Part 6
Extensions

Extensions

The development above has mainly focussed on

- binary treatment,
- linear models,
- average treatment effects,
- single time-point studies,

but extensions can be developed to handle each case.

Continuous treatments

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

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

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

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

Continuous treatments

Example: G-estimation

The model

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

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

This relies on the construction of a model for $\mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X]$

- justified by a focus on *orthogonality* under the *covariance* inner product;
- focus on conditional *uncorrelatedness* rather than conditional *independence*.

Continuous treatments

For IPW estimation, the construction

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

also works in the continuous case as

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

but in practice this estimator can be variable in finite sample.

Continuous treatments

The focus on $\mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X]$ emphasizes the role of *regression* approaches to constructing the treatment model: we denote

$$\xi(X) \equiv \xi(X;\alpha) = \mathbb{E}^{\mathcal{O}}_{Z|X}[Z|X].$$

In most cases, some form of *generalized linear model* will suffice, with estimating equation such as

$$\sum_{i=1}^{n} \mathbf{x}_i^\top (z_i - \xi(\mathbf{x}_i\alpha)) = \mathbf{0}$$

for parameter α so that the fitted values

$$\xi(\mathbf{x}_i\widehat{\alpha}) \quad i = 1, \ldots, n$$

can be computed.

Continuous treatments

In any subsequent procedure, these fitted values are used in estimating the causal effects: for example, in *G-estimation*, we consider the *plug-in* estimator

$$\widehat{\delta}_{\text{G}} = \frac{\sum_{i=1}^{n} Y_i(Z_i - \xi(\mathbf{X}_i\widehat{\alpha}))}{\sum_{i=1}^{n} Z_i(Z_i - \xi(\mathbf{X}_i\widehat{\alpha}))}$$

In IPW estimation, for binary treatment, we use

$$\widehat{\mu}_{\text{IPW}}(1) = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{e(X_i;\widehat{\alpha})}.$$

Advanced modelling

The application of many of the causal adjustment methods rely on regression modelling for

- the outcome mean model

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

- the propensity or balancing score

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

Advanced modelling

In either case we can utilize

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

to construct fitted versions of each model.

Advanced modelling

The advanced methods can be effective, but there are potential pitfalls:

1. The quantification of *uncertainty*;
   - no ready analytic answers, typically relies on bootstrap;
   - large computational burden
2. *Positivity violations*.
   - prediction (of treatment mechanism) is not the fundamental goal;
   - can be overcome using methods that target balance/overlap.

Beyond linear models

If the outcome variable Y is discrete, or it is contended that the outcome mean model

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X = \mathbf{x}, Z = z] = \mu(\mathbf{x}, z; \beta, \psi)$$

is not linear (in the treatment or parameters), it is necessary to extend some of the previous concepts.

In the following,

- X is the random vector of confounders, and x its observed values,
- $\mathbf{x}$ is the vector of functions of x that form the linear predictor,
- $\mathbf{X}$ is the random variable version of $\mathbf{x}$.

Beyond linear models

Suppose that a *log-linear* model is deemed appropriate:

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

Regarding this as the structural model, we then have that

$$\mathbb{E}[Y(z)] \equiv \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y|Z = z] = \mathbb{E}^{\mathcal{E}}_{X}[\exp\{\mathbf{X}_\beta \beta + z\mathbf{X}_\psi \psi\}]$$

which would then invoke the estimator

$$\widetilde{\mu}(z) = \frac{1}{n} \sum_{i=1}^{n} \exp\{\mathbf{x}_{\beta i} \widehat{\beta} + z\mathbf{x}_{\psi i} \widehat{\psi}\}.$$

where $(\widehat{\beta}, \widehat{\psi})$ are estimated from a correctly specified model.

Beyond linear models

Then in the binary treatment case

$$\frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} = \frac{\mathbb{E}_X^{\mathcal{E}}[\exp\{\mathbf{X}_\beta\beta + \mathbf{X}_\psi\psi\}]}{\mathbb{E}_X^{\mathcal{E}}[\exp\{\mathbf{X}_\beta\beta\}]}$$

but notice that unlike in the linear case

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

depend on *both* β and ψ. Also

$$\frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} \neq \mathbb{E}\left[\frac{Y(1)}{Y(0)}\right]$$

in general.

Beyond linear models

The above model assumes that

$$\mu(\mathbf{x}, 1; \beta, \psi) = \mu(\mathbf{x}, 0; \beta, \psi) \exp\{\mathbf{x}_\psi \psi\}.$$

with the effect of Z represented *conditional* on X; *marginally*, ψ alone does not capture the effect of treatment.

We could formulate a model for the potential outcomes where

$$Y(1) = Y(0) \exp\{\mathbf{x}_\psi \psi\}$$

that is

$$\log Y(1) - \log Y(0) = \mathbf{x}_\psi \psi.$$

Beyond linear models

If we can estimate β and ψ consistently, then we can recover APO and ATE estimators from them. By standard regression arguments, we know that *correct specification* of the outcome model is needed.

For the log-linear model,

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

the standard estimating equation takes the form

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^\top \\ z_i \mathbf{x}_{\psi i}^\top \end{pmatrix} (y_i - \exp\{\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi\}) = \mathbf{0}.$$

Beyond linear models

We might try to make inference robust to mis-specification using a *G-estimation*-like strategy, and modify the estimating equation to be

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^{\top} \\ (z_i - e(x_i))\mathbf{x}_{\psi_i}^{\top} \end{pmatrix} (y_i - \exp\{\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi\}) = \mathbf{0}$$

where $e(x_i)$ is the propensity score; if necessary, this can be estimated using a further parametric model

$$\sum_{i=1}^{n} \mathbf{x}_i^{\top}(z_i - e(x_i;\alpha)) = \mathbf{0}.$$

Beyond linear models

In the *linear* case, this modification led to consistent estimation of ψ as the resulting estimating equation is *unbiased*, that is

$$\mathbb{E}_{X,Y,Z}[(Z - e(X))\mathbf{X}_\psi^\top(Y - \mathbf{X}_\beta\beta - Z\mathbf{X}_\psi\psi)] = \mathbf{0}$$

provided $\mathbf{x}_\psi\psi$ correctly captures the effect of treatment, *and either*

- propensity score $e(X)$, *or*
- treatment-free mean component $\mathbf{x}_\beta\beta$

is correctly specified.

Beyond linear models

In the *log-linear* case, the equivalent requirement would be

$$\mathbb{E}_{X,Y,Z}[(Z - e(X))\mathbf{X}_\psi^\top (Y - \exp\{\mathbf{X}_\beta \beta + Z\mathbf{x}_\psi \psi)\}] = \mathbf{0}.$$

However, this requirement is *not* met if the treatment-free mean component is mis-specified, even if the propensity score model is correctly specified.

Beyond linear models

Suppose in reality

$$\mathbb{E}_{Y|X,Z}[Y|X = \mathbf{x}, Z = z] = \mu(\mathbf{x}, z) = \mu_0(\mathbf{x}) \exp\{z\mathbf{x}_\psi\psi\}.$$

Then

$$\begin{aligned}\mathbb{E}_{Y|X,Z}[Y - \exp\{\mathbf{x}_\beta\beta + Z\mathbf{x}_\psi\psi)\}|X = \mathbf{x}, Z = z] \\ = \mu_0(\mathbf{x}) \exp\{z\mathbf{x}_\psi\psi\} - \exp\{\mathbf{x}_\beta\beta + z\mathbf{x}_\psi\psi)\} \\ = \exp\{z\mathbf{x}_\psi\psi\}(\mu_0(\mathbf{x}) - \exp\{\mathbf{x}_\beta\beta\})\end{aligned}$$

which cannot be made independent of z unless

$$\mu_0(\mathbf{x}) \equiv \exp\{\mathbf{x}_\beta\beta\}.$$

Beyond linear models

We must modify the estimating function to be

$$\varphi(X, Y, Z) = (Z - e(X))\mathbf{X}_\psi^\top \exp\{-Z\mathbf{X}_\psi\psi\}(Y - \exp\{\mathbf{X}_\beta\beta + Z\mathbf{X}_\psi\psi\})$$

so that

$$\begin{aligned}
&\mathbb{E}_{Y|X,Z}[\varphi(X, Y, Z)|X = x, Z = z] \\
&\quad = (z - e(x))\mathbf{x}_\psi^\top \exp\{-z\mathbf{x}_\psi\psi\} \exp\{z\mathbf{x}_\psi\psi\}(\mu_0(x) - \exp\{\mathbf{x}_\beta\beta\}) \\
&\quad = (z - e(x))\mathbf{x}_\psi^\top(\mu_0(x) - \exp\{\mathbf{x}_\beta\beta\})
\end{aligned}$$

It then follows that

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

$$=\mathbf{x}_{\psi}^{\top}(\mu_0(\mathbf{x})-\exp\{\mathbf{x}_{\beta}\beta\})\mathbb{E}_{Z|X}[(Z-e(X))|X=\mathbf{x}]=\mathbf{0}$$

if the propensity model is correctly specified. Thus for consistent estimation, we need to solve

$$\sum_{i=1}^{n}\begin{pmatrix}\mathbf{x}_{\beta i}^{\top}\\ (z_i-e(\mathbf{x}_i))\mathbf{x}_{\psi_i}^{\top}\exp\{-z_i\mathbf{x}_{\psi i}\psi\}\end{pmatrix}(y_i-\exp\{\mathbf{x}_{\beta i}\beta+z_i\mathbf{x}_{\psi i}\psi\})=\mathbf{0}$$

Beyond linear models

If outcome Y is *binary*, we might attempt to use a *logistic* model and estimate using

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^{\top} \\ z_i \mathbf{x}_{\psi_i}^{\top} \end{pmatrix} (y_i - \text{expit}\{\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi\}) = \mathbf{0}.$$

where

$$\text{expit}(x) = \frac{e^x}{1 + e^x}.$$

This method is not robust to mis-specification of the mean model, and it cannot be rescued using the previous trick.

Beyond linear models

A solution is found by noting that

$$\exp\{\mathbf{x}_\psi \psi\} = \frac{\Pr[Y=1|Z=1,X=x]/\Pr[Y=0|Z=1,X=x]}{\Pr[Y=1|Z=0,X=x]/\Pr[Y=0|Z=0,X=x]}$$

$$\equiv \frac{\Pr[Z=1|Y=1,X=x]/\Pr[Z=0|Y=1,X=x]}{\Pr[Z=1|Y=0,X=x]/\Pr[Z=0|Y=0,X=x]}$$

that is, ψ parameterizes the conditional *log-odds ratio*.

Beyond linear models

Therefore we can construct a doubly robust estimator assuming that **either**

- the model for Y given X and Z, **or**
- the model for Z given X and Y

is correctly specified, given that the log-odds model is correctly specified.

The R package `drgee` uses the estimating procedure based on first estimating parameters in the two systems

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\gamma i}^{\top} \\ y_i \mathbf{x}_{\psi i}^{\top} \end{pmatrix} \left(z_i - \text{expit}(\mathbf{x}_{\gamma i}\gamma + y_i \mathbf{x}_{\psi i}\psi^{\dagger}) \right) = \mathbf{0}$$

and

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{\beta i}^{\top} \\ z_i \mathbf{x}_{\psi i}^{\top} \end{pmatrix} \left(y_i - \text{expit}(\mathbf{x}_{\beta i}\beta + z_i \mathbf{x}_{\psi i}\psi^{\ddagger}) \right) = \mathbf{0}$$

utilizing two additional nuisance parameters $\psi^{\dagger}$ and $\psi^{\ddagger}$. This yields estimates of β and γ.

Then ψ is estimated using

$$\sum_{i=1}^{n}(z_i - e^*(\mathrm{x}_i;\psi,\widehat{\beta},\widehat{\gamma}))\mathbf{x}_{\psi i}^\top \left(y_i - \text{expit}(\mathbf{x}_{\beta i}\widehat{\beta} + z_i\mathbf{x}_{\psi i}\psi)\right) = \mathbf{0}$$

where

$$e^*(\mathrm{x};\psi,\widehat{\beta},\widehat{\gamma}) = \left[1 + \frac{1-\text{expit}(\mathbf{x}_\gamma\widehat{\gamma})}{\text{expit}(\mathbf{x}_\gamma\widehat{\gamma})}\frac{\text{expit}(\mathbf{x}_\beta\widehat{\beta})}{\text{expit}(\mathbf{x}_\beta\widehat{\beta} + \mathbf{x}_\psi\psi)}\right]^{-1}.$$

It can be shown that this is an unbiased estimating equation.

Average Treatment Effect on the Treated

In the binary case, we have mainly focussed on the estimation of the average treatment effect (ATE)

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y \mid Z = 1] - \mathbb{E}^{\mathcal{E}}_{Y|Z}[Y \mid Z = 0]$$

The *average treatment effect on the treated* (ATT) is

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

using the potential outcome notation.

Average Treatment Effect on the Treated

That is, the ATT aims to identify the causal effect on intervening to change $Z = 0$ to $Z = 1$ but *only* in the subpopulation of individuals who are *observed* to receive treatment

Note that

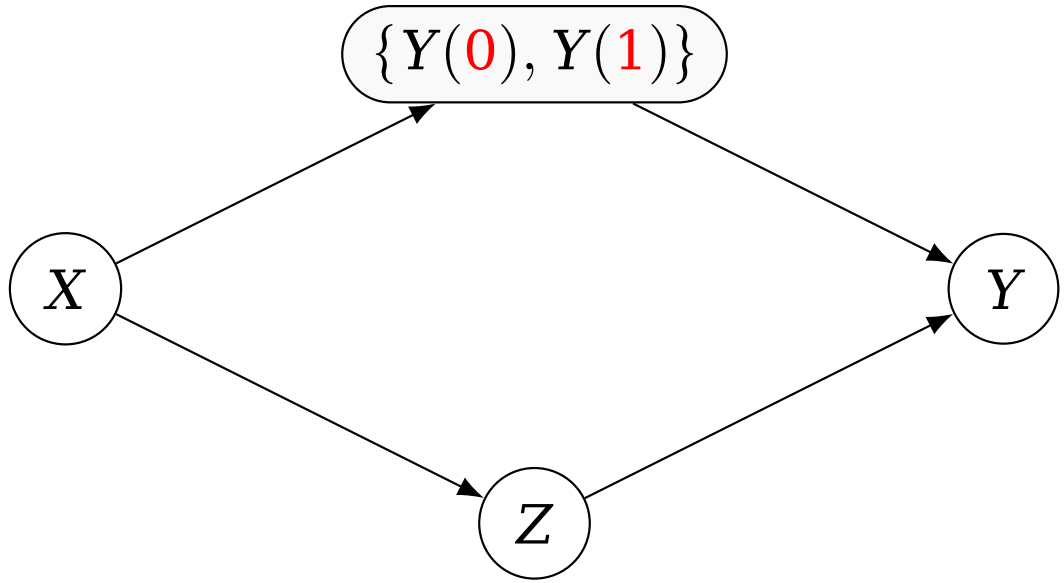
$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1) - Y(0) \mid Z = 0]\Pr[Z = 0]$$
$$+ \mathbb{E}[Y(1) - Y(0) \mid Z = 1]\Pr[Z = 1]$$

Average Treatment Effect on the Treated

In this calculation, we imagine

- the observational distribution $f^{\mathcal{O}}_{X,Y,Z}$ generating the observed data $\{(x_i, y_i, z_i), i = 1, \ldots, n\}$
- in the subgroup observed to have $Z = 1$, we then consider a second (hypothetical) experimental intervention to change Z to z which over-rides the original Z if $z = 0$,
- we then consider comparison of hypothetical outcomes between the two hypothetical subgroups indexed by z.

Average Treatment Effect on the Treated



$$f_X(x)\, f_{Z|X}(z|x)\, f_{Y(0),Y(1)|X}(y_0, y_1|x)\, f_{Y|Z,Y(0),Y(1)}(y|z, y_0, y_1)$$

defines the observational distribution, where

$$f_{Y|Z,Y(0),Y(1)}(y|z, y_0, y_1) = \begin{cases} 1 & y = (1-z)y_0 + zy_1 \\ 0 & \text{otherwise} \end{cases}.$$

Average Treatment Effect on the Treated

Then

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

$$= \int f_{Y|X,Z,Y(0),Y(1)}(y|x,z,y_0,y_1) f_{Y(0),Y(1)|X,Z}(y_0,y_1|x,z)\ dy_0\ dy_1$$

$$= \int f_{Y|Z,Y(0),Y(1)}(y|z,y_0,y_1) f_{Y(0),Y(1)|X}(y_0,y_1|x)\ dy_0\ dy_1$$

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

as the integral reduces to a evaluation at a single point where (y_0, y_1) satisfy

$$y = (1-z)y_0 + zy_1.$$

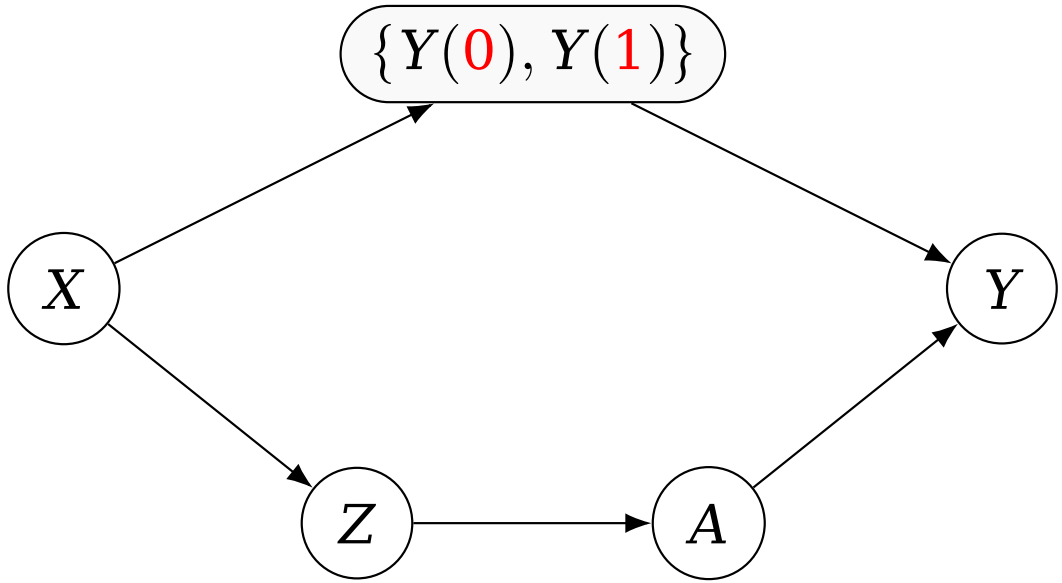
Average Treatment Effect on the Treated

For the ATT, we can represent the quantity of interest using a modified DAG that proposes a second hypothetical binary treatment, A.

We allow Z to cause A, and then allow Z to act as a selection mechanism, but ensure that

$$X \perp\!\!\!\perp A \mid Z.$$

Average Treatment Effect on the Treated



where the implied joint model is

$$f_X(x) f_{Z|X}(z|x) f_{Y(0),Y(1)|X}(y_0, y_1|x) f_{A|Z}(a|z) f_{Y|A,Y(0),Y(1)}(y|a, y_0, y_1)$$

This is the new 'experimental' distribution $\mathcal{E}$.

Average Treatment Effect on the Treated

Again

$$f^{\mathcal{E}}_{Y|A,Y(0),Y(1)}(y|a,y_0,y_1) = \begin{cases} 1 & y = (1-a)y_0 + ay_1 \\ 0 & \text{otherwise} \end{cases} .$$

As A is binary, the model $f_{A|Z}(a|z)$ must take the form

$$f^{\mathcal{E}}_{A|Z}(a|z) = p_z^a(1-p_z)^{1-a} \qquad a,z \in \{0,1\}$$

for $0 \leqslant p_z \leqslant 1$ for $z = 0, 1$.

We can then express the ATT via the new DAG as

$$\mathbb{E}^{\mathcal{E}}_{Y|A,Z}[Y|A=1,Z=1] - \mathbb{E}^{\mathcal{E}}_{Y|A,Z}[Y|A=0,Z=1]$$

Average Treatment Effect on the Treated

We have that from the DAG that

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

and hence as before

$$\begin{aligned} & f^{\mathcal{E}}_{Y|A,X}(y|a,x) \\ &= \int f^{\mathcal{E}}_{Y|A,Y(0),Y(1)}(y|a,y_0,y_1) f^{\mathcal{E}}_{Y(0),Y(1)|A,X}(y_0,y_1|a,x)\ dy_0\ dy_1 \\ &= f^{\mathcal{E}}_{Y(a)|X}(y|x). \end{aligned}$$

Also from the DAG, $X \perp\!\!\!\perp A \mid Z$, so for all a, x, z

$$f^{\mathcal{E}}_{X|A,Z}(x|a,z) \equiv f^{\mathcal{E}}_{X|Z}(x|z).$$

Average Treatment Effect on the Treated

For $\mathsf{a} = 0, 1$, we therefore have

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

$$= \iint y \, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x) f^{\mathcal{E}}_{X|Z}(x \mid z) \, dy \, dx.$$

Note that there is a potential *incompatibility* in the conditioning between

$$f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x) \qquad \text{and} \qquad f^{\mathcal{E}}_{X|Z}(x \mid z)$$

when we try to write the integral in terms of the data generating mechanism.

Average Treatment Effect on the Treated

As before, choosing the form of

$$f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x)$$

is to be avoided if possible.

We seek to resolve the incompatibility using the importance sampling trick, and write the expectation with respect to the observational model

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

Average Treatment Effect on the Treated

First note that

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

so the integral can be rewritten

$$\frac{1}{f^{\mathcal{E}}_{Z}(z)} \iint y \, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, \mathrm{x}) f^{\mathcal{E}}_{Z|X}(z \mid \mathrm{x}) f^{\mathcal{E}}_{X}(\mathrm{x}) \, dy \, dx.$$

Average Treatment Effect on the Treated

For $t = 0, 1$, we can re-write the integrand using the importance sampling trick as

$$y \, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x) \frac{f^{\mathcal{E}}_{Z|X}(z \mid x)}{f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x)} \, f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x) f^{\mathcal{E}}_{X}(x)$$

which can be rearranged to

$$\left\{ y \, \frac{f^{\mathcal{E}}_{Z|X}(z \mid x)}{f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x)} \right\} \, f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x) \, f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x) f^{\mathcal{E}}_{X}(x),$$

Average Treatment Effect on the Treated

Comparing the observational and experimental DAGs, we see that for all x and z

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

Also, we have for any t and y that

$$f^{\mathcal{E}}_{Y|A,X}(y \mid t, x) \equiv f^{\mathcal{O}}_{Y|X,Z}(y \mid x, t).$$

Therefore we have

$$f^{\mathcal{E}}_{Y|A,X}(y \mid \mathsf{a}, x)\, f^{\mathcal{E}}_{Z|X}(\mathsf{a} \mid x) f^{\mathcal{E}}_{X}(x) \equiv f^{\mathcal{O}}_{Y|X,Z}(y \mid x, \mathsf{a})\, f^{\mathcal{O}}_{Z|X}(\mathsf{a} \mid x) f^{\mathcal{O}}_{X}(x).$$

Average Treatment Effect on the Treated

Thus

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

$$= \frac{1}{f^{\mathcal{O}}_{Z}(z)} \iint \left\{ y \frac{f^{\mathcal{O}}_{Z|X}(z \mid x)}{f^{\mathcal{O}}_{Z|X}(\mathsf{a} \mid x)} \right\} f^{\mathcal{O}}_{X,Y,Z}(x, y, \mathsf{a})\, dy\, dx$$

$$= \frac{1}{f^{\mathcal{O}}_{Z}(z)} \iiint \left\{ \mathbb{1}_{\{\mathsf{a}\}}(t) y \frac{f^{\mathcal{O}}_{Z|X}(z \mid x)}{f^{\mathcal{O}}_{Z|X}(t \mid x)} \right\} f^{\mathcal{O}}_{X,Y,Z}(x, y, t)\, dy\, dx\, dt$$

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

Average Treatment Effect on the Treated

For the ATT, we are interested only in $z = 1$. The moment-based estimator is therefore

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|A,Z}[Y \mid A = \mathsf{a}, Z = 1] = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{\mathsf{a}\}}(Z_i) w_1(X_i, Z_i) Y_i}{\sum_{i=1}^{n} \mathbb{1}_{\{1\}}(Z_i)}$$

where

$$w_z(X_i, Z_i) = \frac{f^{\mathcal{O}}_{Z|X}(z \mid X_i)}{f^{\mathcal{O}}_{Z|X}(Z_i \mid X_i)}$$

Average Treatment Effect on the Treated

When $\mathsf{a} = 1$,

$$\mathbb{1}_{\{\mathsf{a}\}}(Z_i) w_1(X_i, Z_i) = \mathbb{1}_{\{1\}}(Z_i) = Z_i \quad \text{w.p. } 1$$

as the weight is identically 1, so therefore

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

that is, the mean in the treated group.

Average Treatment Effect on the Treated

When $\mathsf{a} = 0$,

$$\mathbb{1}_{\{\mathsf{a}\}}(Z_i) w_1(X_i, Z_i) = \mathbb{1}_{\{0\}}(Z_i) \frac{f^{\mathcal{O}}_{Z|X}(1 \mid X_i)}{f^{\mathcal{O}}_{Z|X}(Z_i \mid X_i)}$$

$$= (1 - Z_i) \frac{f^{\mathcal{O}}_{Z|X}(1 \mid X_i)}{f^{\mathcal{O}}_{Z|X}(0 \mid X_i)}.$$

Average Treatment Effect on the Treated

Therefore

$$\widehat{\mathbb{E}}^{\mathcal{E}}_{Y|A,Z}[Y \mid A = 0, Z = 1] = \frac{\sum_{i=1}^{n} (1 - Z_i) w(X_i) Y_i}{\sum_{i=1}^{n} Z_i}$$

where

$$w(X_i) = \frac{f^{\mathcal{O}}_{Z|X}(1 \mid X_i)}{f^{\mathcal{O}}_{Z|X}(0 \mid X_i)} = \frac{e(X_i)}{1 - e(X_i)}$$

That is, this estimator is a weighted sum of contributions from the *untreated* individuals.

Average Treatment Effect on the Treated

Thus the estimator for the ATT is

$$\frac{\sum_{i=1}^{n} \left(Z_i - (1 - Z_i) w(X_i)\right) Y_i}{\sum_{i=1}^{n} Z_i}.$$

Under the standard assumptions, this estimator is consistent for the ATT and asymptotically normally distributed if $e(x)$ is correctly specified; that is, it is *singly robust*.

Average Treatment Effect on the Treated

To achieve *double robustness*, we proceed in the usual fashion and augment the estimand for $\mathsf{a} = 0$ as follows:

$$\mathbb{E}[Y(0) \mid Z = 1] = \mathbb{E}[Y(0) - \mu(X, 0)|Z = 1] + \mathbb{E}[\mu(X, 0)|Z = 1]$$

where

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

is the modelled conditional mean for Y.

Average Treatment Effect on the Treated

To estimate the first term, we use

$$\widehat{\mathbb{E}}[Y(0) - \mu(X, 0) \mid Z = 1] = \frac{\sum_{i=1}^{n} (1 - Z_i) w(X_i)(Y_i - \mu(X_i, 0))}{\sum_{i=1}^{n} Z_i}$$

as in the singly robust case. For the second term, we have

$$\widehat{\mathbb{E}}[\mu(X, 0) \mid Z = 1] = \frac{\sum_{i=1}^{n} Z_i \mu(X_i, 0)}{\sum_{i=1}^{n} Z_i}$$

Average Treatment Effect on the Treated

Therefore

$$\widehat{\mathbb{E}}[Y(0) \mid Z = 1]$$

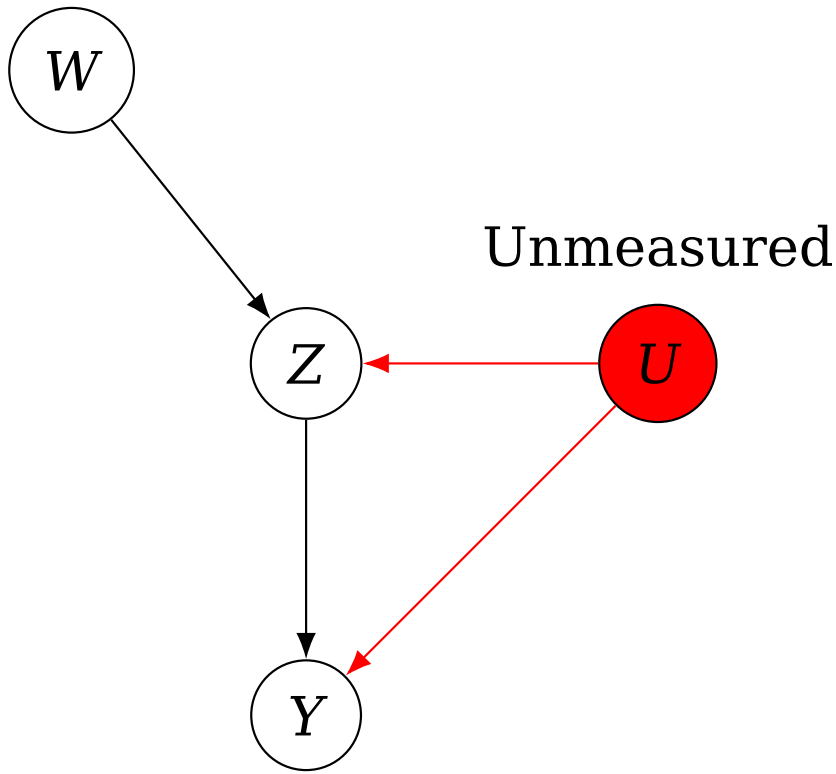
$$= \frac{\sum_{i=1}^{n} (1 - Z_i) w(X_i)(Y_i - \mu(X_i, 0)) + Z_i \mu(X_i, 0)}{\sum_{i=1}^{n} Z_i}$$

which yields the augmented ATT estimator

$$\frac{\sum_{i=1}^{n} (Z_i - (1 - Z_i) w(X_i))(Y_i - \mu(X_i, 0))}{\sum_{i=1}^{n} Z_i}.$$

Instrumental variables

Suppose we have the following system:



where $U \perp\!\!\!\perp W$, and $Y \perp\!\!\!\perp W \mid Z$.

Instrumental variables

We aim to find out the causal effect of Z on Y (both scalars).

- There is an open biasing path ZUY
- We cannot condition on U to block this path as it is unmeasured.

Instrumental variables

Suppose the structural model underlying this system is linear

$$Y_i = \beta_0 + Z_i\psi_0 + \mathbf{u}_i\zeta + \varepsilon_i$$

where

$$\mathbb{E}[\varepsilon_i|\mathbf{u}_i] = 0$$

and for all i, $Z_i \perp\!\!\!\perp \varepsilon_i$.

Clearly, regressing Y on Z alone will lead to inconsistent estimation of ψ_0.

Instrumental variables

Suppose now that W (also scalar) is observed and we in fact have the two-equation structural system

$$Y_i = \beta_0 + Z_i\psi_0 + \mathbf{u}_i\zeta + \varepsilon_i$$

$$Z_i = \gamma_0 + W_i\gamma_1 + \epsilon_i$$

with ϵ_i uncorrelated with W_i and ε_i. Then, by the usual trick, we have

$$W_iY_i = W_i\beta_0 + W_iZ_i\psi_0 + W_i\mathbf{u}_i\zeta + W_i\varepsilon_i$$

$$W_iZ_i = W_i\gamma_0 + W_i^2\gamma_1 + W_i\epsilon_i$$

Instrumental variables

Taking expectations and standardizing, using the usual arguments from linear regression, we have

$$\text{Cov}[W, Y] = \psi_0 \text{Cov}[W, Z]$$

$$\text{Cov}[W, Z] = \gamma_1 \text{Var}[W]$$

as W_i and U_i are *independent* by assumption.

Hence on rearrangement we have that

$$\psi_0 = \frac{\text{Cov}[W, Y]}{\text{Cov}[W, Z]} = \frac{\text{Cov}[W, Y]/\text{Var}[W]}{\text{Cov}[W, Z]/\text{Var}[W]} = \frac{\text{Cov}[W, Y]/\text{Var}[W]}{\gamma_1}.$$

Instrumental variables

This suggests a procedure to estimate ψ_0:

(i) regress Z on W to obtain estimate $\widehat{\gamma}_1$

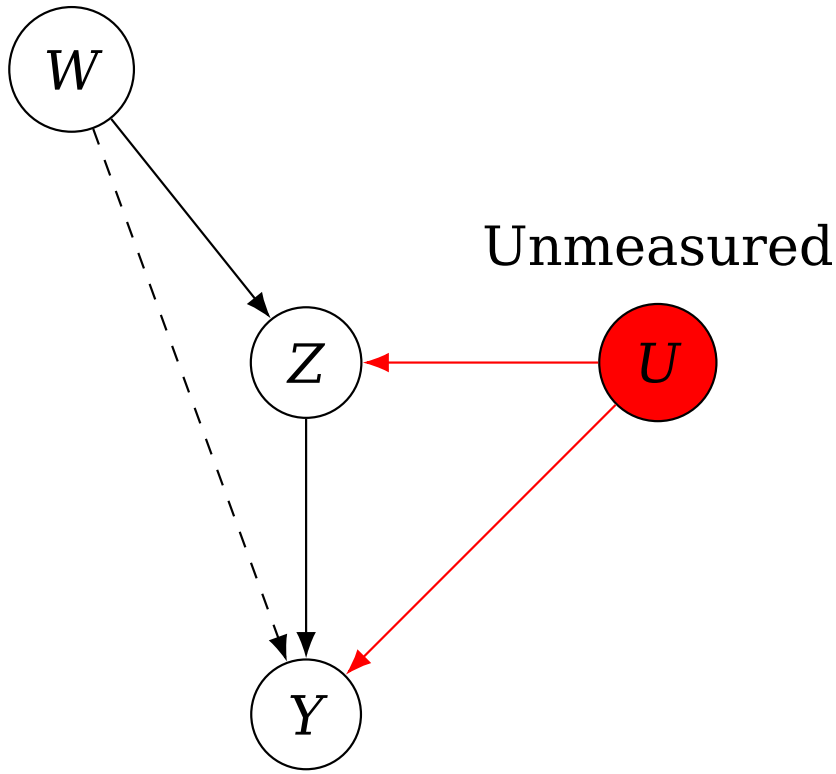
(ii) regress Y on W to obtain estimate $\widehat{\lambda}_1$

(iii) form the estimate of ψ_0 as

$$\widehat{\psi}_0 = \frac{\widehat{\lambda}_1}{\widehat{\gamma}_1}$$

This is a an *instrumental variable ratio estimator* using the *instrument* W.

Instrumental variables



Instrumental variables

Note

- For this method to work, we cannot have an *interaction* between Z and U in the structural model
- If there is only *weak association* between W and Z, so that γ_1 is close to zero, this estimator can have undesirable properties.
- Statistical properties of $\widehat{\psi}_0$ not trivial to derive.

Instrumental variables

A common situation in which the linear model is not quite appropriate is when W and Z are both *binary*. For example, we might have in a randomized study

- W is the randomized *allocated* treatment indicator; eg

$$W = 1 \quad \Longrightarrow \quad \text{“allocated to treatment arm”}$$

- Z is the actual treatment *received* indicator; eg

$$Z = 1 \quad \Longrightarrow \quad \textit{“took treatment”}$$

- Could have $W = 1, Z = 0$ (*non-compliance*);
- In some situations, could have $W = 0$, $Z = 1$.

Instrumental variables

Then we have by iterated expectation, for $w = 0, 1$

$$\begin{aligned}\mathbb{E}[Y|W=w] =& \mathbb{E}[Y|W=w, Z=0]\Pr[Z=0|W=w] \\ &+ \mathbb{E}[Y|W=w, Z=1]\Pr[Z=1|W=w] \\ =& \mathbb{E}[Y|Z=0]\Pr[Z=0|W=w] \\ &+ \mathbb{E}[Y|Z=1]\Pr[Z=1|W=w]\end{aligned}$$

as $Y \perp\!\!\!\perp W \mid Z$.

Instrumental variables

Therefore

$$\mathbb{E}[Y|W=1] = \mathbb{E}[Y|Z=0]\Pr[Z=0|W=1] + \mathbb{E}[Y|Z=1]\Pr[Z=1|W=1]$$

$$\mathbb{E}[Y|W=0] = \mathbb{E}[Y|Z=0]\Pr[Z=0|W=0] + \mathbb{E}[Y|Z=1]\Pr[Z=1|W=0]$$

Instrumental variables

Hence subtracting the second equation from the first

$$\begin{aligned}
&\mathbb{E}[Y|W=1] - \mathbb{E}[Y|W=0] \\
&\quad = \mathbb{E}[Y|Z=0](\Pr[Z=0|W=1] - \Pr[Z=0|W=0]) \\
&\quad + \mathbb{E}[Y|Z=1](\Pr[Z=1|W=1] - \Pr[Z=1|W=0])
\end{aligned}$$

Instrumental variables

But

$$\Pr[Z = 0|W = 1] - \Pr[Z = 0|W = 0]$$

$$= \Pr[Z = 1|W = 0] - \Pr[Z = 1|W = 1]$$

Therefore

$$\mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]$$

$$= (\Pr[Z = 1|W = 1] - \Pr[Z = 1|W = 0])$$

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

Instrumental variables

That is, provided $\Pr[Z = 1|W = 1] \neq \Pr[Z = 1|W = 0]$

$$\begin{aligned}\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0] \\ = \frac{\mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]}{\Pr[Z = 1|W = 1] - \Pr[Z = 1|W = 0]}\end{aligned}$$

or equivalently

$$\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0] = \frac{\mathbb{E}[Y|W = 1] - \mathbb{E}[Y|W = 0]}{\mathbb{E}[Z|W = 1] - \mathbb{E}[Z|W = 0]}.$$

As W is observed, we can deploy the moment-based estimator

$$\frac{\dfrac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i} - \dfrac{\sum_{i=1}^{n} (1 - W_i) Y_i}{\sum_{i=1}^{n} (1 - W_i)}}{\dfrac{\sum_{i=1}^{n} W_i Z_i}{\sum_{i=1}^{n} W_i} - \dfrac{\sum_{i=1}^{n} (1 - W_i) Z_i}{\sum_{i=1}^{n} (1 - W_i)}}$$

This is the binary analogue of the earlier regression-based ratio estimator.

Instrumental variables

In terms of potential outcomes, we can consider potential outcomes for treatment received

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

corresponding to two settings $\{0, 1\}$ of W, and potential outcomes for outcome

$$\{Y(w, z), w = 0, 1, z = 0, 1\}$$

corresponding to the four possible combinations of W and Z.

Instrumental variables

We have the "consistency" requirement relating the potential outcomes to the observed data:

$$Z = (1 - W)Z(0) + W\ Z(1)$$

and

$$Y = \sum_{z=0}^{1} \sum_{w=0}^{1} \mathbb{1}_{\{w\}}(W)\mathbb{1}_{\{z\}}(Z)Y(w, z).$$

Instrumental variables

We must have that for $z = 0, 1$

$$Y(1, z) = Y(0, z)$$

as $Y \perp\!\!\!\perp W | Z$ by assumption. We have *ignorability*

$$\{Y(w, z), Z(w) : \forall\ w, z\} \perp\!\!\!\perp Z.$$

Consequently

$$Y(0, 0) = Y(1, 0) \equiv Y(0)$$

$$Y(0, 1) = Y(1, 1) \equiv Y(1)$$

which simplifies the causal quantity of interest.

Instrumental variables

We are interested in the contrast

$$\mathbb{E}[Y(1) - Y(0)] \equiv \mathbb{E}[Y(W, 1) - Y(W, 0)].$$

We have that

$$\begin{aligned} Y &= (1 - Z)Y(0) + ZY(1) \\ &= Y(0) + (Y(1) - Y(0))Z \end{aligned}$$

which suggests the linear model for the observed data

$$Y_i = \beta_0 + \psi_0 Z_i + \varepsilon_i$$

Instrumental variables

We also need that $W \not\perp\!\!\!\perp Z$, that is

$$\mathbb{E}[Z(1)] \neq \mathbb{E}[Z(0)] \qquad \text{or} \qquad \Pr[Z(1) = 1] \neq \Pr[Z(0) = 1]$$

- Manipulation of W *does* change Z;
- From the assumptions and the DAG, we can state

$$\mathbb{E}[Z(1) - Z(0)] \equiv \mathbb{E}[Z|W = 1] - \mathbb{E}[Z|W = 0].$$

Instrumental variables

Recall the earlier formula

$$\mathbb{E}[Y|Z=1]-\mathbb{E}[Y|Z=0]=\frac{\mathbb{E}[Y|W=1]-\mathbb{E}[Y|W=0]}{\mathbb{E}[Z|W=1]-\mathbb{E}[Z|W=0]}.$$

We now have that the denominator on the right hand side is non-zero.

For the numerator, we have

$$\begin{aligned}\mathbb{E}[Y|W=1] &= \mathbb{E}[Y(0)+(Y(1)-Y(0))Z|W=1]\\ &= \mathbb{E}[Y(0)+(Y(1)-Y(0))Z(1)]\end{aligned}$$

$$\mathbb{E}[Y|W=0] = \mathbb{E}[Y(0)+(Y(1)-Y(0))Z(0)]$$

Instrumental variables

Thus, taking the difference we have

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

so therefore

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

We now try to understand the causal implications of this identity.

Instrumental variables

Stratum	$Z(1)$	$Z(0)$	Description
Compliers	1	0	*Follows* treatment protocol, *takes assigned treatment* w.
Defiers	0	1	*Defies* treatment protocol, takes *other* treatment $1 - w$.
Always Takers	1	1	*Ignores* treatment protocol, takes treatment 1.
Never Takers	0	0	*Ignores* treatment protocol, takes treatment 0.

Instrumental variables

In the *observed* data:

W	Z	Possible strata
1	0	Defier, Never Taker
0	1	Defier, Always Taker
1	1	Complier, Always Taker
0	0	Complier, Never Taker

We can never conclude with certainty to which stratum an individual belongs from the observed data.

Instrumental variables

It is common (and reasonable in most cases) to assume that

$$\Pr[Z(1) \geqslant Z(0)] = \Pr[Z(1) - Z(0) \geqslant 0] = 1$$

that is, being assigned $W = 1$ *cannot decrease* the value of $Z(\mathsf{w})$ compared with being assigned $W = 0$.

i.e. being allocated treatment increases or leaves unchanged the received treatment status variable.

This *monotonicity assumption* rules out the possibility of the *Defier* stratum.

Instrumental variables

Assuming $\Pr[Z(1) - Z(0) \geqslant 0] = 1$

W	Z	Possible strata
1	0	~~Defier~~, Never Taker
0	1	~~Defier~~, Always Taker
1	1	Complier, Always Taker
0	0	Complier, Never Taker

Instrumental variables

To get a further simplification in the expression

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

we use iterated expectation using the partitioning event

$$Z(1) > Z(0)$$

which is an event occurs if and only if

$$Z(1) = 1 \qquad \textit{and} \qquad Z(0) = 0.$$

Instrumental variables

We have in the numerator

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

$$= \mathbb{E}[(Y(1) - Y(0))(Z(1) - Z(0))|Z(1) > Z(0)]\Pr[Z(1) > Z(0)]$$

$$+ \mathbb{E}[(Y(1) - Y(0))(Z(1) - Z(0))|Z(1) \leqslant Z(0)]\Pr[Z(1) \leqslant Z(0)]$$

and similarly in the denominator

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

$$= \mathbb{E}[(Z(1) - Z(0))|Z(1) > Z(0)]\Pr[Z(1) > Z(0)]$$

$$+ \mathbb{E}[(Z(1) - Z(0))|Z(1) \leqslant Z(0)]\Pr[Z(1) \leqslant Z(0)]$$

Instrumental variables

Stratum	$Z(1)$	$Z(0)$	$Z(1) - Z(0)$
Compliers	1	0	1
Defiers	0	1	-1
Always Takers	1	1	0
Never Takers	0	0	0

Instrumental variables

We therefore have in the numerator:

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

$$= \mathbb{E}[(Y(1) - Y(0))|Z(1) > Z(0)]\Pr[Z(1) > Z(0)]$$

as only the contribution $Z(1) = 1$ and $Z(0) = 0$ remains.

Similarly in the denominator:

$$\mathbb{E}[(Z(1) - Z(0))] = \Pr[Z(1) > Z(0)]$$

Instrumental variables

Therefore

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

$$\equiv \mathbb{E}[(Y(1)-Y(0))|Z(1)>Z(0)].$$

It is therefore evident that the IV estimation procedure estimates the causal contrast

$$\mathbb{E}[(Y(1)-Y(0))|Z(1)>Z(0)]$$

Instrumental variables

As

$$Z(1) > Z(0) \quad \Longleftrightarrow \quad Z(1) = 1, Z(0) = 0$$

the quantity

$$\mathbb{E}[(Y(1) - Y(0))|Z(1) > Z(0)]$$

is termed the

- *complier average treatment effect* (CATE), or
- *complier average causal effect* (CACE), or
- *local average treatment effect* (LATE),

as $Z(1) > Z(0)$ identifies the *Complier* stratum.

Instrumental variables

Note

- If we observe $Z = 1$, the subject could be a Complier or Always Taker; we do not know who the Compliers are.
- The Compliers may be different depending on the instrument W chosen.
- In general the three quantities

$$\text{ATE} : \mathbb{E}[(Y(1) - Y(0))]$$

$$\text{ATT} : \mathbb{E}[(Y(1) - Y(0))|Z = 1]$$

$$\text{CATE} : \mathbb{E}[(Y(1) - Y(0))|Z(1) > Z(0)]$$

are not equal.

Note

- It is often in practice realistic to make the assumption

$$\Pr[Z = 1 | W = 0] = 0$$

which asserts that the Always Takers stratum is empty, and that all subjects for whom $Z = 1$ are Compliers.

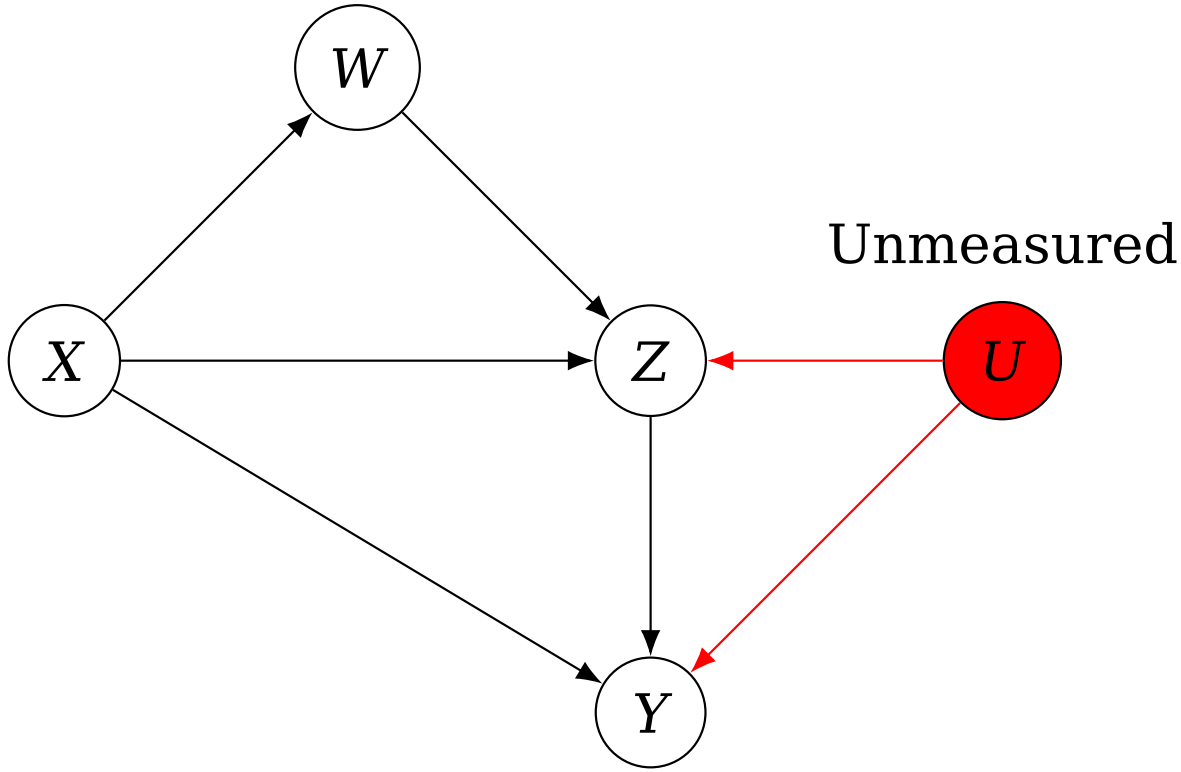
It follows that

$$\frac{\mathbb{E}[Y|Z = 1] - \mathbb{E}[Y|Z = 0]}{\Pr[Z = 1|W = 1]} = \mathbb{E}[(Y(1) - Y(0))|Z = 1]$$

which equates to the ATT.

Instrumental variables

Extension: with measured confounders



We have $U \perp\!\!\!\perp W$, and $Y \perp\!\!\!\perp W \mid Z, X$.

Instrumental variables

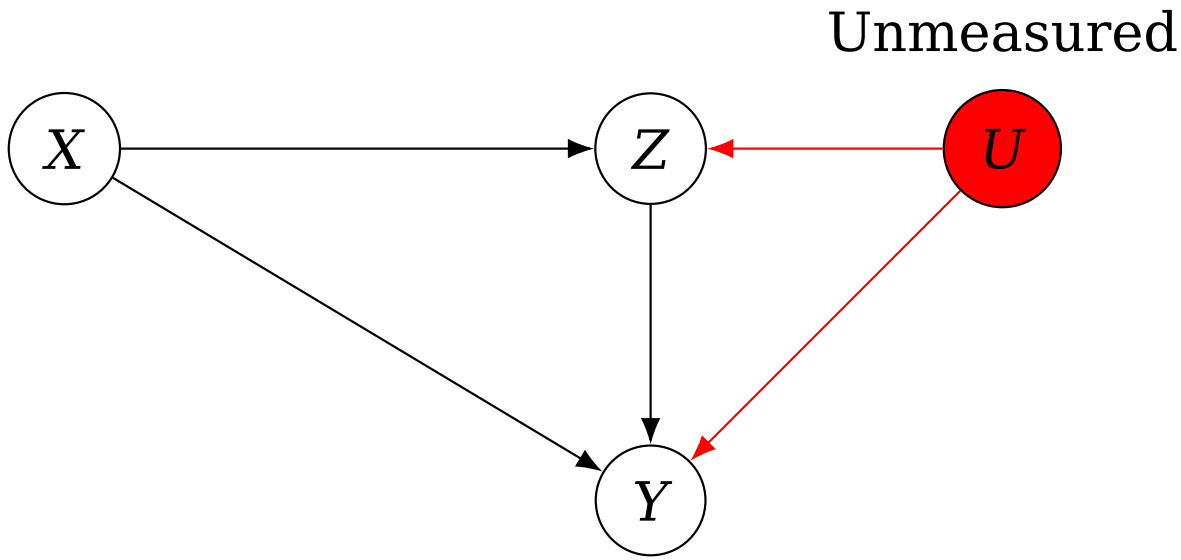
We can extend the previous logic by appealing to *conditional* ignorability

$$\{Y(\mathsf{w}, \mathsf{z}), Z(\mathsf{w}) : \forall\ \mathsf{w}, \mathsf{z}\} \perp\!\!\!\perp Z|X$$

and develop similar methods.

Instrumental variables

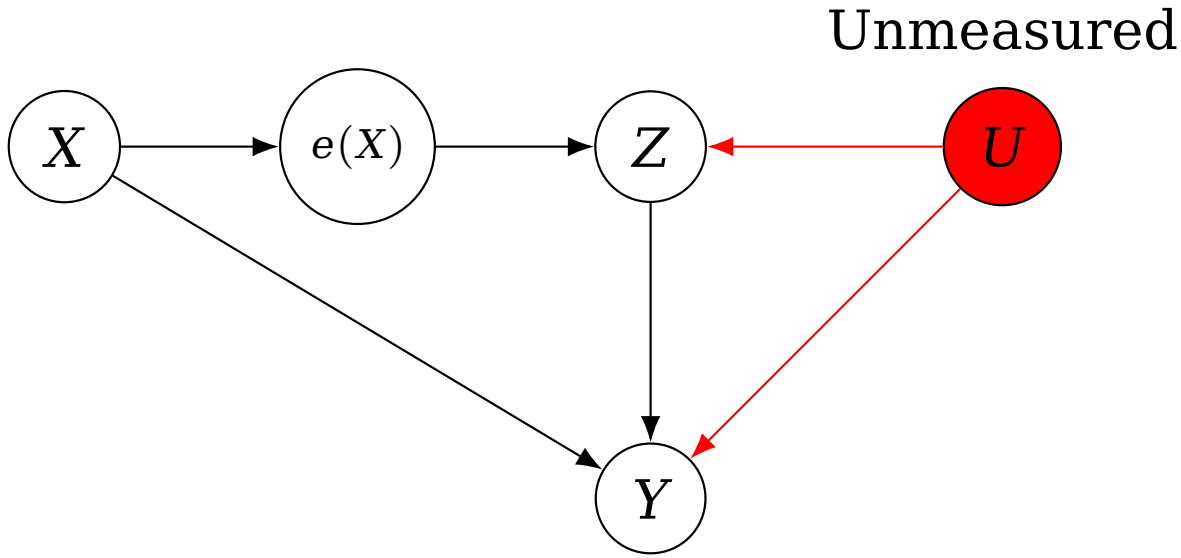
Suppose we have the following confounded system:



No adjustment method is feasible.

Instrumental variables

Including the propensity score, we have



This does not directly help with unmeasured confounding.

Instrumental variables

However, consider a variable W, with $\mathbb{E}[W] = 0$, such that

$$Z = e(X) + W$$

so that

$$W = Z - e(X) = Z - \mathbb{E}[Z|X]$$

We have

$$\mathbb{E}[W\ e(X)] = \mathbb{E}[(Z - e(X))e(X)] = \mathbb{E}[(e(X) - e(X))e(X)] = 0$$

by iterated expectation. Therefore W and $e(X)$ must be uncorrelated.

Instrumental variables

We also have that

$$\mathbb{E}[WZ] = \mathbb{E}[(Z - e(X))Z] = \mathbb{E}[Z - Ze(X)] = \mathbb{E}[e(X)(1 - e(X))]$$

as $Z^2 = Z$ (w.p. 1). Therefore as $\mathbb{E}[W] = 0$,

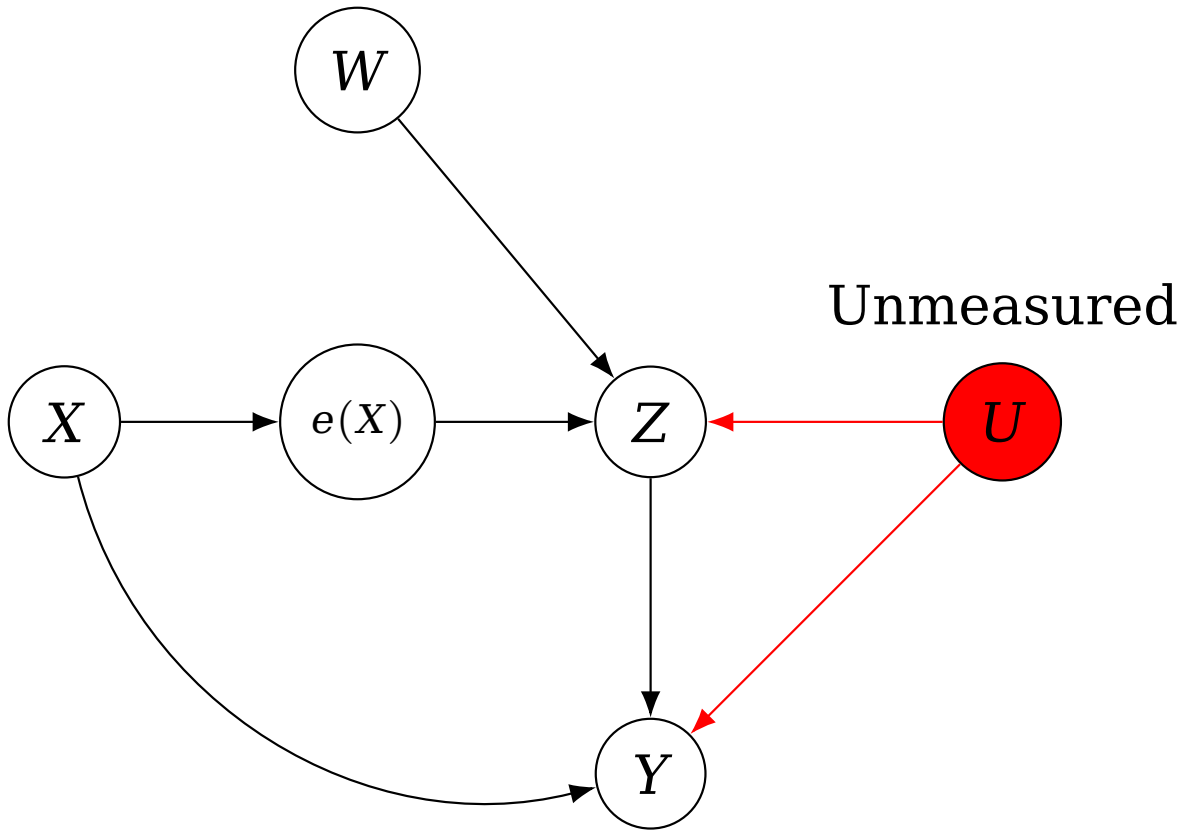
$$\text{Cov}[W, Z] = \mathbb{E}[e(X)(1 - e(X))] > 0$$

under the usual positivity assumption $0 < e(x) < 1$.

Hence $W \not\perp\!\!\!\perp Z$.

Instrumental variables

If we consider W a manipulable variable, we could propose the following DAG:



It is evident that $Y \perp\!\!\!\perp W \mid X, Z$, and also that $U \perp\!\!\!\perp W$.

Instrumental variables

Thus

$$W = Z - e(X) = Z - \mathbb{E}[Z|X].$$

is an instrumental variable. We can compute the ratio estimator based on the earlier identity

$$\psi_0 = \frac{\text{Cov}[W, Y]}{\text{Cov}[W, Z]}$$

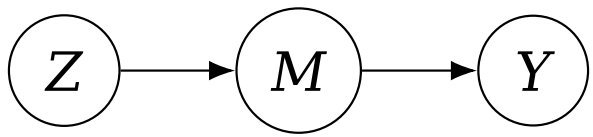
that is

$$\widehat{\psi}_0 = \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i Z_i} = \frac{\sum_{i=1}^{n} (Z_i - e(X_i)) Y_i}{\sum_{i=1}^{n} (Z_i - e(X_i)) Z_i}$$

which is identical to the *singly robust G-estimator*.

Mediation

Statistical methods for causal *mediation* analysis can also be formulated using potential outcomes:



where z and m will represent potential levels of treatment Z and mediator M respectively, with *potential outcome*

$$Y(\mathsf{z}, \mathsf{m})$$

and *potential mediator* value in light of $Z = \mathsf{z}$

$$M(\mathsf{z}).$$

Mediation

In general, both Z and M can take arbitrary values, but we most commonly consider the case when

- Z is binary $\{0, 1\}$
- M is
  - binary,
  - discrete, or
  - continuous.

Mediation

Then following Imai et al. (*Statistical Science,* 2010)

- the *total causal effect* (TCE) of treatment is based on

$$Y(z, M(z))$$

  i.e. set $Z = z$ and observe the consequence.

  For example

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

Mediation

- the *causal mediation effect* (CME) is based on

$$Y(z, M(z^*))$$

i.e. set $Z = z$ and observe the consequence of changing the mediator as if $Z = z^*$

For example

$$\delta(z) = \mathbb{E}[Y(z, M(1)) - Y(z, M(0))] \qquad z = 0, 1.$$

Mediation

- the *direct effect* (DE) of treatment is based on

$$Y(z^*, M(z))$$

i.e. set the mediator as if $Z = z$, and observe the consequence of changing treatment according to z^*.

For example

$$\zeta(z) = \mathbb{E}[Y(1, M(z)) - Y(0, M(z))] \qquad z = 0, 1.$$

Mediation

Terminology:

		Imai et al.	Pearl	Robins
Total effect	τ	ATCE	Total	Total
Indirect effect	$\delta(z)$	ACME	Natural	Pure/Total
Direct effect	$\zeta(z)$	ADE	Natural	Pure/Total

with

$$\text{Pure} : z = 0 \qquad \text{Total} : z = 1.$$

Mediation

That is,

$$\begin{aligned}
&\mathbb{E}[Y(1, M(1)) - Y(0, M(0))] \\
&\qquad = \mathbb{E}[Y(1, M(1)) - Y(1, M(0))] \\
&\qquad\quad + \mathbb{E}[Y(1, M(0)) - Y(0, M(0))]
\end{aligned}$$

so that

$$\tau = \delta(1) + \zeta(0) = \delta(0) + \zeta(1).$$

Mediation

We have in the observed data that

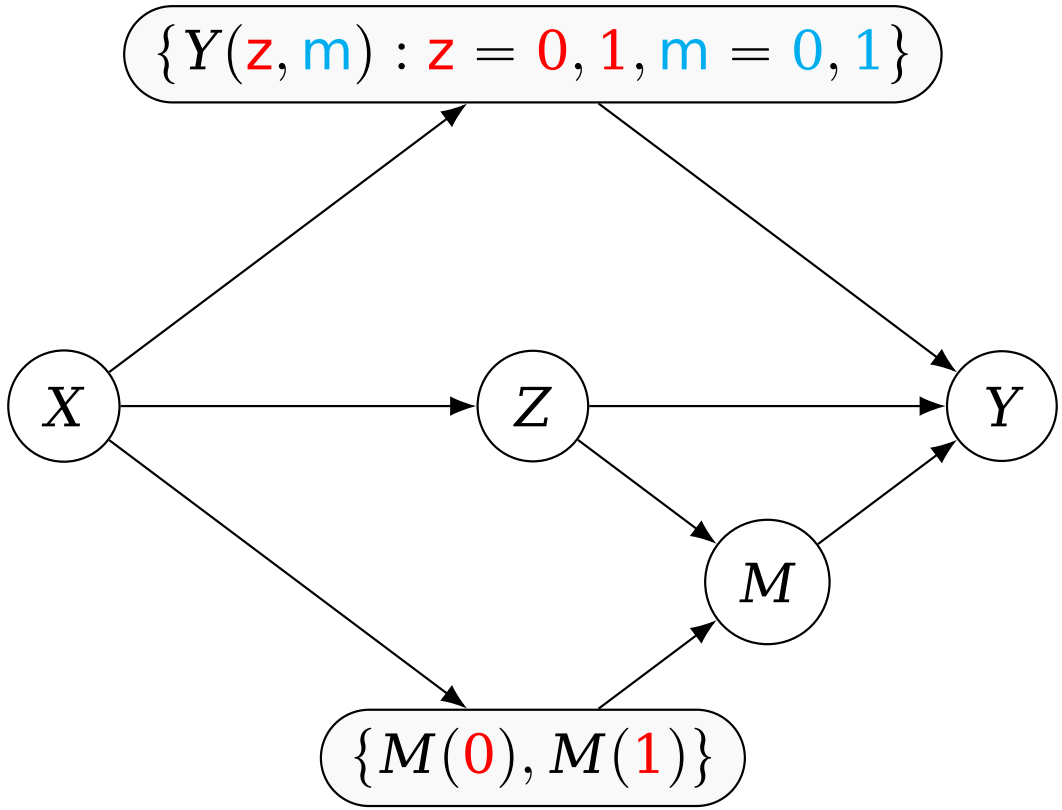
$$M = (1 - Z)M(0) + ZM(1)$$

and, if M is binary say,

$$\begin{aligned} Y = (1 - Z)(1 - M)\ Y(0, 0) \\ + (1 - Z)M\ Y(0, 1) \\ + Z(1 - M)\ Y(1, 0) \\ + Z\ M\ Y(1, 1) \end{aligned}$$

etc.

Mediation



Mediation

The DAG encompasses two ignorability assumptions

(i) $\{Y(\mathsf{z}^*, \mathsf{m}), M(\mathsf{z})\} \perp\!\!\!\perp Z \mid X = x;$

(ii) $Y(\mathsf{z}^*, \mathsf{m}) \perp\!\!\!\perp M(\mathsf{z}) \mid Z = \mathsf{z}, X = x;$

for all $\mathsf{z}, \mathsf{z}^*, \mathsf{m}$ and x, and relies upon two positivity assumptions

$$0 < \Pr[Z = z | X = x] < 1$$

and

$$0 < \Pr[M(z) = m | X = x, Z = z] < 1.$$

Mediation

Imai et al (2010) gives the following identities:

$$\delta(z) = \iiint y f_{Y|M,X,Z}(y|m,x,z) f_{M|X,Z}(m|x,1) f_X(x)\ dy\ dm\ dx$$
$$- \iiint y f_{Y|M,X,Z}(y|m,x,z) f_{M|X,Z}(m|x,0) f_X(x)\ dy\ dm\ dx$$

$$\zeta(z) = \iiint y f_{Y|M,X,Z}(y|m,x,1) f_{M|X,Z}(m|x,z) f_X(x)\ dy\ dm\ dx$$
$$- \iiint y f_{Y|M,X,Z}(y|m,x,0) f_{M|X,Z}(m|x,z) f_X(x)\ dy\ dm\ dx$$

Mediation

As for the ATT case, the challenge in estimating these quantities using moment-based methods is the mis-match in the terms

$$f_{Y|M,X,Z}(y|m,x,z)f_{M|X,Z}(m|x,z^*)$$

whenever $z \neq z^*$.

Mediation

Model-based estimation via the mean-model

$$\mu(m, x, z) = \int y \, f_{Y|M,X,Z}(y|m, x, z) \, dy$$

and the mediator model

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

is possible.

Longitudinal Data

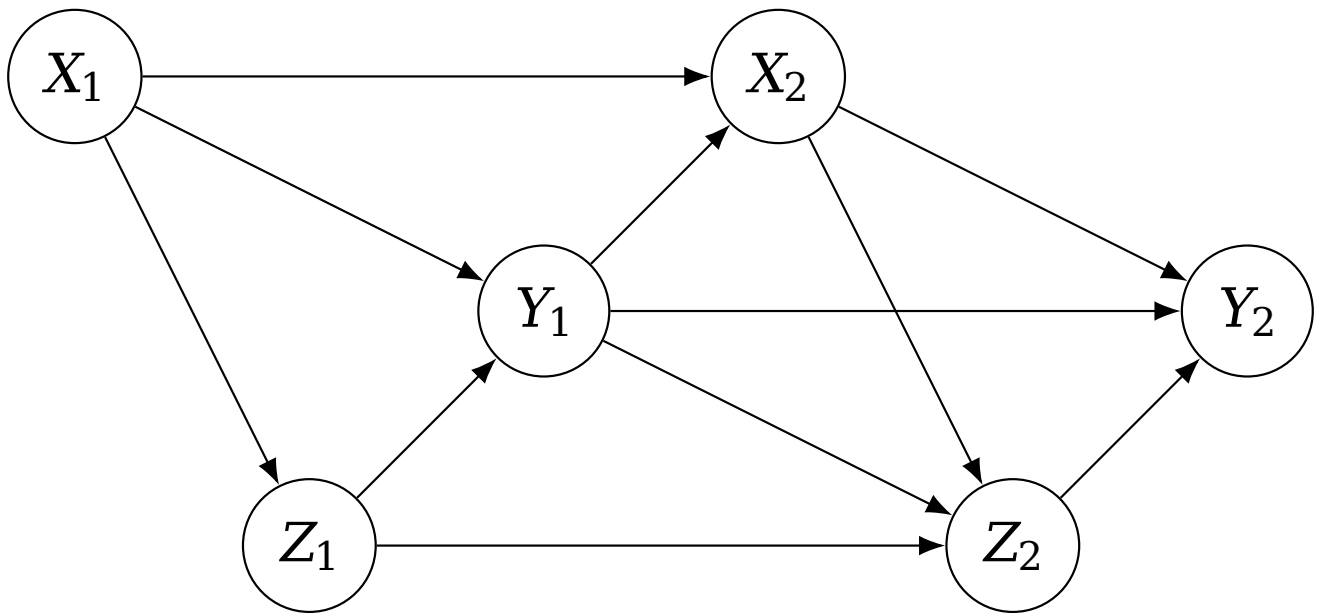
It is quite common to encounter longitudinal data with

- treatment
- confounders
- outcome

recorded over several time points for each individual.

Longitudinal Data

For example

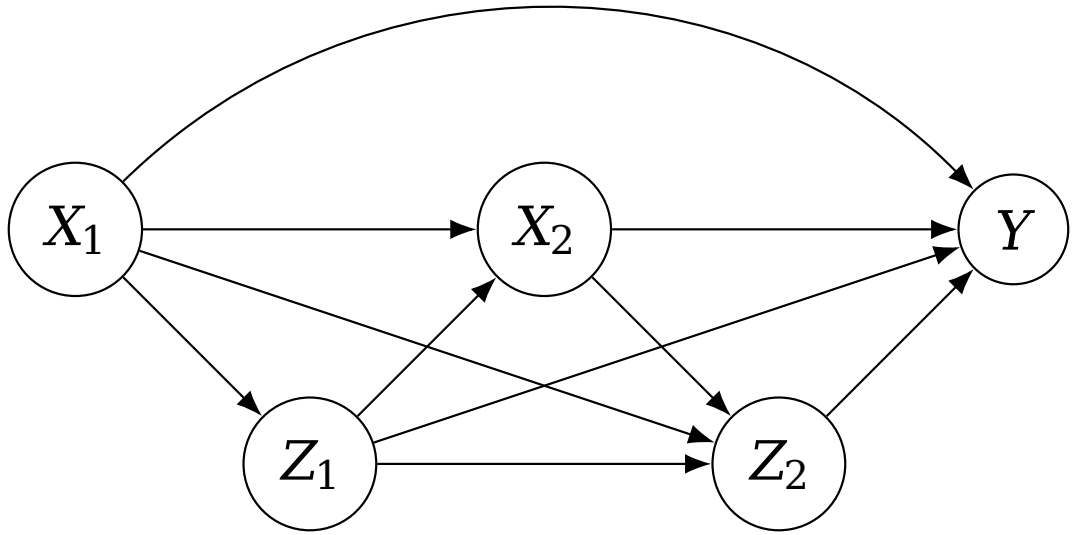


In such cases, and for the analysis of direct effects (within an interval) the methods already established can be utilized sequentially.

For *cumulative* (across interval) effects of treatment on a single terminal outcome, adjustment methods need careful application.

Longitudinal Data

For a two time point setting:



In this formulation, the time ordering

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

delimits the possible causal pathways.

Longitudinal Data

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

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

or equivalently

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

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

Longitudinal Data

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

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

Therefore, we cannot break the confounding by blocking paths by conditioning to get at the aggregate effect.

Marginal structural models

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

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

we may use the estimator

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

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

Marginal structural models

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

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

or, using the *total* treatment

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

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

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

Marginal structural models

That is for example

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

where

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

where

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

Marginal structural models

An alternative weight

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

where

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

is modelled could be used. This generalizes the earlier form of IPW estimator.

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

Marginal structural models

Using a *parametric* model,

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

where

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

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

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

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

Marginal structural models

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

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

for which the so-called *stabilized* weights

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

should be used.

Marginal structural models

Separate models are needed for numerator and denominator

- Denominator:

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

- Numerator:

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

parameterized by α and γ respectively, say.

That is, after cancelling terms,

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

This weight may be less extreme than the unstabilized counterpart.

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

Marginal structural models

Note

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

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

Optimal Dynamic Treatment Regimes

For treatment *sequence* (z_1, z_2), we seek the *optimal* treatment sequence $(z_1^{\text{opt}}, z_2^{\text{opt}})$ to *maximize* expected response.

- The objective is to establish a sequence of decision rules to be applied at each treatment interval;
- For interval $k = 1, 2$, the decision rule for interval k must utilize the information available only up to and including (the start of) that interval;
- The decision rule should be *hyperopic*, that is, prioritize *long-term* outcomes over short-term (*myopic*) outcomes;
- The rules will be learned from observational data, where optimal treatment is not necessarily guaranteed.

Optimal Dynamic Treatment Regimes

To solve this problem, a recursive approach is adopted

- optimal treatment sequence will be *personalized*, that is, dependent on an individual's characteristics;
- the computation will use *reverse* (or *Bellman*) optimization;
- first optimize Stage 2 treatment, and then optimize Stage 1 treatment *assuming optimal treatment at Stage 2*.

Optimal Dynamic Treatment Regimes

Decompose the expected (counterfactual) response as

$$\begin{aligned}\mathbb{E}[Y(z_1, z_2)] = \mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}})] \\ - \left\{\mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}}) - Y(z_1, z_2^{\text{opt}})]\right\} \\ - \left\{\mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(z_1, z_2)]\right\}\end{aligned}$$

This is the basis of the *Structural Nested Mean Model* (SNMM).

Optimal Dynamic Treatment Regimes

Further

$$\mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}}) - Y(z_1, z_2^{\text{opt}})] = \mathbb{E}[Y(z_1^{\text{opt}}, z_2^{\text{opt}}) - Y(0, z_2^{\text{opt}})] - \mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(0, z_2^{\text{opt}})]$$

$$\mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(z_1, z_2)] = \mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(z_1, 0)] - \mathbb{E}[Y(z_1, z_2) - Y(z_1, 0)]$$

Optimal Dynamic Treatment Regimes

We specify models

$$\mathbb{E}[Y(z_1, z_2) - Y(z_1, 0)] = (\mathbf{X}_2\psi_2)z_2 \qquad \text{Stage 2}$$

$$\mathbb{E}[Y(z_1, z_2^{\text{opt}}) - Y(0, z_2^{\text{opt}})] = (\mathbf{X}_1\psi_1)z_1 \qquad \text{Stage 1}$$

where

- $\mathbf{X}_2$ can depend on all data – including observed treatments – observed up to Stage 2.
- $\mathbf{X}_1$ can depend on all data observed up to Stage 1.

Optimal Dynamic Treatment Regimes

1. Use *G-estimation* on Y conditioning on $(\mathbf{x}_1, z_1, \mathbf{x}_2, z_2)$ at second stage to estimate ψ_2 using a proposed mean model

$$Y = \mathbf{X}_{21}\beta_2 + z_2\mathbf{X}_{22}\psi_2 + \varepsilon$$

2. For each individual infer the optimal Stage 2 treatment

$$z_2^{\text{opt}} = \mathbb{1}\{\mathbf{X}_{22}\widehat{\psi}_2 > 0\}$$

3. Form *pseudo-outcome* Y_1

$$Y_1 = Y - (\mathbf{X}_{22}\widehat{\psi}_2)(z_2^{\text{opt}} - z_2)$$

Optimal Dynamic Treatment Regimes

4. Use G-estimation on Y_1 conditioning on $(\mathbf{x}_1, z_1)$ at first stage to estimate ψ_1 using a proposed mean model

$$Y_1 = \mathbf{X}_{11}\beta_1 + z_1\mathbf{X}_{12}\psi_1 + \varepsilon$$

5. For each individual infer the optimal Stage 1 treatment

$$z_1^{\text{opt}} = \mathbb{1}\{\mathbf{X}_{11}\widehat{\psi}_1 > 0\}$$

The method is robust to mis-specification of the *nuisance mean model*, provided the treatment model is correctly specified.

Optimal Dynamic Treatment Regimes

Can infer optimized potential outcome

$$Y + (\mathbf{X}_{12}\widehat{\psi}_1)(z_1^{\text{opt}} - z_1) + (\mathbf{X}_{22}\widehat{\psi}_2)(z_2^{\text{opt}} - z_2)$$

which takes the observed outcome Y and adds

- the additional benefit of optimal treatment at stage 1

$$(\mathbf{X}_{12}\widehat{\psi}_1)(z_1^{\text{opt}} - z_1)$$

- the additional benefit of optimal treatment at stage 2

$$(\mathbf{X}_{22}\widehat{\psi}_2)(z_2^{\text{opt}} - z_2).$$

Causal discovery

Many of the adjustment methods rely upon the construction of estimators assuming knowledge of

- confounder structure
- time-ordering
- propensity score construction

all of which depend on knowledge of the data generating DAG.

In many applications, the DAG is *unknown*.

Causal discovery

There are several approaches to *causal discovery*, that is, learning the DAG from the observed data

- PC algorithm: `pcalg` in `R`
- SGS algorithm

Both rely on recursive identification of conditional independencies, identification of colliders and edge orientation procedures based on statistical tests.

Can be effective, but typically need large sample sizes.