

$\therefore \frac{1}{np} \sum_{i=1}^n z_i Y_i$ is unbiased for $E_{Y|Z}^E[Y|Z=1]$.

$\frac{1}{n(1-p)} \sum_{i=1}^n (1-z_i) Y_i$ is unbiased for $E_{Y|Z}^E[Y|Z=0]$.

Now

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

where

$$w_i = \frac{z_i}{\sum_{j=1}^n z_j}$$

Note $\sum_{i=1}^n w_i = \frac{\sum_{i=1}^n z_i}{\sum_{j=1}^n z_j} = 1.$

$$E_{Z_{1:n}, Y_{1:n}}^{\xi} \left[\sum_{i=1}^n W_i Y_i \right]$$

$$= \sum_{i=1}^n E_{Z_{1:n}, Y_{1:n}}^{\xi} [W_i Y_i]$$

$$= \sum_{i=1}^n E_{Z_{1:n}}^{\xi} \left[E_{Y_{1:n} | Z_{1:n}} [Y_i | Z_{1:n}] W_i \right]$$

$$= \sum_{i=1}^n E_{Z_{1:n}}^{\xi} \left[E_{Y_i | Z_i}^{\xi} [Y_i | Z_i] W_i \right]$$

Now

$E_{Y_i | Z_i} [Y_i | Z_i] W_i$ is zero

unless $Z_i = 1$.

$$\therefore E_{z_{1:n}, Y_{1:n}}^{\varepsilon} \left[\sum_{i=1}^n W_i Y_i \right]$$

$$= \sum_{i=1}^n E_{z_{1:n}}^{\varepsilon} \left[E_{Y_i | z_i}^{\varepsilon} [Y_i | z_i = 1] W_i \right]$$

$$= \sum_{i=1}^n E_{z_{1:n}}^{\varepsilon} [Y_i | z_i = 1] E_{z_{1:n}}^{\varepsilon} [W_i]$$

But

$$E_{z_{1:n}}^{\varepsilon} [W_i] = \frac{1}{n}$$

by symmetry

— we know $\sum_{i=1}^n W_i \equiv 1$

$$\therefore E_{z_{1:n}} \left[\sum_{i=1}^n W_i \right] = 1$$

but the W_i are iid.

∴ expectation is

$$E_{Y|Z}^E[Y|Z=1]$$

as required.

i.e

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

is unbiased for $E[Y(1)] = \mu(1)$
say.

∴ Two possible unbiased estimators

$$\tilde{\mu}(1) = \frac{1}{np} \sum_{i=1}^n z_i Y_i$$

$$\hat{\mu}(1) = \frac{\sum_{i=1}^n z_i Y_i}{\sum_{i=1}^n z_i} = \frac{\sum_{i=1}^n z_i Y_i}{n\hat{p}}$$

∴ An unbiased estimator of

$$\delta = \mu(1) - \mu(0)$$

is either

$$\begin{aligned}\hat{\delta} &= \frac{1}{np} \sum_{i=1}^n z_i y_i - \frac{1}{n(1-p)} \sum_{i=1}^n (1-z_i) y_i \\ &= \frac{1}{np(1-p)} \sum_{i=1}^n (z_i - p) y_i\end{aligned}$$

or

$$\hat{\delta} = \frac{\sum z_i y_i}{\sum z_i} - \frac{\sum (1-z_i) y_i}{\sum (1-z_i)}$$

$$\begin{aligned}&= \frac{1}{(\sum z_i)(\sum (1-z_i))} \left[\left\{ \sum_{i=1}^n (1-z_i) \right\} \sum_{i=1}^n z_i y_i - \left\{ \sum_{i=1}^n (1-z_i) \right\} \sum_{i=1}^n (1-z_i) y_i \right] \\ &= \frac{1}{n\hat{p}(1-\hat{p})} \sum_{i=1}^n (z_i - \hat{p}) y_i\end{aligned}$$

1.e

$$\hat{\delta} = \frac{1}{n} \sum_{i=1}^n \tilde{W}_i Y_i \quad \tilde{W}_i = \frac{Z_i - p}{p(1-p)}$$

$$\hat{\delta} = \frac{1}{n} \sum_{i=1}^n \hat{W}_i Y_i \quad \hat{W}_i = \frac{Z_i - \hat{p}}{\hat{p}(1-\hat{p})}$$

$$\text{Var}[\tilde{W}_i] = \frac{\text{Var}[Z_i]}{(p(1-p))^2} = \frac{1}{p(1-p)}$$

$$\text{Var}[\hat{W}_i]:$$

$$\text{Var}[Z_i - \hat{p}] = \text{Var}[Z_i] + \text{Var}[\hat{p}] - 2\text{Cov}[Z_i, \frac{1}{n} \sum_{j=1}^n Z_j]$$

$$\text{Var}[\hat{p}] = \frac{p(1-p)}{n}$$

$$\text{Cov}[Z_i, \frac{1}{n} \sum_{j=1}^n Z_j] = \frac{1}{n} \text{Var}[Z_i] = \frac{p(1-p)}{n}$$

$$\therefore \text{Var}[\bar{z}_i - \hat{p}] = \text{Var}[\bar{z}_i] - \frac{p(1-p)}{n}$$

$$< \text{Var}[\bar{z}_i]$$

It follows that

$$\text{Var}[\hat{\delta}] > \text{Var}[\hat{\delta}^*]$$

or more usefully

$$\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} N(0, \tilde{V}(\delta))$$

$$\sqrt{n}(\hat{\delta}^* - \delta) \xrightarrow{d} N(0, \hat{V}(\delta))$$

$$\lim_{n \rightarrow \infty} \text{Var}[\sqrt{n}\hat{\delta}] > \lim_{n \rightarrow \infty} \text{Var}[\sqrt{n}\hat{\delta}^*]$$

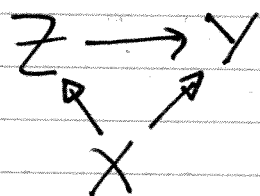
i.e. $\hat{\delta}^*$ is (asymptotically) preferable to $\hat{\delta}$ as it is more efficient.

Observational Studies

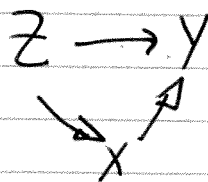
In an observational studies, we observe data on 'exposure' Z , but do not assign it experimentally.

We may denote the data generating mechanism

$$f_{X,Y,Z}^0(x,y,z)$$



(a)



(b)

(a) — backdoor path from Z to Y (arrow into Z).

$$f_X^0(x) f_{Z|X}^0(z|x) f_{Y|X,Z}^0(y|x,z)$$

$$X \not\perp Z$$

→ also a confounding path (arrow into Y).

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

(b) Two paths (directed) from Z to Y

$Z \rightarrow Y$ direct path

$Z \rightarrow X \rightarrow Y$ indirect path.

Conditioning on X here blocks the indirect path.

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

$X \not\perp Z$

Let us try to compute $E_{Y|Z}^0[Y|Z=z]$: first consider

$$E_{Y|Z}^0[Y|Z=z] = \iint y f_{Y|X,Z}^0(y|x,z) f_{X|Z}^0(x|z) dx dy.$$

We can readily assume

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

$$\text{But } f_{X|Z}^0(x|z) \neq f_X^{\varepsilon}(x)$$

in general for all (x, z) .

$$E_{Y|Z}^0[Y|Z=z] = \sum_{z'=0}^1 \iint \mathbb{I}_{\{z\}}(z') y f_{Y|X,Z}^0(y|x,z) f_{X|Z}^0(x|z') f_z^0(z') dx dy$$

$$f_z^0(z).$$

$$= E_{X,Y,Z}^0[\mathbb{I}_{\{z\}}(Z)Y]$$

$$E_Z^0[\mathbb{I}_{\{z\}}(Z)]$$

as before.

$\therefore$ In general

$$E_{Y|Z}^0[Y|Z=z] \neq E_{Y|Z}^{\varepsilon}[Y|Z=z]$$

and the estimator

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

will be biased for $\mu(z) = E_{Y|Z}^{\varepsilon}[Y|Z=z] \equiv E[Y(z)]$

NOTE $f_{X|Z}^0(x|z)f_Z^0(z) = f_{Z|X}^0(z|x)f_X^0(x)$.

(conditioning, Bayes Thm etc)

∴ Can write

$$E_{Y|Z}^0[Y|Z=z] = \sum_{z'=0}^1 \iint \frac{\pi_{Y|Z}(z')}{f_Z^0(z)} f_{Y|X,Z}^0(y|x,z) f_{Z|X}^0(z'|x) f_X^0(x) dx dz'.$$

$$f_Z^0(z).$$

but this is still not quite the due to the

$$f_{Z|X}^0(z|x)$$

term. It is ~~quite~~ the case though that $f_X^0(x) \equiv f_X^E(x)$.

Suppose we know that

$$E_{Y|X,Z}^0[Y|X,Z] = E_{Y|X,Z}^E[Y|X,Z] \\ = \mu(X,Z).$$

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

$\therefore$ an estimator of the two parts are

$$\hat{\mu}_{OR}^{(1)} = \frac{1}{n} \sum_{i=1}^n \mu(X_i, 1) \quad \hat{\mu}_{OR}^{(0)} = \frac{1}{n} \sum_{i=1}^n \mu(X_i, 0)$$

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

This estimator is unbiased. It is based on conditional modelling of Y given X and Z .

$\rightarrow$ an (outcome) regression model (or model).

However, suppose we mistakenly assume

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

Then

$$\hat{\delta} = \frac{1}{n} \sum_{i=1}^n [m(X_i, Z_i) - m(X_i, \bar{Z}_i)]$$

is in general biased. However, if Z is binary we may always write

$$\mu(X, Z) = \mu_0(X) + Z\mu_1(X).$$

$$m(X, Z) = m_0(X) + Zm_1(X).$$

$$\therefore \hat{\delta} = \frac{1}{n} \sum_{i=1}^n [m_1(X_i)]$$

is unbiased provided

$$E_X[m_1(X)] \equiv E_X[\mu_1(X)].$$

i.e if

$$m(x, z) = m_0(x) + z\mu_1(x).$$

$$\begin{aligned} E_x^\pi[m(x, 1) - m(x, 0)] &= E^\pi[\mu_1(x)] \\ &= E^\pi[\mu(x, 1) - \mu(x, 0)] \end{aligned}$$

i.e we can mis-specify $m(x, 0) \equiv m_0(x)$
provided we correctly specify $m_1(x) \equiv \mu_1(x)$

$$m_1(x) = E_x^\pi[\mu(x, 1) - \mu(x, 0)]$$

Can we beat estimator $\hat{\delta}_{0e2}$?

Consider

$$\hat{\delta}_{0e2} = \frac{1}{n\hat{p}} \sum_{i=1}^n z_i \mu(x_i, z_i) - \frac{1}{n(1-\hat{p})} \sum_{i=1}^n (1-z_i) \mu(x_i, z_i)$$

The two terms are comprised of independent
subsamples. with $z_i = 1$ respectively

$$z_i = 0$$

The estimator is unbiased

However, it can be shown that

$$\text{Var}^E[\hat{\delta}_{002}] < \text{Var}^E[\hat{\delta}_{002}]$$

if $\mu(x, z)$ is truly the
conditional expected response.

given X and Z .

Typically we must model $E_{Y|X, Z}^E(Y|X, Z)$

3. SEMIPARAMETRIC INFERENCE

A semiparametric model is a model that is part parametric and part ~~non~~^{not} parametric.

eg $f_x(x; \theta) = f_0(x - \theta)$ (LOCATION FAMILY)

θ - parametric part (finite dimensional)

f_0 - nonparametric part (infinite dimensional)

~~eg $f_x(x; \beta) = f_0(x - \beta)$~~

eg $f_{Y|X}(y|x; \beta) = f_0(y - x\beta)$

$f_X(x)$

parametric : $x\beta$ $\sum_{j=1}^p x_j \beta_j$

nonparametric : f_0, f_X

eg Cox PH model.

Hazard function $h(t) = \frac{f'(t)}{1-F(t)}$

~~$h(t; \beta) = h_0(t) \exp\{x\beta\}$~~

$$h_{y|x}(y|x; \beta) = h_0(y) \exp\{x\beta\} \cdot f_x(x).$$

$$h_0(y) = \frac{f_0(y)}{1-F_0(y)} \quad \text{— 'baseline hazard'!}$$

— nonparametric part

Semiparametric Regression

Recall the linear model estimate by OLS

$$E_{Y|X}[Y|X=x] = X\beta \quad \begin{matrix} X = (n \times p) \\ \beta = (p \times 1) \end{matrix}$$

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n (y_i - x_i \beta)^2$$

$$= \arg \min_{\beta} (y - X\beta)^T (y - X\beta)$$

$$= (X^T X)^{-1} X^T y$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad (n \times p)$$

This is a semiparametric inference method, as the probability distribution of Y given $X=x$ is not specified completely - only the expectation is specified.

Recall, our solve the estimating equation

$$X^T(y - X\beta) = 0_p.$$

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

- This is an orthogonality result.

i.e we choose β to make
 $y - X\beta$

orthogonal to the space spanned by the columns of X

If X is random.

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

ε ($n \times 1$)
- zero mean
- finite variance

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

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

Assuming $E[X^T \varepsilon] = E_x[E_{\varepsilon/x}[X^T \varepsilon/x]] = 0$.

we have at $\beta = \beta_0$ (the true value).

$$E[X^T Y - X^T X \beta_0] = O_p$$

i.e. $E[X^T (Y - X\beta_0)] = O_p$. - orthogonality

In the sample this becomes

$$\frac{1}{n} \sum_{i=1}^n x_i^T (y_i - x_i \hat{\beta}) = O_p$$

$$E[X^T(Y - X\beta_0)] = 0_p$$

$$\Rightarrow E_X \left[E_{Y|X} [X^T(Y - X\beta_0) | X] \right] = 0_p$$

or

$$E_X \left[X^T E_{Y|X} [Y - X\beta_0 | X] \right] = 0_p$$

- could write

$$\varepsilon = Y - X\beta_0$$

We do not specify the distribution of ε .

- we have a location family model.

For a Generalized Linear Model (GLM) consider
 the Exponential-Dispersion family (natural form $E(y)=\eta$,
 canonical parameter θ ,
 dispersion ϕ).

$$f_Y(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right\}$$

$$E_Y[Y] = b(\theta) \quad \text{Var}_Y[Y] = \phi b''(\theta).$$

GLM: set $\theta_i = x_i^T \beta$ $x_i^T, \beta (\phi \times 1)$.

- define model through β which defines the
 mean-variance relation.

$$\begin{aligned} \log f_{Y_i}(y_i; \theta_i, \phi) &= \log f_{Y_i}(y_i; x_i, \beta, \phi) \\ &= \frac{y_i x_i^T \beta - b(x_i^T \beta)}{\phi} + \text{constant} \end{aligned}$$

$$\begin{aligned} \text{Log Likelihood: } \ell(\beta, \phi) &= \sum_{i=1}^n \left(\frac{y_i x_i^T \beta - b(x_i^T \beta)}{\phi} \right) + \text{const.} \end{aligned}$$

Score function

$$\frac{\partial}{\partial \beta} \ell(\beta; \phi) = \sum_{i=1}^n \frac{(y_i x_i^T - x_i^T \dot{b}(x_i \beta))}{\phi}$$

$$= \frac{1}{\phi} \sum_{i=1}^n x_i^T (y_i - \dot{b}(x_i \beta))$$

$$= \frac{1}{\phi} \sum_{i=1}^n x_i^T (y_i - \mathbb{E}_{Y|X}[Y_i | x_i; \beta])$$

$$= 0.$$

- more 'exotic' model ~~uses~~ uses more general model for $\mathbb{E}_{Y|X}[Y_i | x_i; \beta]$

$$\text{i.e. } \mathbb{E}_{Y|X}[Y_i | x_i; \beta] = \dot{b}(0) = \dot{b}(m(x_i \beta)).$$

Essentially this is still a 'orthogonality' result

i.e we solve

$$X^T(y - \mu(x\beta)) = 0$$

for β where

$$\mu(x\beta) = \begin{bmatrix} \mu(x_1\beta) \\ \mu(x_2\beta) \\ \vdots \\ \mu(x_n\beta) \end{bmatrix}$$

—non linear; numeric methods needed
(IRLS, Fisher scoring etc).

We have at $\beta = \beta_0$

$$E_{Y|X} [X^T(Y - \mu(x\beta_0)) | X] = 0.$$

provided $\mu(x\beta_0)$ is the correct mean model.

The estimating equation

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

could also be considered a 'semiparametric' approach.

- we specify

$$E_{Y|X}[Y_i | x_i; \beta] = \mu(x_i \beta)$$

• find the score by $\frac{\partial}{\partial \beta}$

• equate to zero.

The distribution of Y_i is never used.

Constructing a semiparametric model

Consider $Y \in \mathbb{R}$, $X \in \mathbb{R}^p$

$$Y = \mu(X; \beta) + \varepsilon$$

with $\mathbb{E}_{\varepsilon|X}[\varepsilon|X] = 0$.

(X_i, Y_i) , $i=1, \dots, n$ drawn from density.

$$f_{Y,X}(y, x; \beta, f_0)$$

β - $p \times 1$ parameter

f_0 - joint density (infinite dimensional)

As $\varepsilon = Y - \mu(X; \beta)$ we have

$$f_{Y,X}(y, x; \beta, f_0) \equiv f_{\varepsilon, X}(y - \mu(X; \beta), x).$$

The only assumption (beyond assumptions about probability densities) is that

$$\mathbb{E}_{\varepsilon|X}[\varepsilon|X] = 0.$$

We have

$$f_{\varepsilon, x}(\varepsilon, x) = f_{\varepsilon|x}(\varepsilon|x) f_x(x).$$

with

$$f_{\varepsilon|x}(\varepsilon|x) \geq 0 \quad \forall \varepsilon, x.$$

$$\int f_{\varepsilon|x}(\varepsilon|x) d\varepsilon = 1 \quad \forall x.$$

$$\int \varepsilon f_{\varepsilon|x}(\varepsilon|x) d\varepsilon = 0 \quad \forall x.$$

$$f_x(x) \geq 0 \quad \forall x.$$

$$\int f_x(x) dx = 1.$$

1. choose an arbitrary (integrable) function

$$h_0(\varepsilon, x) > 0$$

2. Define

$$h_1(\varepsilon, x) = \frac{h_0(\varepsilon, x)}{\int h_0(t, x) dt.}$$

$h_1(\varepsilon, x)$ is a conditional density for ε given x and

$$\int h_1(\varepsilon, x) d\varepsilon = 1 \quad \forall x.$$

3. The expectation wrt $h_1(\varepsilon, x)$ is

$$\mu(x) = \int \varepsilon h_1(\varepsilon, x) d\varepsilon.$$

$\therefore$ Define the new variable ϵ

$$\epsilon = \varepsilon - \mu(x) \Rightarrow \varepsilon = \epsilon + \mu(x).$$

$$\therefore \mathbb{E}_{\epsilon|x}[\epsilon|x] = \mathbb{E}_{\varepsilon|x}[\varepsilon|x] - \mu(x) = 0.$$

We can deduce the conditional density of ϵ to be, by transformation of variables

$$f_0(\epsilon, x) = h_1\left(\epsilon + \int t h_1(t, x) dt, x\right)$$

By construction

$$\int \epsilon f_0(\epsilon, x) d\epsilon = 0.$$

as ϵ has expectation zero.

$$\text{Also } \int f_0(\epsilon, x) d\epsilon = 1.$$

4 Also, can choose f_x in a similar nonparametric way

$\therefore$ a ~~non~~ semiparametric construction can be based on

$$\beta, \underbrace{f_0(\varepsilon, x), f_x}_{\text{non-parametric.}}$$

and

$$\begin{aligned} f_{Y,X}(y, x; \beta, f_0, f_x) \\ &= f_{Y|X}(y|x, \beta, f_0) f_x(x) \\ &= f_0(y - \mu(x; \beta), x) f_x(x). \end{aligned}$$

Inference :

In a (regular) parametric model, we typically seek estimators $\hat{\beta}_n$ st

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

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Sigma(\beta_0))$$

as $n \rightarrow \infty$, where β_0 is the true data generating model, and $\Sigma(\beta_0)$ is the asymptotic variance.

This theory typically holds under 'correct specification', that is, when the data generating model matches the model used for inference.

In the semiparametric setting, we can examine similar results under the assumption that the semiparametric class considered contains the truth

e.g. f_0 , f_X is 'flexibly' specified such that the true model is captured.

For estimation, we need a framework to mimic estimation procedures in the parametric case.

• Parametric estimation : MLE
Method of moments

• Non parametric:

$$F_X(x) \text{ estimated by } \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{\{X_i \leq x\}}$$

- pointwise, uniformly in x etc.

In LM/GLM we consider optimal estimation.
through the concept of projection

In an LM:

$\mathcal{M}$: 'Model space' - linear combination of
elements of predictor vector.

$$\text{i.e. } \hat{y} = x\beta = \sum_{j=1}^p x_j \beta_j$$

$\mathcal{Y}$: 'Data space' $y \in \mathbb{R}$.

For all data together

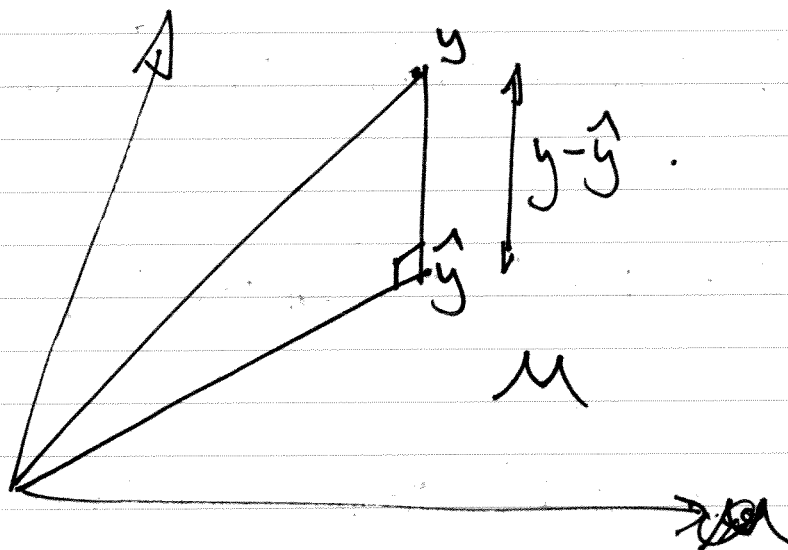
$$\mathcal{M}: \left\{ \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \dots + \beta_p \underline{x}_p, \beta \in \mathbb{R}^p \right\}$$

$$\underline{x}_j = j^{\text{th}} \text{ column of } X$$

$$Y: \{y: y \in \mathbb{R}^n\}.$$

$$\text{Fitted value: } \hat{y} = X\hat{\beta}$$

— 'closest point in M to y '.



By construction,

$y - \hat{y}$ is orthogonal to M .

i.e. orthogonal to all $\underline{x} \in M$

i.e. orthogonal to all $\underline{x}_j$

We 'project' y onto $\mathcal{M}$

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

H is the 'hat matrix'.

Recall $H = H^T$

$$HH = H^T H = HH^T = H. \quad (\text{idempotent}).$$

$$y - \hat{y} = y - Hy = (I_n - H)y.$$

Here $\mathcal{M}$ is a finite dimensional space.

What happens if $\mathcal{M}$ is infinite dimensional?

— Hilbert space methods.

— need to define 'distance', 'orthogonality', 'projection'.

Nuisance parameters

Suppose we have the model.

$$Y = X_\beta \beta + X_\theta \theta + \varepsilon$$

θ - parameter of interest
 β - nuisance parameter.

write

$$(Y - X_\theta \theta) = X_\beta \beta + \varepsilon.$$

Then by our

$$\hat{\beta}(\theta) = (X_\beta^T X_\beta)^{-1} X_\beta^T (Y - X_\theta \theta).$$

$$\therefore Y = X_\beta \hat{\beta}(\theta) + X_\theta \theta + \varepsilon.$$

$$= \underbrace{X_\beta (X_\beta^T X_\beta)^{-1} X_\beta^T}_{H_\beta} (Y - X_\theta \theta) + X_\theta \theta + \varepsilon.$$

$$\therefore (I - H_\beta) Y = (I - H_\beta) X_\theta \theta + \varepsilon.$$

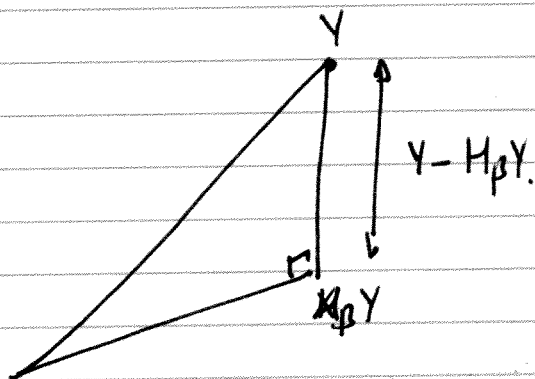
$$\therefore \hat{\theta} = (X_\theta^T (I - H_\beta)^T (I - H_\beta) X_\theta)^{-1} \{X_\theta^T (I - H_\beta)\}^T (I - H_\beta) Y.$$

$$= (X_\theta^T (I - H_\beta) X_\theta)^{-1} X_\theta (I - H_\beta) Y.$$

$$\text{as } (I - H_\beta)^T = (I - H_\beta) \quad (I - H_\beta)^2 = (I - H_\beta).$$

$$\hat{\theta} = (X_0^T (I - H_p) X_0)^{-1} X_0^T (I - H_p) Y$$

What is $(I - H_p)Y = Y - H_p Y = Y - \hat{Y}_p$.



and we have that

$$\begin{aligned} X_p^T (I - H_p) Y &= X_p^T Y - X_p^T H_p Y \\ &= X_p^T Y - X_p^T X_p (X_p^T X_p)^{-1} X_p^T Y \\ &= 0! \end{aligned}$$

$\therefore (I - H_p)Y$ is orthogonal to the space spanned by the columns of X_p .

i.e. $\hat{\theta}$ is a linear combination of the residuals $y_i - \hat{y}_p$.

Random Functions: consider r.v.s $Z \in \mathcal{Z}$
defined on some probability space $(\mathcal{Z}, \mathcal{A}, P)$.

Consider

$$h: \mathcal{Z} \rightarrow \mathbb{R}^2$$

(h is a (measurable) transform, or mapping).

st

$$(i) \quad \mathbb{E}[h(Z)] = \mathbf{0}_2 \quad \text{zero mean}$$

$$(ii) \quad \mathbb{E}[\underbrace{\{h(Z)\}^T h(Z)}_{\text{covariance}}] = \underbrace{\text{scribble}}_{\text{finite variance}} < \infty$$

and $\mathcal{H}$, the space of all such functions.

- linear space (closed under linear combinations).

Random functions $h \in \mathcal{H}$.

Consider $h_1, h_2 \in \mathcal{H}$. (random functions.)

Consider a mapping $\langle h_1, h_2 \rangle$ 'inner product'.

$$\langle, \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{R}$$

st

$$(i) \langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$$

$$(ii) \langle h_1 + h_2, h_3 \rangle = \langle h_1, h_3 \rangle + \langle h_2, h_3 \rangle$$

$$(iii) \langle \lambda h_1, h_2 \rangle = \lambda \langle h_1, h_2 \rangle \quad \lambda \in \mathbb{R}.$$

$$(iv) \langle h_1, h_1 \rangle \geq 0. \text{ with}$$

$$\langle h_1, h_1 \rangle = 0 \iff h_1 \equiv 0_{\mathcal{H}}.$$

For $\mathcal{H}$ the space of random functions in q dimensions, consider

$$\langle h_1, h_2 \rangle = \mathbb{E}[h_1^T h_2] \equiv \mathbb{E}[\{h_1(z)\}^T h_2(z)].$$

Define 'length' $\|h\| = \{\langle h, h \rangle\}^{1/2}$.

'orthogonality' $\langle h_1, h_2 \rangle = 0$.

For Hilbert space $\mathcal{H}$, a linear subspace $\mathcal{V} \subset \mathcal{H}$

is a space where $u_1, u_2 \in \mathcal{V} \Rightarrow au_1 + bu_2 \in \mathcal{V}$
 $a, b \in \mathbb{R}$.

eg $h_1, h_2, \dots, h_k$ arbitrary functions $\in \mathcal{H}$.

Define $\mathcal{V} = \left\{ \sum_{j=1}^k a_j h_j : (a_1, \dots, a_k)^T \in \mathbb{R}^k \right\}$.

- linear subspace spanned by $h_1, \dots, h_k$.

Projection Th^m Let $\mathcal{H}$ be a Hilbert space, and $\mathcal{V}$ be a linear subspace that contains all limit points. Then for each $h \in \mathcal{H}$, $\exists u_0 \in \mathcal{V}$ that is closest to h

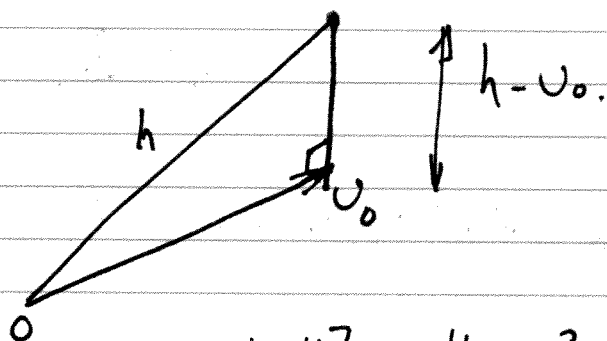
$$\text{i.e. } \|h - u_0\| \leq \|h - u\| \quad \forall u \in \mathcal{V}.$$

and $h - u_0$ is orthogonal to $\mathcal{V}$

$$\text{i.e. } \langle h - u_0, u \rangle = 0 \quad \forall u \in \mathcal{V}.$$

u_0 is the projection of h onto U .

~~It~~ - it is the only element in U for which $h - u$ is orthogonal to u .



$$\|h\|^2 = \|u_0\|^2 + \|h - u_0\|^2.$$

Consider the Hilbert space of mean zero, finite variance q -dimensional random functions with covariance inner product

$$\langle h_1, h_2 \rangle = E[h_1^T h_2]$$

NOTE $\|h - Bv\|^2 = \sum_{i=1}^q E\left[\left(h_i - \sum_{j=1}^p B_{ij} v_j\right)^2\right]$

- componentwise projection.

Let $v \equiv v(z)$ be an r dimensional random function, mean zero, finite ~~variance~~ second moment

$$E[v(z)] = 0 \quad E[v^T v] \leq \infty.$$

Let

$$\mathcal{V} \equiv \{Bv, B \text{ an arbitrary } q \times r \text{ matrix}\}$$

- finite dimensional. (q dimensional)

$$v = \begin{bmatrix} v_1 \\ \vdots \\ v_r \end{bmatrix}$$

For $h \in \mathcal{H}$, the projection of h onto $\mathcal{V}$ is $B_0 v$ st

$$\langle h - B_0 v, Bv \rangle = 0 \quad \forall B \in \mathbb{R}^{q \times r}.$$

$$\text{i.e. } E[(h - B_0 v)^T (Bv)] = 0 \quad \forall B.$$

$$\Leftrightarrow E[(h - \beta_0 v) v^T] = 0^{2 \times r}$$

$$\uparrow E[(h - \beta_0 v)^T (\beta v)]$$

$$= \sum_{i=1}^2 \sum_{j=1}^r \beta_{ij} E[(h - \beta_0 v) v_j]$$

Define the β_{ij} terms as the $2 \times r$ series of matrices with 1 in ij^{th} position.

$$\Leftrightarrow E[h v^T] = \beta_0 E[v v^T]$$

$$\therefore \beta_0 = \{E[\cancel{v v^T}]\}^{-1} E[h v^T] \{E[v v^T]\}^{-1}$$

$$\therefore \beta_0 v = \underline{E[h v^T] \{E[v v^T]\}^{-1} v}$$

Setup: $z_1, \dots, z_n$ iid, density $p_z(z; \theta)$.

$$\theta = \begin{bmatrix} \beta \\ \eta \end{bmatrix} \quad \begin{matrix} g \times 1 \\ r \times 1 \end{matrix} \quad p = g + r.$$

$$\theta_0 = \begin{bmatrix} \beta_0 \\ \eta_0 \end{bmatrix} \text{ - true value.}$$

① Asymptotically linear estimator: $\exists$ influence function $\varphi(z)$ st.

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(z_i) + o_p(1).$$

$$\text{with } E_{\theta_0}[\varphi(z)] = 0_{g \times 1}.$$

$$E_{\theta_0}[\varphi(z)\{\varphi(z)\}^T] \neq \infty, \text{ nonsingular.}$$

φ - 'influence function'.

eg

$$\sqrt{n}(\bar{z}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (z_i - \mu).$$

$$\therefore \varphi(z) = z - \mu.$$

As

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \varphi(z_i) \xrightarrow{d} N(0_{r \times 1}, [E\{\varphi(z)\varphi(z)^T\}])$$

by CLT, we also have that $\sqrt{n}(\hat{\beta}_n - \beta_0)$ has the distribution.

— the influence function is unique.

② Regular estimators Consider the local data generating process (LDGP):

$$Z_n, \dots, Z_n \sim p(z; \theta_n), \quad n=1, 2, \dots$$

$$\text{st } n^{1/2}(\theta_n - \theta^*) \rightarrow \text{constant } n \rightarrow \infty.$$

$\hat{\beta}_n$ is regular if for each $\theta^* \in \Theta$, $\sqrt{n}(\hat{\beta}_n - \beta_n)$ has a limiting distⁿ that does not depend on θ_n .

$$\text{i.e. } \sqrt{n}(\hat{\beta}_n - \beta^*) \xrightarrow{d} N(0, \Sigma^*) \text{ under } p(z; \theta^*)$$

$$\Rightarrow \sqrt{n}(\hat{\beta}_n - \hat{\beta}_n) \xrightarrow{d} N(0, \Sigma^*) \text{ under } p(z; \theta_n).$$

③ Score function

$$S_{\theta}(z, \theta_0) = \frac{\partial}{\partial \theta} \left\{ \log p_z(z, \theta) \right\}_{\theta = \theta_0} \quad p \times 1$$

$$= \begin{bmatrix} S_{\beta}(z, \theta_0) \\ S_{\eta}(z, \theta_0) \end{bmatrix} \quad \begin{matrix} q \times 1 \\ r \times 1 \end{matrix}$$

④ Theorem If $\beta(\theta)$ is a q -dimensional parameter of interest, let

$$\Gamma(\theta) = \frac{\partial \beta(\theta)}{\partial \theta^T} \quad (q \times p).$$

Suppose $\hat{\beta}_n$ is RAL with influence function $\psi(\cdot)$ st $E[\psi(z)^T \psi(z)] < \infty$.

Then

$$E_{\theta_0}[\psi(z) \{S_{\theta}(z, \theta_0)\}^T] = \Gamma(\theta_0).$$

If $\theta = \begin{bmatrix} \beta \\ \eta \end{bmatrix}$, then

$$\Gamma(\theta) = \begin{bmatrix} I_{q \times q} & 0 \\ 0 & 0_{r \times r} \end{bmatrix}$$

Then

$$(i) E_{\theta_0} [\psi(z) \{S_{\beta}(z, \theta_0)\}^T] = \mathbb{I}_{2 \times 2}$$

$$(ii) E_{\theta_0} [\psi(z) \{S_{\eta}(z, \theta_0)\}^T] = \mathbf{0}_{2 \times r}.$$

i.e

$\psi(z)$ is orthogonal to $S_{\eta}(z, \theta_0)$

⑤ M-estimation

In likelihood estimation, the log-likelihood is

$$\ell_n(\theta) = \log \hat{L}_n(\theta) = \sum_{i=1}^n \log p_z(z_i, \theta).$$

$\therefore$ under regularity condition

$$\ell_n(\theta) = \ell_n(\theta_0) + \dot{\ell}_n(\theta')(\theta - \theta_0)$$

by the mean value theorem, where $\|\theta' - \theta_0\| < \|\theta - \theta_0\|$

We (*) evaluate at $\theta = \hat{\theta}_n$ (MLE)

(so that $\dot{\ell}_n(\hat{\theta}_n) = 0_{p \times 1}$).

(ii) multiply through by $\frac{1}{\sqrt{n}}$

$$\underbrace{\left\{ -\frac{1}{n} \ddot{\ell}_n(\theta') \right\}}_{p \times p} \sqrt{n} (\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \dot{\ell}_n(\theta_0).$$

$$\text{As } n \rightarrow \infty, \hat{\theta}_n \xrightarrow{P} \theta_0, \left\{ -\frac{1}{n} \ddot{\ell}_n(\theta') \right\} \xrightarrow{P} E_{\theta_0} [\ddot{\ell}(\theta_0)]$$

$$\therefore \sqrt{n} (\hat{\theta}_n - \theta_0) = \left\{ E_{\theta_0} [\ddot{\ell}(\theta_0)] \right\}^{-1} \frac{1}{\sqrt{n}} \dot{\ell}_n(\theta_0) + o_p(1)$$

$$= \left\{ E_{\theta_0} [\ddot{\ell}(\theta_0)] \right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n S_{\theta}(z_i, \theta_0).$$

$$\therefore \varphi(z) = \left\{ E_{\theta_0} [\ddot{\ell}(\theta_0)] \right\}^{-1} S_{\theta}(z, \theta_0).$$

$$\ddot{\ell}(\theta_0) = \frac{\partial^2}{\partial \theta \partial \theta^T} \left\{ \log p_z(z, \theta) \right\}_{\theta = \theta_0}.$$

In m -estimation, we replace the score equation

$$\sum_{i=1}^n S_0(z_i, \theta) = O_{p \times 1}$$

by the general form

$$\sum_{i=1}^n m(z_i, \theta) = O_{p \times 1}$$

with an arbitrary function m st

$$E_\theta[m(z, \theta)] = O_{p \times 1}$$

$$E_\theta[m(z, \theta)^T m(z, \theta)] < \infty$$

$$E_\theta[m(z, \theta) m(z, \theta)^T] \text{ is nonsingular}$$

for all θ .

Define $\hat{\theta}_n$ by

$$\sum_{i=1}^n m(z_i, \hat{\theta}_n) = O_{p \times 1}$$

Using exactly the same logic, we have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(z_i) + o_p(1)$$

with

$$\psi(z_i) \equiv \left\{ -E_{\theta_0}[\dot{m}(z, \theta_0)] \right\}^{-1} m(z, \theta_0)$$

$\therefore$ By the CLT we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \text{Normal}_p(0_{p \times 1}, B^{-1} A B^{-1})$$

$$B \equiv B(\theta_0) = E_{\theta_0}[\dot{m}(z, \theta_0)]$$

$$A \equiv A(\theta_0) = E_{\theta_0}[m(z, \theta_0) \{m(z, \theta_0)\}^T]$$

-estimated by

$$\hat{B}_n = \frac{1}{n} \sum_{i=1}^n \dot{m}(z_i, \hat{\theta}_n) \quad \hat{A}_n = \frac{1}{n} \sum_{i=1}^n m(z_i, \hat{\theta}_n) \{m(z_i, \hat{\theta}_n)\}^T$$

⑥ The Geometry of Influence functions

$\mathcal{H}$: Hilbert space of zero mean $g \times 1$ fns
 $\langle h_1, h_2 \rangle = E[h_1^T h_2]$

$\mathcal{T}$: p -dimensional linear subspace of $\mathcal{H}$

$$\mathcal{T} = \{ B S_0(z, \theta_0) : B \text{ is } g \times p \}$$

- 'tangent space'.

Λ : linear subspace of $\mathcal{H}$

$$\Lambda = \{ B S_\eta(z, \theta_0) : B \text{ is } g \times r \}$$

- 'nuisance tangent space'.

By the above theorem (ii), for RAC estimator $\hat{\beta}_n$

$$E_{\theta_0} [\Psi_{\hat{\beta}_n}(z) \{S_\eta(z, \theta_0)\}^T] = O_{g \times r}.$$

i.e. $\Psi_{\hat{\beta}_n}(z)$ is orthogonal to Λ .

$$\Psi_{\hat{\beta}_n}(z) \in \Lambda^\perp.$$

We have that

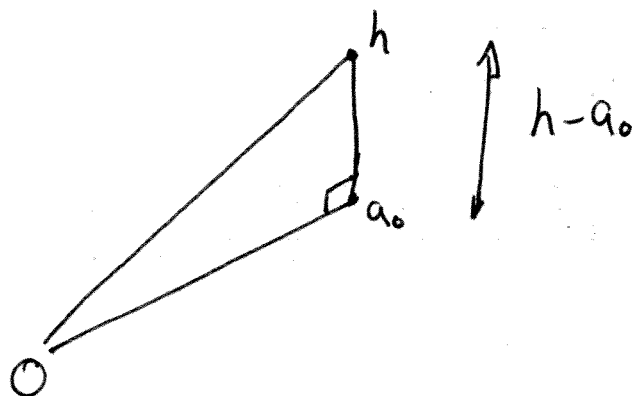
$$\mathcal{H} = \Lambda \oplus \Lambda^\perp$$

i.e. $h \in \mathcal{H}$ is the sum of elements

- one in Λ

- one in $\Lambda^\perp$

For $h \in \mathcal{H}$, denote by $\Pi(h/\Lambda)$ the projection of h onto Λ , and by $\Pi(h/\Lambda^\perp)$ the 'residual'



$a_0 \in \Lambda$ is the 'closest' point in Λ to h .

$$\Pi(h/\Lambda) = a_0$$

$$\Pi(h/\Lambda^\perp) = h - a_0.$$

$$\tau = \tau_{\beta} \oplus \Lambda$$

$$\tau_{\beta} \equiv \{BS_{\beta}(z, \alpha_0) : B \text{ } \varepsilon \times \varepsilon\}$$

By the earlier result, we know

$$\pi(h|\Lambda) = E[hS_{\eta}^T] \{E[S_{\eta}S_{\eta}^T]\}^{-1} S_{\eta}$$

$\therefore$ All elements of $\Lambda^{\perp}$ take the form

$$h - \pi(h|\Lambda)$$

$$= h - E[hS_{\eta}^T] \{E[S_{\eta}S_{\eta}^T]\}^{-1} S_{\eta}$$

⑦ Constructing the Estimator $\psi(z)$ satisfies the Theorem

Set

$$m(z, \beta, \gamma) = \psi(z) - E_{\beta, \gamma}[\psi(z)]$$

(mean zero, finite variance).

$\hat{\beta}_n$ satisfying

$$\sum_{i=1}^n m(z_i, \hat{\beta}_n, \hat{\gamma}_n(\hat{\beta}_n)) = O_{p \times 1}.$$

is RAL with influence function $\psi(z)$.

⑧ Efficient Influence Function.

$\psi_{\text{eff}}(z)$ is the influence function with the 'smallest' variance matrix. We have

$$\begin{aligned}\psi_{\text{eff}}(z) &= \psi^*(z) - \Pi(\psi^*(z) | \gamma^\perp) \\ &\equiv \Pi(\psi^*(z) | \gamma)\end{aligned}$$

for an arbitrary influence function $\psi^*(z)$

More precisely

$$\eta_{\text{eff}}(z) = \Gamma(\theta_0) I^{-1}(\theta_0) S_\theta(z, \theta_0)$$

where

$$I(\theta_0) = E_{\theta_0} [S_\theta(z, \theta_0) \{S_\theta(z, \theta_0)\}^T]$$

The efficient score function for β is

$$S_{\text{eff}}(z, \theta_0) = S_\beta(z, \theta_0) - \Pi(S_\beta(z, \theta_0) | \Lambda)$$

— obtained by projecting $S_\beta(z, \theta_0)$ onto the nuisance target space Λ , and taking the residual. Recall

$$\Pi(S_\beta(z, \theta_0) | \Lambda)$$

$$= E[S_\beta S_\eta^T] \{E[S_\eta S_\eta^T]\}^{-1} S_\eta(z, \theta_0)$$

The efficient influence function for β is

$$\psi_{\text{eff}}(z, \theta_0) = \left\{ E_{\theta_0} \left[S_{\text{eff}} S_{\text{eff}}^T \right] \right\}^{-1} S_{\text{eff}}(z, \theta_0).$$

with variance

$$\left\{ E_{\theta_0} \left[S_{\text{eff}} S_{\text{eff}}^T \right] \right\}^{-1}$$

If $I = \begin{bmatrix} I_{\beta\beta} & I_{\beta\eta} \\ I_{\eta\beta} & I_{\eta\eta} \end{bmatrix}$

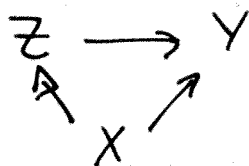
the minimum variance is

$$\left\{ I_{\beta\beta} - I_{\beta\eta} I_{\eta\eta}^{-1} I_{\eta\beta} \right\}^{-1}.$$

⑨ Semiparametric Case

4 CAUSAL ADJUSTMENT METHODS

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



- modelling $Y|X, Z$ correctly yields

$$E_{Y|Z}^E[Y|Z=z] = E_X^E[E_{Y|X,Z}^E[Y|X, Z=z]]$$

How else can we correct the inference to make it robust to misspecification?

Recall the 'G-computation' formula:

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

$$\neq E_{Y|Z}^0[Y|Z=z]$$

In general, as $X \not\perp Z$.

How can we break the dependence?

- if $f_X^E(x)$ is degenerate at $x=x_0$.

say, then

$$f_{Z|X}^0(z|x) \equiv f_Z^0(z) \quad \forall z, x=x_0.$$

- there is no variation in x

$$E_{Y|Z}^0[Y|Z=z] = \int y f_{Y|X,Z}^0(y|x_0,z) dy.$$

$$\equiv E_{Y|Z}^E[Y|Z=z].$$

- suppose that $f_{z|x}^0(z|x)$ does not depend on x
for $x \in \mathcal{X}_1 \subset \mathcal{X}$

Then we can write the 'local' causal effect

$$\mathbb{E}_{y|z}^{\varepsilon, \mathcal{X}_1}[Y|Z=z] = \int \int_{\mathcal{X}_1} y f_{y|x,z}^{\varepsilon}(y|x,z) f_x^{\varepsilon}(x) dy dx$$

$$= \int \int \int_{\mathcal{X}_1} \mathbb{I}_z(z') y f_{y|x,z}^{\varepsilon}(y|x,z') f_z^{\varepsilon}(z') f_x^{\varepsilon}(x) dy dz' dx$$

$$= \int \int \int_{\mathcal{X}_1} \mathbb{I}_z(z') y f_{y|x,z}^{\varepsilon}(y|x,z') f_z^{\varepsilon}(z') f_x^{\varepsilon}(x) dy dz' dx$$

$$= \int \int_{\mathcal{X}_1} f_z^0(z') \int y f_{y|x,z}^0(y|x,z') f_x^0(x) dy dz' dx$$

$$= \mathbb{E}_{y|z}^{0, \mathcal{X}_1}[Y|Z=z] = \mu_{\mathcal{X}_1}^Y(z) \text{ say.}$$

This inspires the 'local' estimator.

$$\hat{\nu}^{x_i}(z) = \frac{\sum_{i=1}^n \mathbb{I}_{x_i}(x_i) \mathbb{I}_z(z) x_i}{\sum_{i=1}^n \mathbb{I}_{x_i}(x_i) \mathbb{I}_z(z)}.$$

• suppose that $\mathcal{X} \equiv \bigcup_{j=1}^J \mathcal{X}_j$ $\mathcal{X}_j \cap \mathcal{X}_k = \emptyset$.

and that

$$f_{z|x}^0(z|x)$$

does not depend on x for $x \in \mathcal{X}_j$

i.e.

$$f_{z|x}^0(z|x) \equiv f_j(z) \quad x \in \mathcal{X}_j.$$

Then we can write by the above argument

$$E_{\pi/z}^{\varepsilon}[Y|Z=z] = \sum_{j=1}^J \nu^{x_j}(z) P[X \in \mathcal{X}_j].$$

This construction encourages us to consider matching

- we can directly compare Y values for different z values if we hold X (almost) constant

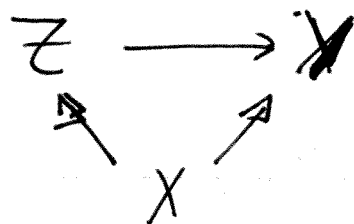
However if X is continuous, we must define the 'strata' $\mathcal{X}_1, \dots, \mathcal{X}_J$, or carefully choose the 'matched' subjects.

- difficult to study statistically.
- very popular method.

(method of last resort!)

Constructing Balance

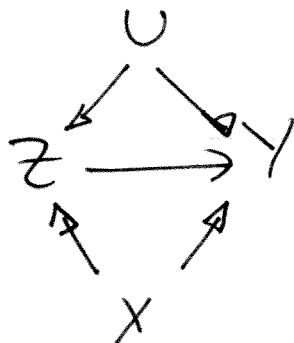
The lack of independence between X and Z is the principal obstacle to inferring the causal effect.



Conditioning on X blocks the path.

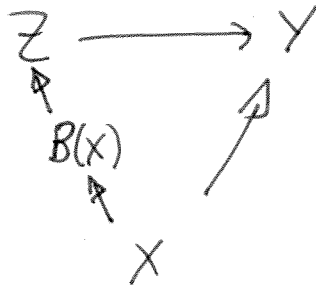
→ correct inference (under correct spec).

NOTE 'Unmeasured confounding'



If U is not included in the model, there is an open, confounding path.

Consider



$B(x)$ - some function of X st

$$Z \perp X \mid B(x).$$

$\therefore$ By conditioning on $B(x)$, we block the backdoor path from Z to Y .

but $B(x)$ does not have an arrow into Y .

$\therefore$ within 'strata' of $B(x)$, we can examine the difference between ~~between~~

Y values for different Z values
and know that there is no influence of X .

Can we find such a $B(x)$? i.e. st

$$f_{z|x, B(x)}(z|x, B(x)) \equiv f_{z|B(x)}(z|B(x)) \quad \forall z, x.$$

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

Suppose first $z \in \{0, 1\}$.

$$P_{z|x}[Z=1|X=x] = 1 - P_{z|x}[Z=0|X=x]$$

$\therefore$ Must have a Bernoulli-like model.

$$f_{z|x}(z|x) = \{p(x)\}^z \{1-p(x)\}^{1-z} \quad z \in \{0, 1\}$$

Now

$$\cancel{f_{z|x, B(x)}} f_{z|x, B(x)}(z|x, B(x))$$

Must also be Bernoulli-like.

$$\{g(x, b(x))\}^z \{1-g(x, b(x))\}^{1-z}.$$

We can make μ independent of x ^{given $B(x)=b$} iff

$$q(x, B(x)) = q(B(x)).$$

i.e. this probability is constant on the 'contour'

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

in which case $q(B(x)) = q(b) \quad x \in \mathcal{B}_b.$

Then

$$\begin{aligned} f_{Z|X, B(x)}(z | x, B(x)) &\equiv f_{Z|B(x)}(z | B(x)) \\ &= \{q(b)\}^z \{1 - q(b)\}^{1-z} \\ &\quad z \in \{0, 1\}. \end{aligned}$$

$$q(b) = P[Z=1 | B=b]$$

Now if we define

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

We then have

$$f_{z|x,b}(z|x,b) \equiv b^z(1-b)^{1-z} \quad z \in \{0,1\}$$

In this case, we have the definition of $B(x)$ as

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

and

$$Z \perp X | B(x).$$

$B(x)$ is the propensity score

-denoted by $e(x)$ by
Rosenbaum and Rubin (1983).

Write $B = B(x)$, $b = B(x)$.

We have

$$\begin{aligned}
 f_{B,x,z}(b,x,z) &= f_x(x) f_{B|x}(b|x) f_{z|B,x}(z|b,x) \\
 &= \begin{cases} f_x(x) f_{z|B,x}(z|b,x) & b = B(x) \\ 0 & \text{otherwise} \end{cases}
 \end{aligned}$$

$$f_{B,z}(b,z) = \int_{\mathcal{X}_b} f_{z|B,x}(z|b,x) f_x(x) dx \quad \forall (b,z) \in \mathcal{B} \times \mathcal{Z}$$

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

$$f_{x|z,B}(x|z,b) = \frac{f_{z|B,x}(z|b,x) f_{x|B}(x|b)}{\int f_{z|B,x}(z|b,x') f_{x|B}(x'|b) dx'}$$

$$(b,z) \in \mathcal{Z} \times \mathcal{B}$$

$$= \frac{f_{z|B,X}(z|b,x) f_{X|B}(x|b)}{\int_{\mathcal{X}_b} f_{z|B,X}(z|b,x') f_{X|B}(x'|b) dx'}$$

If $f_{z|B,X}(z|b,x) \equiv f_{z|B}(z|b) \quad \forall x \in \mathcal{X}_b$.

then the LHS $\Rightarrow X \perp Z | B$.

For the binary case $Z \in \{0,1\}$.

$$B \equiv B(X) = f_{z|x}(1|x) \equiv \mathbb{E}_{Z|X}[Z|X]$$

$B(X)$ is a 'balancing score'.

- conditional on a fixed value of $B(X)$, b , say.
 Z and X are independent.

Suppose $B^* \equiv B^*(x)$ is another balancing score.

THEOREM

B^* is a balancing score if and only if

$$B = g(B^*)$$

i.e. B is a function of B^*

i.e. B is 'coarser' than B^*

B^* is 'finer' than B .

(i.e. you can deduce the value of B if given B^*).

Proof Suppose B^* is finer than B : $B^* = b^* \Rightarrow B = b$.

We have

$$\begin{aligned} P[Z=1 | B^* = b^*] &= E_{Z|B^*} [Z | B^* = b^*] \\ &= E_{B|B^*} \left[E_{Z|B} [Z | B^* = b^*] \mid B^* = b^* \right] \end{aligned}$$

$$= E_{B|B^*} [E_{Z|B} [Z | B=b] \mid B^* = b^*]$$

$$= E_{Z|B} [Z | B=b] = b = P[Z=1 | B=b]$$

$\Rightarrow B^*$ is a balancing score.

Conversely, suppose B^* is a balancing score
but $\exists x_1, x_2$ st

$$B^*(x_1) = B^*(x_2) \quad \underline{\text{but}} \quad B(x_1) \neq B(x_2) \\ = b^*$$

$$\text{But } B(x_1) \neq B(x_2) \Rightarrow P[Z=1/X=x_1] \neq P[Z=1/X=x_2]$$

$\Rightarrow X$ and Z not independent given $B^*(X)$.

$\Rightarrow \underline{B^*}$ is not a balancing score — contradiction.

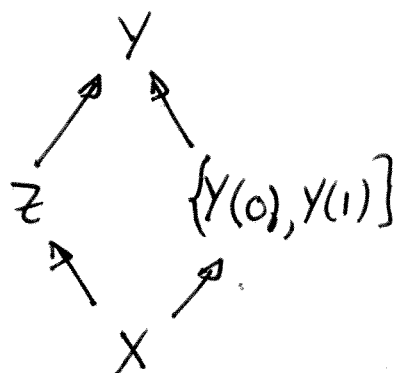
This theorem says that $B(X)$ is the
~~best~~ balancing score capable of the
greatest 'dimension reduction'

— akin to unimod sufficiency.

Ignorability

Treatment assignment is strongly ignorable if

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

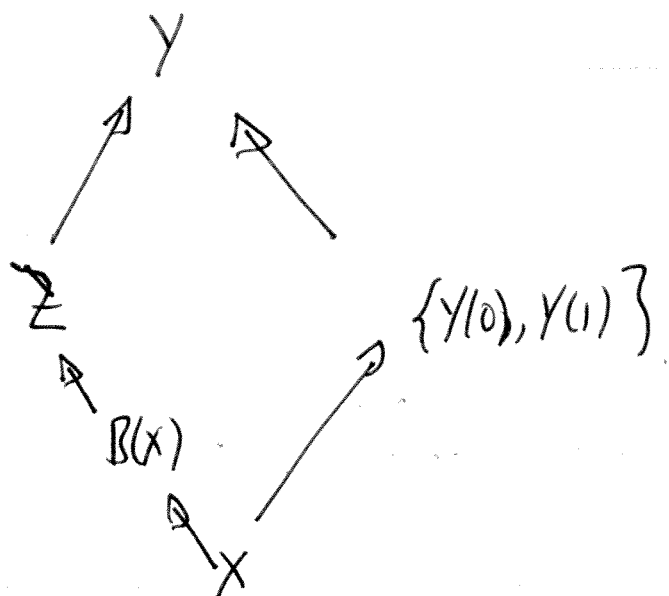


$$Y = ZY(1) + (1-Z)Y(0).$$

$$Y(z) = g(X, z) \quad (\text{or } g(X, z, U) \text{ possibly})$$

$$\text{eg } Y(z) = \beta_0 + \beta_1 z + \beta_2 X + U. \quad z \text{ nonrandom}$$

$$z = 0, 1.$$



It follows that

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

i.e. ignorable given $X \Rightarrow$ ignorable given $B(x)$.

BEYOND THE BINARY CASE

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

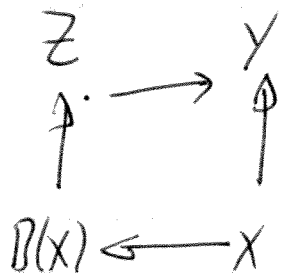
$$f_{Z|B}(z|b) = \int_{\mathcal{X}_{b,z}} f_{Z|B,X}(z|b, x) f_{X|B}(x|b) dx$$

$$\mathcal{X}_{b,z} = \{x : f_{Z|X}(z|x) = b\}$$

But clearly $f_{Z|B,X}(z|b, x) \equiv f_{Z|X}(z|x) = b$
on $\mathcal{X}_{b,z}$

$$\begin{aligned} f_{Z|B}(z|b) &= b \int_{\mathcal{X}_{b,z}} f_{X|B}(x|b) dx \\ &= b. \end{aligned}$$

which does not depend on x .



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

~~we~~

We observe data

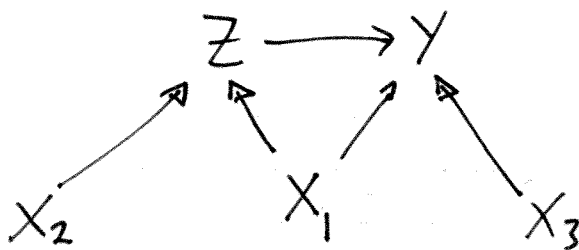
$\{(x_i, y_i, z_i) : i=1, \dots, n\}$ from

$$f_{X,Y,Z}^0(x, y, z).$$

Adjustment via the propensity score : Binary case.

$e(x) = P[Z=1|X]$ can be used to construct causal adjustment procedures by blocking biasing/backdoor/confounding paths.

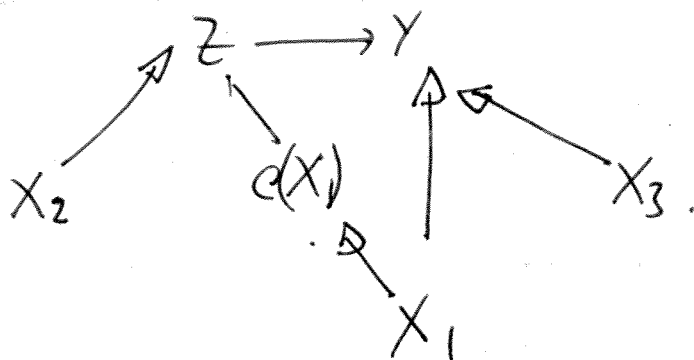
- Note that if there are more variables in the graph, this can alter the nature of the construction of $e(x)$



X_1 : influences both Z and Y — confounders

X_2 : influences only Z — 'instruments'

X_3 : influences only Y — 'pure predictors of outcome'.



$\therefore e(x) \equiv e(x_1)$ i.e. only confounders
need enter the calculation of $e(x)$

$$e(x) \equiv e(x_1) = P[Z=1/X_1]$$

i.e. it is evident from the graph that

$$\int_{z|x_1, x_2}^0 (z|x_1, x_2) \equiv \int_{z|e(x), x_2}^0 (z|e, x_2)$$

BUT we may still define

$$e(x) \equiv e(x_1) \equiv \int_{z|x_1}^0 (z|x_1)$$

Stratification

Note that $e(x)$ is scalar irrespective of the dimension of X . We have

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

$$\mu(z_1) - \mu(z_0)$$

$$\begin{aligned}&= \mathbb{E}_X^{\mathcal{E}} \left[\mathbb{E}_{Y|X,Z}^{\mathcal{E}}[Y|X,Z=z_1] - \mathbb{E}_{Y|X,Z}^{\mathcal{E}}[Y|X,Z=z_0] \right] \\ &= \mathbb{E}_X^{\mathcal{E}} \left[\mathbb{E}_{Y|e(X),Z}^{\mathcal{E}}[Y|e(X),Z=z_1] - \mathbb{E}_{Y|e(X),Z}^{\mathcal{E}}[Y|e(X),Z=z_0] \right] \\ &= \mathbb{E}_X^0 \left[\underbrace{\mathbb{E}_{Y|e(X),Z}^0[Y|e(X),Z=z_1] - \mathbb{E}_{Y|e(X),Z}^0[Y|e(X),Z=z_0]}_{\text{Y with the same } e(X) \text{ values but different } Z \text{ values are directly comparable.}} \right]\end{aligned}$$

Y with the same $e(x)$ values but different z values are directly comparable.

$\therefore$ For ~~estimation~~ estimation, let

$$\mathcal{X}_1 \equiv \{x : e(x) = e_1\} \quad \text{fixed } e_1.$$

$$\mu^{x_1}(z) = E_{y/z}^{e, x_1}[Y | Z=z]$$

can be estimated by

$$\hat{\mu}^{x_1}(z) = \frac{\sum_{i=1}^n \mathbb{I}_{\mathcal{X}_1}(x_i) \mathbb{I}_{\{z\}}(z_i) y_i}{\sum_{i=1}^n \mathbb{I}_{\mathcal{X}_1}(x_i) \mathbb{I}_{\{z\}}(z_i)}$$

$$\hat{\mu}(z) = \sum_{j=1}^J \hat{\mu}^{x_j}(z) P[X \in \mathcal{X}_j].$$

Matching: can construct a set of paired values matched on $e(X)$.

Regression: 'conditioning on $e(X)$ ' can be achieved by regression.

eg $E_{Y|X,Z}^0[Y|X,Z]$ replaced by

$$E_{Y|e(X),Z}^0[Y|e(X),Z]$$

eg

$$E_{Y|e(X),Z}^0[Y|e(X),Z] = \beta_0 + \beta_1 Z + \beta_2 e(X)$$

Then $E_{Y|Z}^0[Y|Z=z]$ can be computed

$$E_{e(X)}^0 \left[E_{Y|e(X),Z}^0[Y|e(X),Z=z] \right]$$

ad estimated by

$$\frac{1}{n} \sum_{i=1}^n \{ \hat{\beta}_0 + \hat{\beta}_1 z + \hat{\beta}_2 e(x_i) \}$$

where $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ are estimated via OLS
using the regression model.

'Propensity score regression'.

Inverse Weighting From an earlier result on Importance Sampling.

$$E_{Y|Z}^E[Y|Z=z] = E_{X,Y,Z}^0 \left[\frac{\mathbb{I}_{Z=1}(z)}{f_{Z|X}^0(z|X)} Y \right] / E_{X,Y,Z}^0 \left[\frac{\mathbb{I}_{Z=1}(z)}{f_{Z|X}^0(z|X)} \right]$$

$$E_{Y|Z}^E[Y|Z=z] = \frac{\iiint \mathbb{I}_{Z=1}(z') y f_{X,Y,Z}^E(x,y,z) dx dy dz}{\iiint \mathbb{I}_{Z=1}(z') f_{X,Y,Z}^E(x,y,z) dx dy dz}$$

Write $f_{x,y,z}^{\mathcal{E}}(x,y,z) = \frac{f_{x,y,z}^{\mathcal{E}}(x,y,z)}{f_{x,y,z}^{\mathcal{O}}(x,y,z)} f_{x,y,z}^{\mathcal{O}}(x,y,z)$.

$$\Rightarrow \underbrace{\frac{f_x^{\mathcal{E}}(x)}{f_x^{\mathcal{O}}(x)}}_{=1} \cdot \frac{f_z^{\mathcal{E}}(z)}{f_{z|x}^{\mathcal{O}}(z|x)} \cdot \frac{f_{y|x,z}^{\mathcal{E}}(y|x,z)}{\underbrace{f_{y|x,z}^{\mathcal{O}}(y|x,z)}_{=1}}$$

$\therefore$ The 'Inverse Weights' estimator is

$$\frac{\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}_{\{z_i\}}(z_i) y_i}{f_{z|x}^{\mathcal{O}}(z_i|x_i)}}{\frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}_{\{z_i\}}(z_i)}{f_{z|x}^{\mathcal{O}}(z_i|x_i)}}$$

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

say.

$$W_i = \frac{\mathbb{I}_{\{z_i\}}(z_i)}{f_{z|x}^0(z_i | x_i)}$$

$$= \sum_{i=1}^n W_i^* Y_i$$

z_i binary

$$\hat{\mu}(1) = \frac{\sum_{i=1}^n \frac{z_i Y_i}{e(x_i)}}{\sum_{i=1}^n \frac{z_i}{e(x_i)}}$$

$$\hat{\mu}(0) = \frac{\sum_{i=1}^n \frac{(1-z_i) Y_i}{1-e(x_i)}}{\sum_{i=1}^n \frac{(1-z_i)}{1-e(x_i)}}$$

NOTE

$$\begin{aligned} E[W_i] &= E_{x_i, z_i} \left[\frac{z_i}{e(x_i)} \right] = E_{x_i} \left[E_{z_i | x_i} \left[\frac{z_i}{e(x_i)} \right] x_i \right] \\ &= E_{x_i} [1] = 1. \end{aligned}$$

Could use the simpler estimator

$$\hat{\mu}(z) = \frac{1}{n} \sum_{i=1}^n \frac{\Pi_i(z) Y_i}{f_{Z|X}^0(z|X_i)}$$

as we have

$$\hat{\mu}(z) = \sum_{i=1}^n W_i^* Y_i$$

$$W_i^* = \frac{W_i}{\sum_{j=1}^n W_j}$$

$$\text{with } E_{X,Z}^0[W_i^*] = \frac{1}{n}$$

Note

$$E_{X,Y,Z}^0[\hat{\mu}(z)] = E_{X,Y,Z}^0\left[\frac{\Pi_z(z) Y_i}{f_{Z|X}^0(z|X)}\right]$$

Write

$$\begin{aligned}\nu(x, z) &= E_{Y|X, Z}^0[Y|X=x, Z=z] \\ &= E_{Y|X, Z}^{\varepsilon}[Y|X=x, Z=z].\end{aligned}$$

Then

$$E_{X, Y, Z}^0\left[\frac{\pi_{12}(z)}{f_{X|Z}^0(z|X)} Y\right] = E_{X, Z}^0\left[\frac{\pi_{12}(z)}{f_{X|Z}^0(z|X)} \nu(X, z)\right]$$

For $z \in \{0, 1\}$:

$$= E_{X, Z}^0\left[\frac{z}{e(x)} \nu(X, z)\right] \quad z=1.$$

$$= E_X^0\left[\frac{1}{e(x)} \nu(X, 1) P_{\mathcal{Z}}^0[\bar{z}=1|X]\right]$$

$$\equiv E_X^0[\nu(X, 1)] \equiv E_X^{\varepsilon}[\nu(X, 1)] = \mu(1)$$

$\Rightarrow \tilde{\mu}(1)$ is an unbiased estimator of $\mu(1)$

- similar result for $\hat{\mu}(1)$

- similar for $\tilde{\mu}(0), \hat{\mu}(0)$.

- write $\tilde{\mu}_{IPW}(z), \hat{\mu}_{IPW}(z)$.

NOTE Estimating equation

$$E_{x,y,t}^0 \left[\frac{zY}{e(x)} \right] = \mu(1)$$

$$\text{i.e. } E_{x,y,t}^0 \left[\frac{zY}{e(x)} - \mu(1) \right] = 0.$$

$$\rightarrow \sum_{i=1}^n \left[\frac{z_i Y_i}{e(x_i)} - \mu(1) \right] = 0.$$

$$E_{x,y,t}^0 \left[\left\{ \frac{zY}{e(x)} - \frac{(1-z)Y}{1-e(x)} \right\} - \{ \mu(1) - \mu(0) \} \right] = 0.$$

$$\therefore \hat{\mu}(1) - \hat{\mu}(0) = \frac{1}{n} \sum_{i=1}^n \frac{z_i Y_i}{e(x_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1-z_i) Y_i}{1-e(x_i)} = 0.$$

Is this the most efficient estimator?

Augmentation

Note that

$$\begin{aligned} E_{Y|Z}^E[Y|Z=z] &= E_X^E \left[E_{Y|X,Z}^E[Y|X,z=z] \right] \\ &= E_X^E \left[E_{Y|X,Z}^E[Y - \mu(X,z) + \mu(X,z) | X, z=z] \right] \end{aligned}$$

$$\text{where } \mu(X,z) = E_{Y|X,Z}^E[Y|X,z]$$

$$\begin{aligned} &= E_X^E \left[E_{Y|X,Z}^E[(Y - \mu(X,z)) | X, z=z] \right] \\ &\quad + E_X^E [\mu(X,z)] \end{aligned}$$

$$= E_X^0 \left[\frac{I_{(1)}(z)(Y - \mu(X,z))}{e(X) (1 - e(X))^{1-z}} \right]$$

$$\equiv E_{X,Y,Z}^0 \left[\frac{\mathbb{I}_{\{Z\}}(Z)}{f_{Z|X}^0(Z|X)} (Y - \mu(X, Z)) \right] \\ + E_X^0 [\mu(X, Z)].$$

$\therefore$ we have estimators

$$\hat{\mu}_{AIPW}(Z) = \frac{\sum_{i=1}^n \frac{\mathbb{I}_{\{Z\}}(Z_i)}{f_{Z|X}^0(Z_i|X_i)} (Y_i - \mu(X_i, Z)) + \frac{1}{n} \sum_{i=1}^n \mu(X_i, Z)}{\sum_{j=1}^n \frac{\mathbb{I}_{\{Z\}}(Z_j)}{f_{Z|X}^0(Z_j|X_j)}} \\ = \sum_{i=1}^n W_i^* (Y_i - \mu(X_i, Z)) + \frac{1}{n} \sum_{i=1}^n \mu(X_i, Z)$$

$$\tilde{\mu}_{AIPW}(Z) = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{I}_{\{Z\}}(Z_i)}{f_{Z|X}^0(Z_i|X_i)} (Y_i - \mu(X_i, Z)) + \frac{1}{n} \sum_{i=1}^n \mu(X_i, Z) \\ = \frac{1}{n} \sum_{i=1}^n W_i (Y_i - \mu(X_i, Z)) + \frac{1}{n} \sum_{i=1}^n \mu(X_i, Z).$$

AIPW - Augmented PW

- model $\mu(x, z)$ to be specified
 $\mu(x, 1), \mu(x, 0)$.

- unbiased for $\mu(z)$.

Estimating eqⁿ $\sigma \tilde{\mu}_{AIPW}(z)$

$$\sum_{i=1}^n \left[W_i (y_i - \mu(x_i, z)) + \mu(x_i, z) - \mu(z) \right] = 0.$$

- we have that

$$\text{Var}_{x,y,z}^0 \left[\tilde{\mu}_{AIPW}(z) \right] \leq \text{Var}_{x,y,z}^0 \left[\tilde{\mu}_{IPW}(z) \right]$$

with equality iff $\mu(x, z) = 0$.

- exercise.

Write

$$\begin{aligned}\tilde{\mu}_{AIPW}(z) &= \frac{1}{n} \sum_{i=1}^n W_i (Y_i - \mu(X_i, z)) + \frac{1}{n} \sum_{i=1}^n \mu(X_i, z) \\ &= \frac{1}{n} \sum_{i=1}^n \{W_i Y_i + (1-W_i) \mu(X_i, z)\}\end{aligned}$$

$$\text{Var}_{X,Y,z}^{\circ} \left[\tilde{\mu}_{AIPW}(z) \right] = \frac{1}{n} \text{Var} [WY + (1-W) \mu(X, z)]$$

$$\begin{aligned}\text{Var} [WY + (1-W) \mu(X, z)] &= \text{Var} [WY] + \text{Var} [(1-W) \mu(X, z)] \\ &\quad + 2 \text{Cov} [WY, (1-W) \mu(X, z)]\end{aligned}$$

$$\begin{aligned}\text{Var} [(1-W) \mu(X, z)] &= \mathbb{E} \left[\{(1-W) \mu(X, z)\}^2 \right] - \underbrace{\left\{ \mathbb{E} [(1-W) \mu(X, z)] \right\}^2}_{=0} \\ &\quad \text{as } \mathbb{E} [(1-W) \mu(X, z)] = 0\end{aligned}$$

$$\begin{aligned}\text{Cov} [WY, (1-W) \mu(X, z)] &= \mathbb{E} [WY (1-W) \mu(X, z)] \\ &\quad - \underbrace{\mathbb{E} [WY] \mathbb{E} [(1-W) \mu(X, z)]}_{=0}\end{aligned}$$

$$\begin{aligned}
& \therefore \text{Var}[WY + (1-W)\mu(x, z)] \\
&= \text{Var}[WY] + \mathbb{E}[(1-W)^2 \{\mu(x, z)\}^2] + 2 \mathbb{E}[W(1-W) \{\mu(x, z)\}^2] \\
&= \text{Var}[WY] + \mathbb{E}[\{(1-W)^2 + 2W(1-W)\} \{\mu(x, z)\}^2] \\
&= \text{Var}[WY] + \mathbb{E}[(1-W^2) \{\mu(x, z)\}^2] \\
&\leq \text{Var}[WY]
\end{aligned}$$

$$\begin{aligned}
& \text{as } \mathbb{E}_{z|X}[(1-W^2) \{\mu(x, z)\}^2 | X] \\
&= \{\mu(x, z)\}^2 \mathbb{E}_{z|X}[1 - W^2 | X] \\
&= \{\mu(x, z)\}^2 \mathbb{E}_{z|X}\left[\left[1 - \left\{\frac{z_i}{e(x)}\right\}^2\right] | X\right] \\
&= \{\mu(x, z)\}^2 \left(1 - \cancel{\mathbb{E}_{z|X}}\left(\frac{1}{e(x)}\right)\right) \\
&= \{\mu(x, z)\}^2 \left(\frac{e(x) - 1}{e(x)}\right) \leq 0.
\end{aligned}$$

NOTE The APU estimator can be constructed by projecting the IPW influence function onto the appropriate nuisance target space, and taking the residual.

$$\cdot \Lambda = \{ (z - e(x)) h(x) : h(x) \text{ arbitrary} \}$$

— logistic regression.

• 'Antithetic variable' trick.

$$E[T(Y)] = 0$$

Suppose $g(Y)$ w st $E[g(Y)] = 0$.

$$\therefore E[T(Y) + g(Y)] = 0 \quad \text{still unbiased!}$$

$$\begin{aligned} \text{Var}[T(Y) + g(Y)] &= \text{Var}[T(Y)] + \text{Var}[g(Y)] + 2\text{Cov}[T(Y), g(Y)] \\ &\leq \text{Var}[T(Y)] \end{aligned}$$

$$\text{if } \text{Var}[g(Y)] + 2\text{Cov}[\pi(Y), g(Y)] \leq 0.$$

$$\text{i.e. } \frac{\text{Var}[g(Y)]}{-2\text{Cov}[\pi(Y), g(Y)]} \leq 1.$$

Here for $\tilde{\mu}(1)$

$$\psi^{\text{eff}}(X, Y, Z) = \frac{Z}{e(X)} - \underbrace{\pi\left[\frac{ZY}{e(X)} \mid 1\right]}_{\text{}} - \mu(1).$$

$$= \frac{(Z - e(X))E[Y|X, Z=1]}{e(X)}.$$

For both $\hat{\mu}_{IPW}(z)$ and $\hat{\mu}_{OLS}(z)$

$$E_{x,y,z}^0 [W_i (Y_i - \mu(x_i, z))] = 0$$

if $\mu(x, z)$ is the true mean model.

But we know that the optimal estimator in these cases is

$$\hat{\mu}(z) = \frac{1}{n} \sum_{i=1}^n \mu(x_i, z)$$

(see regression results)

$\therefore$ Why would we use IPW at all if $\mu(x, z)$ is correctly specified?

Robustness to Misspecification

Suppose, in reality, the conditional mean model is $\mu(x, z)$, but we mistakenly use $m(x, z)$.

eg use estimator

$$\frac{1}{n} \sum_{i=1}^n W_i (Y_i - m(x_i, z)) + \frac{1}{n} \sum_{i=1}^n m(x_i, z)$$

$$E_{x, y, z}^0 [W_i (Y_i - m(x_i, z))]$$

$$= E_{x, z}^0 [W_i (\mu(x_i, z) - m(x_i, z))]$$

provided $E_{x, z}^0 [W_i (\mu(x_i, z) - m(x_i, z))]$

$$= E_{x,z}^0 \left[\frac{z_i}{e(x_i)} (\nu(x_i, z) - m(x_i, z)) \right]$$

$$= E_x^0 [\nu(x_i, 1) - m(x_i, 1)]$$

∴

$$E_{x,y,z}^0 [W_i(Y_i - m(x_i, z)) + m(x_i, z)]$$

$$= E_x^0 [\nu(x_i, 1)] \equiv E_{y|z}^E [Y | Z=1]$$

i.e. provided $e(x_i)$ is correctly specified as $f_{z|x}^0(1|x_i)$ we still have a unbiased estimator.

- similar result for $\hat{\nu}_{AIDW}(z)$.

Misspecifying $e(x_i)$.

Suppose instead of $e(x_i)$, we use another function $g(x_i)$ as an incorrect model for $P[z=1/x_i]$

$$\text{Then } E_{x,z}^0 \left[\frac{z_i}{g(x_i)} \right] = E_x^0 \left[\frac{e(x_i)}{g(x_i)} \right] \neq 1.$$

Then the IPW estimator based on $g(\cdot)$

$$\tilde{\mu}_{IPW}(\mu) = \frac{1}{n} \sum_{i=1}^n \frac{z_i}{g(x_i)} y_i$$

is biased as its expectation is

$$E_x^0 \left[\frac{e(x_i)}{g(x_i)} \mu(x_i, 1) \right]$$

However, if $\mu(X_i, z)$ is correctly specified

$$E_{X,Y,Z}^0 \left[\frac{Z}{g(X)} (Y - \mu(X, 1)) \right] = 0$$

as $E_{Y|X,Z}^0 [Y/X, Z=1] = \mu(X, 1)$.

$\therefore$ AIPW estimator is still unbiased.

i.e. AIPW estimator is unbiased even if one of

- $e(x)$

- $\mu(X, z)$

is incorrectly specified.

- 'robust' to mis-specification of $e(x)$ or $\mu(X, z)$

- 'doubly robust'.

- parameter estimation part.

G-estimation

We construct another estimating equation with expectation zero ('unbiased estimating equation').

Consider the structural model for binary z .

$$Y(z) = \mu_0(x) + z \mu_1(x).$$

i.e.

$$(Y(z) - z \mu_1(x)) = \mu_0(x)$$

$$\therefore E_{x, Y, z}^{\bullet} [Y(z) - z \mu_1(x) - \mu_0(x)] = 0$$

and

$$E_{Y|X, z}^{\bullet} [Y(z) - z \mu_1(x) | x] = \mu_0(x).$$

- RHS does not depend on z .

We can construct a semiparametric model as before:

$$f_{x,y,z}(x,y,z) = f_{\varepsilon}(y - \mu_0(x; \beta) - z\mu_1(x; \psi)) f_{x,z}(x,z).$$

- (r.x.1) β -parameterizes μ_0 ; true value β_0
 (g.x.1) ψ -parameterizes μ_1 ; true value ψ_0 $\supset$ parametric submodel

$$\mu_0(x; \beta_0) \equiv \mu_0(x) \quad \text{data generating model.}$$

S_{β}, S_{ψ} yield the estimation procedure.

Parametric submodel: $f_{\varepsilon}(\varepsilon) \equiv f_{\varepsilon}(\varepsilon, \sigma_{\bullet}) \equiv N(0, \sigma^2)$.
 (i.e. our procedure)

$$S_{\beta} = \frac{1}{\sigma^2} (y - \mu_0(x) - z\mu_1(x; \psi_0)) \mu_{0\beta}(x; \beta_0).$$

$$S_{\psi} = \frac{1}{\sigma^2} (y - \mu_0(x) - z\mu_1(x; \psi_0)) z \mu_{1\psi}(x; \psi_0).$$

$$\mu_{0p}(x; \beta_0) = \frac{\partial}{\partial \beta} \{ \mu_0(x; \beta) \}_{\beta = \beta_0}$$

$$\mu_{1p}(x; \psi_0) = \frac{\partial}{\partial \psi} \{ \mu_1(x; \psi) \}_{\psi = \psi_0}$$

$$S_\psi = \frac{1}{\sigma^2} \sum z \mu_{1p}(x; \psi_0)$$

$$S_p = \frac{1}{\sigma^2} \sum \mu_{0p}(x; \beta_0)$$

S_p defines the nuisance target space

$$\Lambda = \left\{ \sum h(x) : h(x) \text{ arbitrary}, h(x) \in \mathbb{R}^n \right\}$$

The projection of S_ψ onto Λ takes the form

$$\sum h_0(x)$$

where

$$\mathbb{E}_{x,y,z} \left[\left(\varepsilon z \mu_{1\psi}(\mathbf{x}; \psi_0) - \varepsilon h_0(\mathbf{x}) \right)^T \varepsilon h(\mathbf{x}) \right] = 0.$$

$$\Rightarrow \mathbb{E}_{x,y} \left[\varepsilon^2 \left(\cancel{e(x)} \mu_{1\psi}(X; \psi_0) - h_0(X) \right)^T h(X) \right] = 0$$

$$\forall h(X).$$

$$\Rightarrow h_0(X) = e(X) \mu_{1\psi}(X; \psi_0).$$

$\therefore$ projection is $h_0(X) = e(X) \mu_{1\psi}(X; \psi_0)$
residual is

$$\varepsilon (z - e(X)) \mu_{1\psi}(X; \psi_0)$$

$\therefore$ Efficient score is

$$S_{\psi}^{\text{eff}} = \varepsilon (z - e(X)) \mu_{1\psi}(X; \psi_0).$$

$\therefore$ Estimating equation is

$$\sum_{i=1}^n (z_i - e(x_i)) (y_i - \mu_0(x_i; \beta) - z_i \mu_1(x_i; \psi))_{\mu, \psi} = 0.$$

i.e we solve jointly

$$\sum_{i=1}^n \begin{pmatrix} S_{\beta i} \\ S_{\psi i}^{\text{eff}} \end{pmatrix} = 0.$$

$$\sum_{i=1}^n \begin{pmatrix} \mu_{0\beta}(x_i; \beta) \\ \mu_{1\psi}(x_i; \psi)(z_i - e(x_i)) \end{pmatrix} (y_i - \mu_0(x_i; \beta) - z_i \mu_1(x_i; \psi)).$$

In the linear case

$$\mu_0(x_i; \beta) = x_i \beta$$

$$\mu_1(x_i; \psi) = x_{\psi i} \psi.$$

$$\sum_{i=1}^n \begin{pmatrix} x_{pi}^T \\ x_{\psi i}^T (z_i - e(x_i)) \end{pmatrix} (y_i - x_{pi}^T \beta - x_{\psi i}^T \psi) = 0.$$

- 'G - estimation'

- $\mu_1(x; \psi)$ - known functional form.

- doubly robust to misspecification of

- $e(x_i)$
- $\mu_0(x_i; \beta)$.

Connection to PSR

$$E[Y_i | X_i = x_i, Z_i = z_i] = \mu_0(x_i; \beta) + z_i \mu_1(x; \psi) + e(x_i) \mu_1(x; \gamma).$$

- fit using OLS.

In this case.

$$\sum_{i=1}^n \begin{pmatrix} \mu_{0\beta}(x_i; \beta) \\ \mu_{1\psi}(x_i; \psi) z_i \\ \mu_{1\delta}(x_i; \delta) e(x_i) \end{pmatrix} (y_i - \mu_{0\beta}(x_i; \beta) - z_i \mu_{1\psi}(x_i; \psi) - e(x_i) \mu_{1\delta}(x_i; \delta)) = 0.$$

— subtract third equation from second

— G - estimating equations

NOTE If $e(x_i) \equiv e(x_i; \alpha)$, we estimate α using another (parametric) model.

Outcome regression

- optimal under correct specification of $\mu(x; \beta, \psi)$.

AIPW

- optimal if only one of $e(x)$ or $\mu(x; \beta, \psi)$ can be guaranteed correct.

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

G-estimation

- optimal if $\mu_1(x; \psi)$ correct and either one of $e(x)$ or $\mu_0(x; \beta)$ correct.

NOTE Should use the semiparametric
(sandwich) variance estimator for the
efficient score

- should use adjustnet for
estimation of all nuisance parameters.

(i.e. β and α).

- need to compute

$$\text{Var}[\eta_{\psi}^{\text{eff}}] = \left\{ E[\{S_{\psi}^{\text{eff}}\}\{S_{\psi}^{\text{eff}}\}^T] \right\}^{-1}$$

Var

ALPW Estimation by Regression

Consider the model for binary z

$$E_{y|x,z}^0[Y|X=x, z=z] = \mu(u, z) + \phi_0 \frac{(1-z)}{1-e(u)} + \phi_1 \frac{z}{e(u)}$$

fitted by OLS.

Score equations for ϕ_0, ϕ_1

$$\textcircled{1} \frac{\partial}{\partial \phi_0} : \sum_{i=1}^n \left(\frac{1-z_i}{1-e(u_i)} \right) \left(y_i - \mu(u_i, z_i) - \phi_0 \frac{(1-z_i)}{1-e(u_i)} - \phi_1 \frac{z_i}{e(u_i)} \right) = 0.$$

$$\textcircled{2} \frac{\partial}{\partial \phi_1} : \sum_{i=1}^n \frac{z_i}{e(u_i)} \left(y_i - \mu(u_i, z_i) - \phi_0 \frac{(1-z_i)}{1-e(u_i)} - \phi_1 \frac{z_i}{e(u_i)} \right) = 0.$$

$$w_{0i} = \left(\frac{1-z_i}{1-e(u_i)} \right) \quad w_{1i} = \frac{z_i}{e(u_i)}.$$

$$\Rightarrow \hat{\phi}_z = \frac{\sum_{i=1}^n w_{zi} (y_i - \mu(x_i, z_i))}{\sum_{i=1}^n w_{zi}^2} \quad z = 0, 1.$$

Consider the predictions for $z=0, z=1$ case

$$\mu(x_i, 0) + \hat{\phi}_0 \frac{1}{1 - e(x_i)}$$

$$\mu(x_i, 1) + \hat{\phi}_1 \frac{1}{e(x_i)}$$

$\therefore$ Average across data as

$$\tilde{\mu}(0) = \frac{1}{n} \sum_{i=1}^n \left\{ \mu(x_i, 0) + \hat{\phi}_0 \frac{1}{1 - e(x_i)} \right\}$$

$$= \frac{1}{n} \sum_{i=1}^n \mu(x_i, 0) + \hat{\phi}_0 \frac{1}{n} \sum_{i=1}^n \frac{1}{1 - e(x_i)}$$

$$= \frac{\frac{1}{n} \sum_{i=1}^n \mu(x_i, 0) + \frac{\sum_{i=1}^n \frac{1}{1 - e(x_i)}}{\sum_{i=1}^n w_{0i}^2}}{\frac{1}{n} \sum_{i=1}^n w_{0i} (y_i - \mu(x_i, 0))}$$

$$\tilde{\mu}(1) = \frac{\frac{1}{n} \sum_{i=1}^n \mu(x_i, 1) + \frac{\sum_{i=1}^n \frac{1}{e(x_i)}}{\sum_{i=1}^n w_{1i}^2}}{\frac{1}{n} \sum_{i=1}^n w_{1i} (y_i - \mu(x_i, 1))}$$

Now

$$\frac{\sum_{i=1}^n \frac{1}{e(x_i)}}{\sum_{i=1}^n W_{1i}^2} = \frac{\frac{1}{n} \sum_{i=1}^n \frac{1}{e(x_i)}}{\frac{1}{n} \sum_{i=1}^n \left(\frac{z_i}{e(x_i)}\right)^2} \xrightarrow{P} \frac{E_x^0\left[\frac{1}{e(x)}\right]}{E_{x|z}^0\left[\frac{z^2}{e(x)^2}\right]}.$$

But

$$E_{x|z}^0\left[\frac{z^2}{e(x)^2}\right] = E_{x|z}^0\left[\frac{z}{e(x)^2}\right] = E_x^0\left[\frac{1}{e(x)}\right].$$

$$\therefore \frac{\sum_{i=1}^n \frac{1}{e(x_i)}}{\sum_{i=1}^n W_{1i}^2} \xrightarrow{P} 1 \quad \text{as } n \rightarrow \infty.$$

$$\therefore \hat{\mu}(1) \xrightarrow{P} \tilde{\mu}_{HPW}(1) + o_p(1).$$

Similarly

$$\hat{\mu}(0) = \tilde{\mu}_{HPW}(0) + o_p(1).$$

∴ Can obtain a consistent APLW estimator
by regressing Y on

$$\mu(x, z) + \phi_0 \frac{(1-z)}{(1-e(x))} + \phi_1 \frac{z}{e(x)}$$

and estimating ϕ_0, ϕ_1 by OLS.

NOTE Can also use

$$\mu(x, z) + \phi \left(\frac{z}{e(x)} - \frac{(1-z)}{1-e(x)} \right)$$

to estimate

$$\delta = \mu(1) - \mu(0)$$

—uses a single ϕ .

5. EXTENSIONS

Beyond the linear model

The causal effect $E[Y(1) - Y(0)] \equiv E_{Y|Z}[Y|Z=1] - E_{Y|Z}[Y|Z=0]$ required may be amended to other functions.

$$E[\log Y(1) - \log Y(0)].$$

$$\log E[Y(1)] - \log E[Y(0)] = \log \left\{ \frac{E[Y(1)]}{E[Y(0)]} \right\}$$

— causal methods loosely extended to these cases but must remember the issue of collapsability

eg in a loglinear model

$$\log \{E[Y|X, Z]\} = X\beta + Z\psi$$

ψ captures the influence of Z .

However, note that for a average effect for α or

$$\begin{aligned} E_{y|z}[Y|z=z] &= E_x \left[E_{y|x,z}[Y|X, z=z] \right] \\ &= E_x \left[\exp\{X_\beta \beta + z X_\psi \psi\} \right] \end{aligned}$$

would be estimated by

$$\frac{1}{n} \sum_{i=1}^n \exp\{X_{\beta i} \hat{\beta} + z X_{\psi i} \hat{\psi}\}$$

i.e. must perform the transformation.

-techniques listed previously
still valid, must be applied with care

Continuous Treatments

In general

$$B \equiv B(X, Z) \equiv f_{Z|X}^0(Z|X)$$

is a balancing score: can be utilized in causal adjustment.

However, suppose

$$f_{Z|X}^0(Z|X) \equiv f_{Z|S}^0(Z|S)$$

for some sufficient statistic $S \equiv S(X)$.

Then $B_S \equiv B(S, Z)$ is also a balancing score, and we may proceed using it.

Furthermore, in a parametric model, suppose we have

$$f_{z|x}^0(z|x;\alpha) \equiv f_{z|x}^0(z|x\alpha).$$

i.e. dependent on a linear combination of x components

eg

$$f_{z|x}^0(z|x) = \text{Normal}(x\alpha, \sigma^2)$$

say.

Then $f_{z|x}^0(z|x)$ is entirely determined by $x\alpha$.

$\therefore$ We may 'balance' using (approximately) equal $x\alpha$ values.

- regress z on X using OLS - estimate α

- stratify on values of $x\hat{\alpha}$.

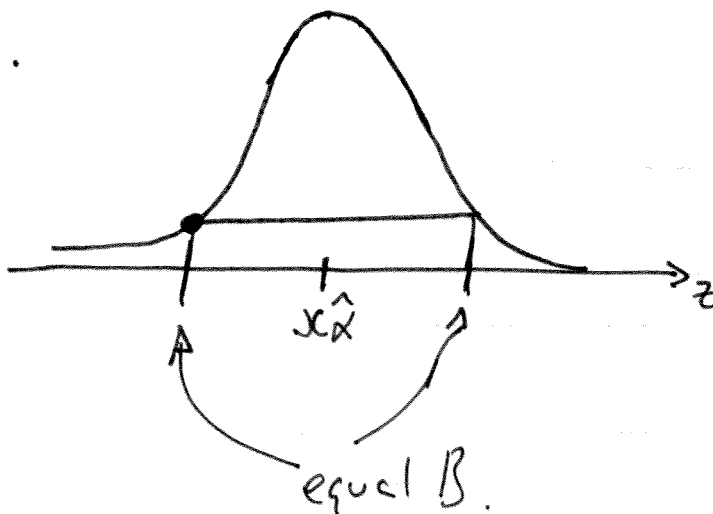
- compare sets of equal

$$(a) \quad B = f_{z|x}^0(z|x) = \left(\frac{1}{2\pi}\right)^{1/2} \exp\left\{-\frac{1}{2}(z - x\hat{\alpha})^2\right\}.$$

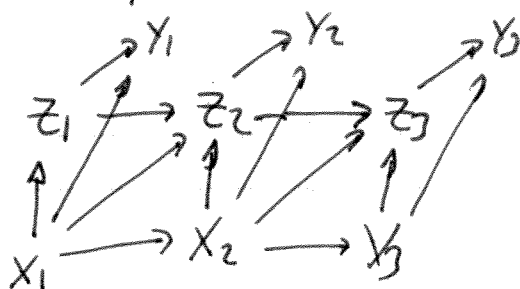
with sets of equal

$$(b) \quad x\hat{\alpha}$$

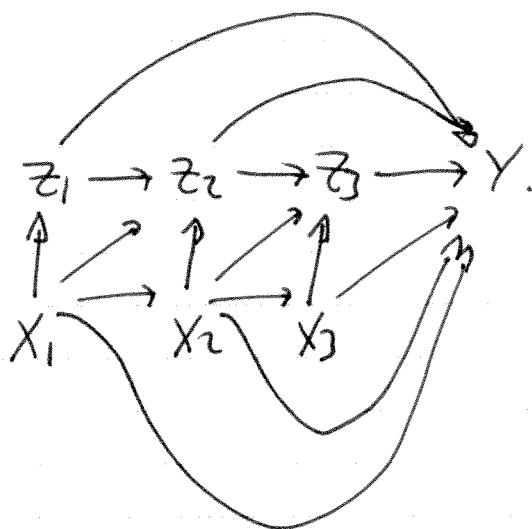
For (a)



Longitudinal Exposures.



or



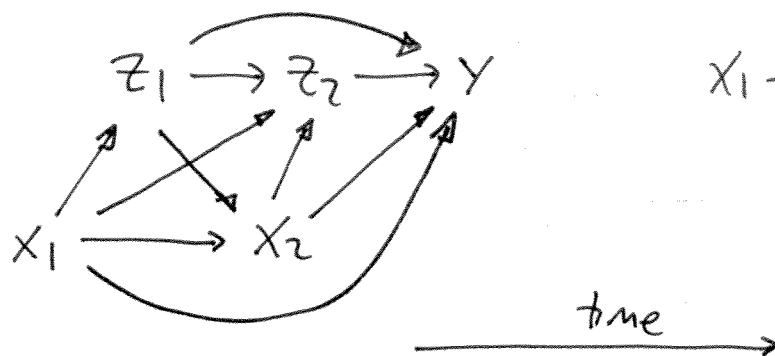
etc. Only 'forward in time' arrows.

Many possible investigations.

Consider n general K successive treatments

Investigate the 'total' effect of $z_1, \dots, z_K$.

K=2



x_1 - 'baseline'.

$$f_{x_1}(x_1) f_{z_1|x_1}(z_1|x_1)$$

$$f_{x_2|x_1, z_1}(x_2|x_1, z_1) f_{z_2|x_1, z_1, x_2}(z_2|x_1, z_1, x_2).$$

$$f_{y|x_1, z_1, x_2, z_2}(y|x_1, z_1, x_2, z_2).$$

- with knowledge of all of these models, under correct specification can recover

$$E_{y|z}^{\epsilon}[Y|Z=z] \quad z = (z_1, z_2) \quad z = (z_1, z_2).$$

$$= \frac{\iiint \mathbb{I}_{\{z\}}(z') y f_{y|x,z}^{\epsilon}(y|x,z') f_x^{\epsilon}(x) f_z^{\epsilon}(z') dy dx dz}{\iint \mathbb{I}_{\{z\}}(z') f_z^{\epsilon}(z') dz'}.$$

$$= \frac{\mathbb{E}_{X,Y,Z}^{\circ} \left[\frac{\mathbb{I}_{\{Z\}}(Z) Y}{f_{Z|X}^{\circ}(Z|X)} \right]}{\mathbb{E}_Z^{\circ} \left[\frac{\mathbb{I}_{\{Z\}}(Z)}{f_{Z|X}^{\circ}(Z|X)} \right]}.$$

Here

$$f_{Z|X}^{\circ}(Z|X) = f_{Z_1|X_1}^{\circ}(Z_1|X_1) f_{Z_2|X_1, Z_1, X_2}^{\circ}(Z_2|X_1, Z_1, X_2).$$

- in a parametric analysis, need to fit

$$f_{Z_1|X_1}^{\circ}(Z_1|X_1; \alpha_1)$$

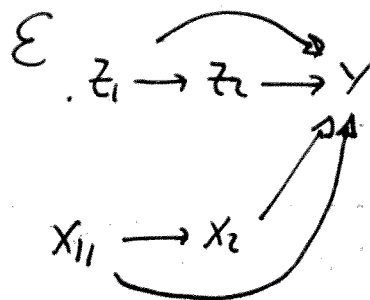
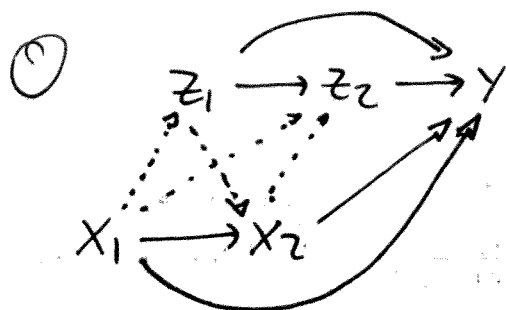
$$f_{Z_2|X_1, Z_1, X_2}^{\circ}(Z_2|X_1, Z_1, X_2; \alpha_2).$$

Typically, also model

$$\mathbb{E}_{Y|X_1, Z}^{\circ} [Y | X_1 = x_1, Z_1 = z_1, X_2 = x_2, Z_2 = z_2].$$

using regression approaches.

But here the 'experimental' calculation needs some justification: we need the unconfounded (total) effect of z_1 and z_2 .



- conditioning on X_2 blocks the indirect effect z_1 (via $z_1 \rightarrow X_2 \rightarrow Y$) $\therefore$ don't get the total effect.
- conditioning on X_1 is possible but influences the extracted effect.

$\therefore$ Can model

$$f_{Y|X,Z}^{\epsilon}(y|x,z) \equiv f_{Y|Z_1,Z_2}^{\epsilon}(y|z_1,z_2).$$

$$\text{or } f_{Y|X,Z}^{\epsilon}(y|x,z) \equiv f_{Y|X,Z_1,Z_2}^{\epsilon}(y|x,z_1,z_2).$$

$$\text{but not } f_{Y|X,Z}^{\epsilon}(y|x,z) \equiv f_{Y|X,Z_1,X_2,Z_2}^{\epsilon}(y|x,z_1,z_2,x_2,z_2).$$

∴ Can consider

$$E_{X,Z}^0 \left[\frac{I_{\{z\}}(Z)}{f_{X|Z}^0(Z|X)} \cdot E_{Y|X,Z}^0 [Y | Z_1 = z_1, Z_2 = z_2] \right]$$

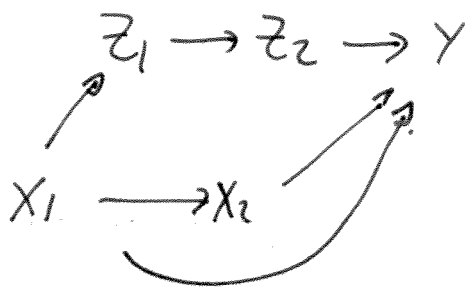
or

$$E_{X,Z}^0 \left[\frac{I_{\{z\}}(Z)}{f_{X|Z}^0(Z|X)} E_{Y|X,Z}^0 [Y | X_1 = x_1, Z_1 = z_1, Z_2 = z_2] \right]$$

Also, sometimes the effect of non time-varying mediating variables is considered.

1-2

$$E_{Y|X,Z}^e [Y | X_1 = x_1, Z = z].$$



- conditioning on X_1 does not affect assessment of the 'total' effect.

$$E_{Y|X_1, Z}^E[Y|X_1=x_1, Z=z].$$

$$= \frac{\iiint \mathbb{I}_{\{z\}}(z') y f_{Y|X_1, Z}^E(y|x_1, z) f_{Z|X_1}^E(z'|x_1) f_{X_2|X_1}^E(x_2|x_1) dz' dx_2}{\int f_{Z|X_1}^E(z'|x_1) dz'}$$

$\therefore$ 'Weight' is based on

$$\frac{f_{Z|X_1}^E(z|x_1)}{f_{Z|X}^0(z|x)}$$

1.2

$$f_{z|x_1}^{\epsilon}(z|x_1) = f_{z_1|x_1}^{\epsilon}(z_1|x_1) f_{z_2|x_1, z_1}^{\epsilon}(z_2|x_1, z_1).$$

$$f_{z|x}^0(z|x) = f_{z_1|x_1}^0(z_1|x_1) f_{z_2|x_1, x_{z_1}}^0(z_2|x_1, x_{z_1}).$$

- 'stabilized weights'.

- different causal effect!

- need a model of $f_{z|x_1}^{\epsilon}(z|x_1)$