frac{\frac{1}{n} \sum_{i=1}^n z_i y_i}{\frac{1}{n} \sum_{i=1}^n z_i} = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i} \\ &= \frac{\sum_{i=1}^n z_i y_i}{N_1} \end{aligned}$$

— sample mean amongst those assigned  $Z_i = 1$  in the randomized study. Similarly for  $Z=0$

$$\frac{\sum_{i=1}^n (1-z_i) y_i}{\sum_{i=1}^n (1-z_i)} = \frac{\sum_{i=1}^n (1-z_i) y_i}{N_0}$$

NOTE We have

$$E_z[\Pi_{\{1\}}(z)] = p.$$

$$E_z[\Pi_{\{0\}}(z)] = (1-p)$$

$\therefore$  could use alternative estimators.

$$\frac{1}{np} \sum_{i=1}^n z_i y_i$$

$$\frac{1}{n(1-p)} \sum_{i=1}^n (1-z_i) y_i.$$

Now

$$E_{x,y,z}^{\xi}[z_i y_i] = E_z^{\xi}[E_{y|z}^{\xi}(y|z) z]$$

$$= E_{y|z}^{\xi}[y|z=1] p^{\xi}[z=1]$$

$$= E_{y|z}^{\xi}[y|z=1] p$$