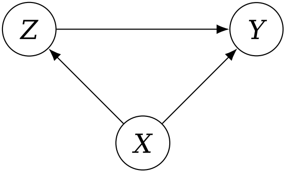
Part 1
Causal Adjustment Methods

Causal Adjustment Methods

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



Model-based estimation will be successful if

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

is correctly specified.

Causal Adjustment Methods

Recall that, under the previous assumptions

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

that is, the treatment data are ignored.

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

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

Causal Adjustment Methods

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

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

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

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

How can we break the dependence ?

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

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

Then (trivially)

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

and

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

Matching

Now suppose X takes values on the finite set

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

with

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

determining the distribution of X.

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

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

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

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

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

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

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

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

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

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

where

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

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

Matching

This is a *matching* estimator:

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

Note

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

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

Matching

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

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

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

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

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

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

Matching

Note

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

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

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

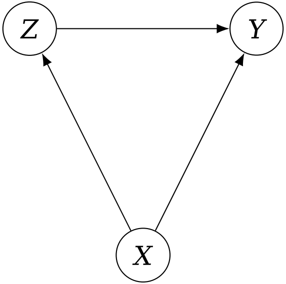
which then leads to an estimator

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

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

Constructing Balance

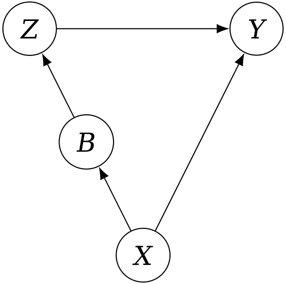
Consider the basic confounding set up:



We know that conditioning on X blocks the confounding path.

Constructing Balance

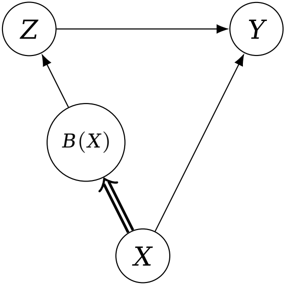
Now suppose we could find a new variable B so that



Conditioning on B also *blocks* the confounding path.

Constructing Balance

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



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

Constructing Balance

Note that

- if we can find such a B, then

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

- there is no arrow from B to Y, and

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

  as conditioning on X and Z blocks the open paths from B to Y.

Constructing Balance

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

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

Constructing Balance

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

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

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

provided the conditional densities are well-defined.

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

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

Constructing Balance

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

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

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

That is,

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

Constructing Balance

Similarly, for any proposed B, we must also have

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

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

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

We must now ensure that

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

Constructing Balance

We can do this by requiring

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

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

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

Constructing Balance

Define the function

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

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

$$q(b) = b$$

so that

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

Constructing Balance

Hence by construction

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

that is, if we consider the 'contour'

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

then

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

Thus

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

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

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

$$f_{Z|X}(1|x) = \Pr[Z = 1|X = x] = \frac{\exp\{x_1 + 2x_2 x_3^2\}}{1 + \exp\{x_1 + 2x_2 x_3^2\}} \equiv B(x).$$

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

Constructing Balance

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

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

by Rosenbaum and Rubin (1983).

Constructing Balance

We have for the joint pdf

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

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

as

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

and

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

Constructing Balance

Thus if

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

then for all (z, b)

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

Constructing Balance

Therefore

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

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

as the b terms cancel. Thus

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

Constructing Balance

Note

Note that for the binary case

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

Constructing Balance

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

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

$$B^* = b^* \implies B = b.$$

Constructing Balance

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

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

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

Constructing Balance

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

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

Consider two values x_1 and x_2. We have that

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

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

Constructing Balance

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

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

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

But as for all x

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

this implies that

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

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

Constructing Balance

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

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

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

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

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

Then directly

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

Constructing Balance

Example:

Suppose

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

Then define

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

so that

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

which does not depend on (x_1, x_2).

For the binary case

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

which suggests another possible balancing score construction involves inspection of

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

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

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

Constructing Balance

Note

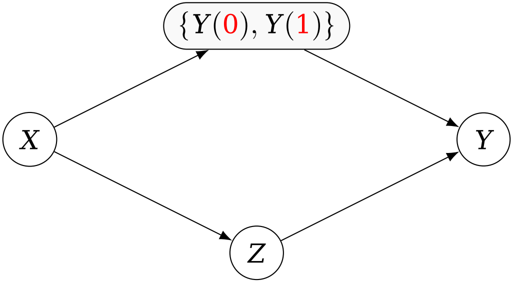
For the moment, we will assume that $e(X)$ or $B(X)$ is *known precisely*.

In practice, we will typically have to

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

Constructing Balance

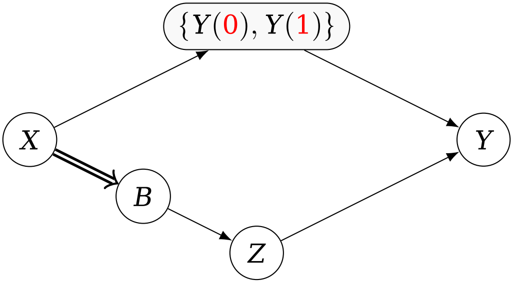
Recall the assumption of *strong ignorability*:



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

Constructing Balance

This can be considered in the balancing score case:



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

Adjustment via the Propensity Score

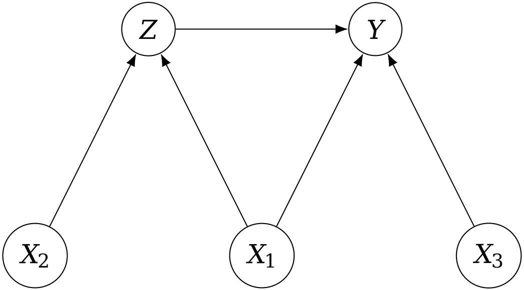
Adjustment methods based on balancing scores can be developed; the balancing score is used to block backdoor (confounding) paths.

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

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

Adjustment via the Propensity Score

The basic set up we consider is the following:



Adjustment via the Propensity Score

There are three types of covariate:

X_1 *confounders*

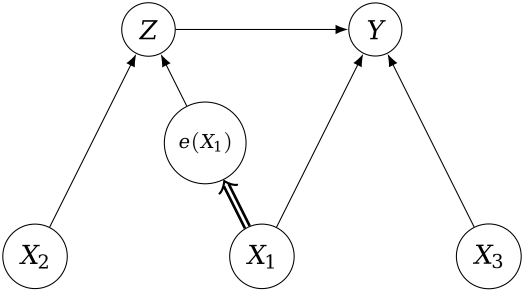
X_2 *instruments* (pure predictors of treatment)

X_3 *pure predictors of outcome*

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

Adjustment via the Propensity Score

Including the propensity score retains the previous feature:



Note

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

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

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

and

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

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

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

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

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

Stratification and Matching

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

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

for different values of z.

For fixed value e_1, let

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

and define a 'local' APO estimator as

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

Stratification and Matching

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

Consider a partition constructed using

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

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

We can define local estimator

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

and global estimator

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

Note

To construct such estimators of the ATE, say

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

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

This is termed an *overlap* condition.

Stratification and Matching

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

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

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

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

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

  for some constant c.

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

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

  is *minimized*.

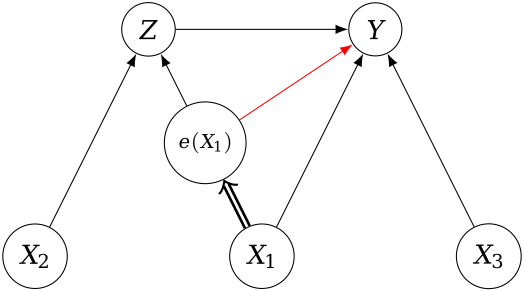
Stratification and Matching

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

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

Propensity Score Regression

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



Propensity Score Regression

We may consider the regression model

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

which, as

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

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

Propensity Score Regression

Example:

Suppose that we have the following data generating model:

- Confounders: $(X_1, X_2)^\top \sim \mathit{Normal}_2((1,1)^\top, \Sigma)$ with

$$\Sigma = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.5 \end{bmatrix} \begin{bmatrix} 1.0 & 0.8 \\ 0.8 & 1.0 \end{bmatrix} \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.5 \end{bmatrix}$$

- Treatment: $Z|X_1, X_2 \sim \mathit{Bernoulli}(e(X_1, X_2))$, where

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

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

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

Propensity Score Regression

Example:

We consider fitting the parametric model

$$m(x, z; \beta, \psi) = (\beta_0 + \beta_1 x_1) + z\psi_0$$

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

Propensity Score Regression

```
#n=1000
#Correct specification
> round(coef(summary(lm(Y~X1+X2+X1:X2+Z))),4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.7134     0.6222  4.3608   0.0000
X1            2.2156     0.6869  3.2254   0.0013
X2            0.2882     0.4807  0.5996   0.5489
Z             1.0150     0.0674 15.0572   0.0000
X1:X2         1.7421     0.4721  3.6905   0.0002
```

```
#Incorrect specification
> round(coef(summary(lm(Y~X1+Z))),4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  -4.2613     0.4034 -10.5631        0
X1           11.4990     0.3888  29.5760        0
Z             0.6366     0.0762   8.3523        0
```

Propensity Score Regression

Example:

In the correctly specified model, we have

$$\widehat{\psi}_0 \ : \ 1.0150 \ (0.0674)$$

however in the incorrectly specified model we have

$$\widehat{\psi}_0 \ : \ 0.6366 \ (0.0762)$$

This effect persists at even larger sample sizes.

Propensity Score Regression

Example:

Now consider fitting the parametric model

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

which considers the additional final term that depends on the propensity score.

Initially, we will set

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

that is, using the true value.

Propensity Score Regression

```
#Propensity score regression
> round(coef(summary(lm(Y~X1+Z+eX))),4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   4.1718     0.5609   7.4377        0
X1            5.1662     0.4701  10.9907        0
Z             1.0172     0.0682  14.9069        0
eX           -4.6374     0.2430 -19.0815        0
```

Propensity Score Regression

Example:

We now have

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

and so correct estimation of ψ_0 has been recovered.

Propensity Score Regression

Example:

Now suppose

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

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

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

Propensity Score Regression

```
#n=1000
#Correct specification
> round(coef(summary(lm(Y~X1+X2+X1:X2+Z+Z:X1+Z:X2))),4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.5674     0.9486  3.7609   0.0002
X1            1.2812     1.0109  1.2675   0.2053
X2            0.1155     0.6004  0.1923   0.8475
Z            -0.2023     0.9313 -0.2173   0.8281
X1:X2         1.9903     0.5672  3.5090   0.0005
X1:Z          2.3744     1.0558  2.2488   0.0247
X2:Z          0.8420     0.2091  4.0272   0.0001
```

Propensity Score Regression

```
#Incorrect specification
> round(coef(summary(lm(Y~X1+Z+Z:X1+Z:X2))),4)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.4874     0.5541 -8.0981        0
X1           11.7187     0.5363 21.8503        0
Z             6.4906     0.8778  7.3941        0
X1:Z         -6.5766     0.9644 -6.8196        0
Z:X2          2.8785     0.1642 17.5344        0
```

Propensity Score Regression

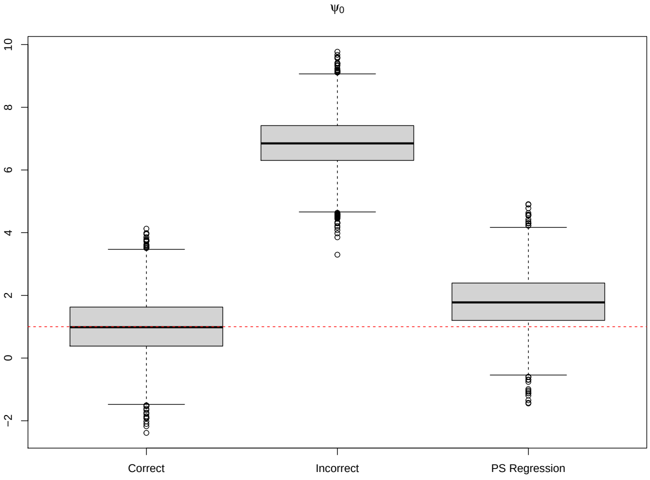
```
#Propensity score regression
> round(coef(summary(lm(Y~X1+Z+Z:X1+Z:X2+eX))),4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   3.9848     0.7752   5.1403   0.0000
X1            5.4002     0.6565   8.2252   0.0000
Z             1.4774     0.8716   1.6951   0.0904
eX           -4.7679     0.3313 -14.3913   0.0000
X1:Z          0.6533     1.0113   0.6460   0.5184
Z:X2          0.8889     0.2036   4.3664   0.0000
```

Example:

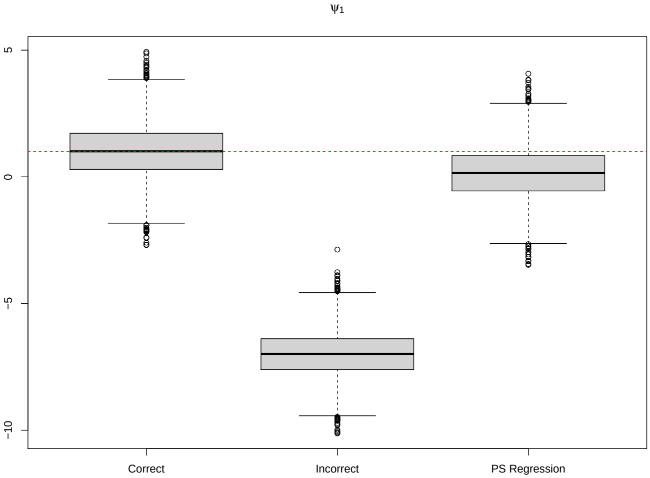
Hard to conclude anything due to the inherent variability, but it seems that including the propensity score does improve the estimation of (ψ_0, ψ_1, ψ_2).

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

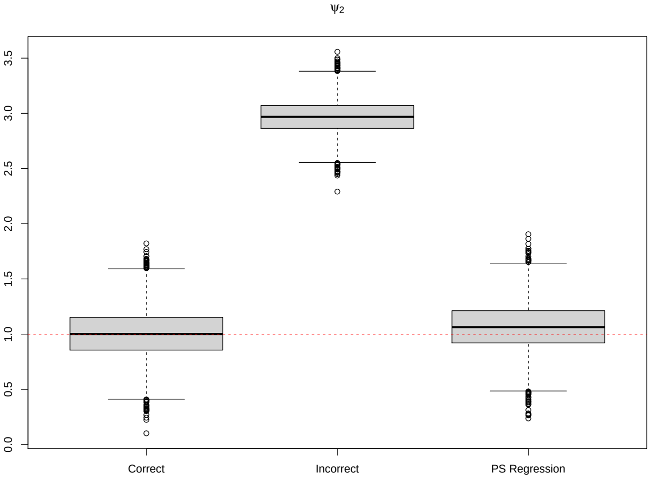
Propensity Score Regression



Propensity Score Regression



Propensity Score Regression



Example:

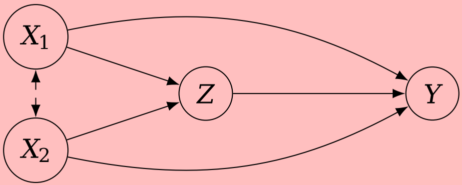
This confirms that including the propensity score *does* improve the estimation of (ψ_0, ψ_1, ψ_2), even if the treatment-free model component is incorrectly specified.

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

Propensity Score Regression

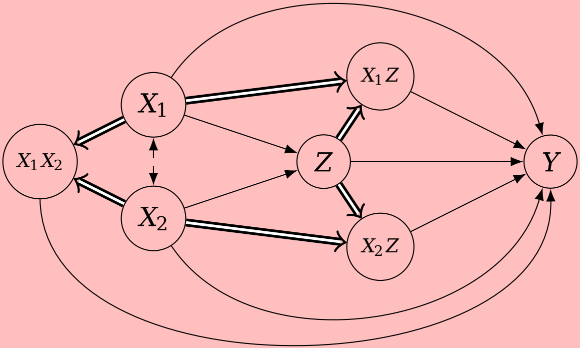
Example:

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



Example:

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



Example:

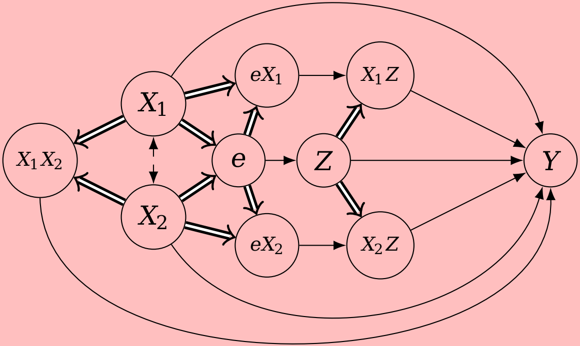
We need to block the open paths via the interactions. This can be achieved by using the model

$$m(x, z; \beta, \psi, \phi) = (\beta_0 + \beta_1 x_1) + z(\psi_0 + \psi_1 x_1 + \psi_2 x_2) + e(x_1, x_2)(\phi_0 + \phi_1 x_1 + \phi_2 x_2)$$

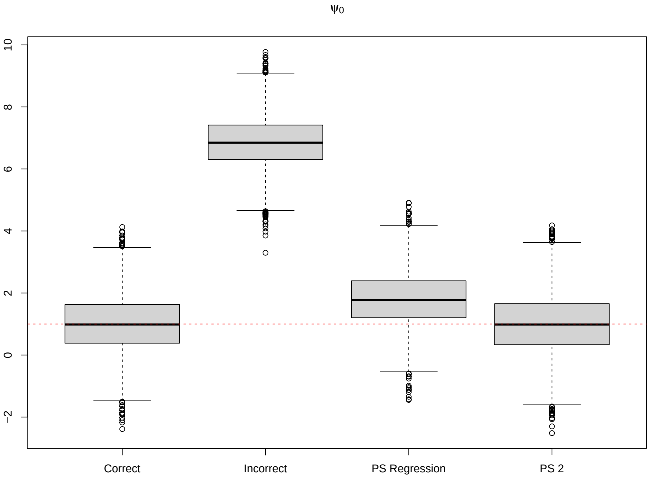
Conditioning on $e(X), e(X)X_1$ and $e(X)X_2$ blocks the paths.

Propensity Score Regression

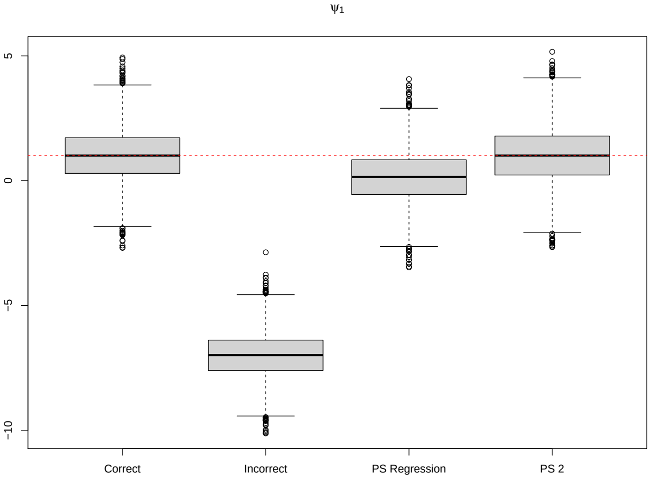
Example:



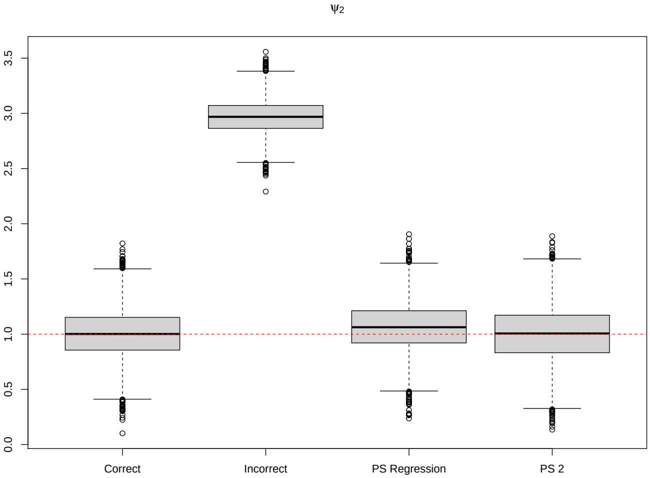
Propensity Score Regression



Propensity Score Regression



Propensity Score Regression



Example:

The augmented propensity score regression model (PS 2) improves the performance.

Note, however, that the variances of the estimators from propensity score regression model are slightly *larger* than those arising from the correctly specified model.

- 10% to 20% larger in this simulation.

Propensity Score Regression

Example:

In this analysis, we may estimate the ATE by taking the average difference of the two fitted values under the proposed model, that is

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

Example:

Note, however that we need to take care in estimating the APO. In the data generating model, with

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

we have that

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

This cannot in general be estimated correctly using

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

Propensity Score Regression

Note

This type of adjustment works for a *linear* outcome model; however, for other types of model such as

- log-linear
- logistic

more care needs to be taken.

Inverse Probability Weighting

We have from a previous result that

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

and also that

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

Inverse Probability Weighting

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

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

Inverse Probability Weighting

That is

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

or equivalently

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

Inverse Probability Weighting

We therefore have the estimator

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

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

Inverse Probability Weighting

We may also write

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

where

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

is a weight, where

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

In the binary case, we have

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

where

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

by iterated expectation.

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

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

is

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

Inverse Probability Weighting

Note

The two estimators solve two slightly different estimating equations:

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

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

  i.e. reweights the datum Y_i.

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

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

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

Inverse Probability Weighting

Note

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

$$Y_i^* = \left(\frac{Z_i}{e(X_i)} + \frac{1 - Z_i}{1 - e(X_i)} \right) Y_i$$

and

$$X_i^* = \left(\frac{Z_i}{e(X_i)} + \frac{1 - Z_i}{1 - e(X_i)} \right) X_i$$

which represent a sample from a pseudo-population in which

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

The new data set does not suffer from confounding.

Inverse Probability Weighting

Note

- $\widehat{\mu}_{\text{IPW}}(\mathsf{z})$ and $\widetilde{\mu}_{\text{IPW}}(\mathsf{z})$ are *unbiased* estimators of $\mu(\mathsf{z})$ by construction.
- In these estimators
$$f^{\mathcal{O}}_{Z|X}(Z|X)$$
plays a critical role; this is the function that determines the propensity score.
- There is an important requirement that
$$f^{\mathcal{O}}_{Z|X}(z|x) > 0$$
for any (x, z). This is termed a *positivity* requirement.

Inverse Probability Weighting

Note

- Positivity requires that for all z under consideration

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

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

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

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

Augmentation

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

Suppose that we have knowledge of the conditional model

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

We could use this model for *outcome regression*.

However, note that

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

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

Now under the standard assumption, we can write

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

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

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

as

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

Augmentation

This suggests the alternative moment-based estimator

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

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

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

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

where as before

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

Note

Note that

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

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

Note also that

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

identical to the outcome regression estimator.

Augmentation

Note

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

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

and

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

Augmentation

Note

However, note that

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

and

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

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

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

Augmentation

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

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

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

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

Write for the expectation of the second term

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

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

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

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

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

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

and hence we have

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

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

Augmentation

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

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

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

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

$$m(x, z)$$

or

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

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

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

Augmentation

Note

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

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

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

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

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

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

so that

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

and that

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

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

Augmentation

Note that

$$\begin{aligned}\mathrm{Var}[RY + (1-R)\mu(X, \mathsf{z})] \\ = \mathrm{Var}[RY] + \mathrm{Var}[(1-R)\mu(X, \mathsf{z})] + 2\mathrm{Cov}[RY, (1-R)\mu(X, \mathsf{z})].\end{aligned}$$

Fr the second term

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

as by iterated expectation

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

Augmentation

Similarly by iterated expectation

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

Therefore

$$\begin{aligned}
&\mathrm{Var}[RY + (1-R)\mu(X, \mathsf{z})] \\
&\quad = \mathrm{Var}[RY] + \mathbb{E}[(1-R)^2\{\mu(X, \mathsf{z})\}^2] + 2\mathbb{E}[R(1-R)\{\mu(X, \mathsf{z})\}^2] \\
&\quad = \mathrm{Var}[RY] + \mathbb{E}[(1-R^2)\{\mu(X, \mathsf{z})\}^2]
\end{aligned}$$

Augmentation

However

$$
\begin{aligned}
\mathbb{E}_{R|X}[(1-R^2)\{\mu(X,\mathsf{z})\}^2] &= \{\mu(X,\mathsf{z})\}^2\mathbb{E}_{R|X}[(1-R^2) \mid X] \\
&= \{\mu(X,\mathsf{z})\}^2\mathbb{E}_{Z|X}\left[1-\left(\frac{\mathbb{1}_{\{\mathsf{z}\}}(Z)}{f^{\mathcal{O}}_{Z|X}(Z|X)}\right)^2\middle|X\right] \\
&= \{\mu(X,\mathsf{z})\}^2\mathbb{E}_{Z|X}\left[1-\frac{\mathbb{1}_{\{\mathsf{z}\}}(Z)}{\left(f^{\mathcal{O}}_{Z|X}(Z|X)\right)^2}\middle|X\right] \\
&= \{\mu(X,\mathsf{z})\}^2\left(1-\frac{1}{f^{\mathcal{O}}_{Z|X}(\mathsf{z}|X)}\right) \\
&\leqslant 0 \qquad \text{(w.p. 1.)}
\end{aligned}
$$

Augmentation

Therefore

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

and hence

$$\mathrm{Var}[\widetilde{\mu}_{\text{AIPW}}] \leqslant \mathrm{Var}[\widetilde{\mu}_{\text{IPW}}].$$

Similar result for $\widehat{\mu}(\mathsf{z})$.

Augmentation

Note

This variance result *can* hold if $\mu(x,z)$ is mis-specified as the same argument follows for any estimator

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

We have that

$$\begin{aligned}\mathrm{Var}[RY + (1-R)m(X,z)] = \ & \mathrm{Var}[RY] + \mathbb{E}[(1-R)^2\{m(X,z)\}^2] \\ & + 2\mathbb{E}[R(1-R)\mu(X,z)m(X,z)]\end{aligned}$$

Augmentation

Note

Thus we get variance reduction over the IPW estimator if

$$\mathbb{E}[(1-R)^2\{m(X,\mathsf{z})\}^2] \geqslant -2\mathbb{E}[R(1-R)\mu(X,\mathsf{z})m(X,\mathsf{z})]$$

which will hold if $m(X,\mathsf{z})$ and $\mu(X,\mathsf{z})$ are sufficiently positively correlated.

Note

If *neither* of the models

$$\mu(x, z) \qquad e(x) = f^{\mathcal{O}}_{Z|X}(1|x)$$

is correctly specified, then the AIPW estimator is *biased*. If we instead use

$$m(x, z) \qquad g(x)$$

for these two models, the expectation of $\widetilde{\mu}(1)$ is

$$\mathbb{E}^{\mathcal{O}}_{X}\left[\frac{e(X)}{g(X)}(\mu(X, 1) - m(X, 1))\right] + \mathbb{E}^{\mathcal{O}}_{X}\left[m(X, 1)\right]$$

Augmentation

Note

The bias is therefore

$$\mathbb{E}_X^{\mathcal{O}}\left[\left(\frac{e(X)}{g(X)} - 1\right)(\mu(X, 1) - m(X, 1))\right]$$

Augmentation

Note

In the inverse weighting estimators, it has been assumed that the model

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

is known *precisely*. This can be replaced by a parametric model

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

with α then estimated using maximum likelihood or other methods. The IPW estimators then proceed using the fitted values

$$f^{\mathcal{O}}_{Z|X}(z|x;\widehat{\alpha}).$$

AIPW via regression

Consider the binary treatment case, and the model

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y \mid X = x, Z = z] = \mu(x, z) + \phi_0 \frac{1 - z}{1 - e(x)} + \phi_1 \frac{z}{e(x)}$$

for parameters ϕ_0 and ϕ_1. We consider estimating these parameters using ordinary least squares. Let

$$R_{0i} = \frac{1 - Z_i}{1 - e(X_i)} \qquad R_{1i} = \frac{Z_i}{e(X_i)}$$

with corresponding observed values r_{0i} and r_{1i}.

AIPW via regression

The OLS score equations are

$$\frac{\partial}{\partial \phi_0} \quad : \quad -2\sum_{i=1}^{n} r_{0i}(y_i - \mu(x, z) - \phi_0 r_{0i} - \phi_1 r_{1i}) = 0$$

$$\frac{\partial}{\partial \phi_1} \quad : \quad -2\sum_{i=1}^{n} r_{1i}(y_i - \mu(x_i, z_i) - \phi_0 r_{0i} - \phi_1 r_{1i})$$

and we may solve these directly to obtain

$$\widehat{\phi}_z = \frac{\sum_{i=1}^{n} r_{zi}(y_i - \mu(x_i, z_i)))}{\sum_{i=1}^{n} r_{zi}^2} \qquad z = 0, 1.$$

AIPW via regression

Predictions from this fit for the $z = 0, 1$ cases are

$$\mu(x_i, 0) + \widehat{\phi}_0 \frac{1}{1 - e(x_i)} \qquad \mu(x_i, 1) + \widehat{\phi}_1 \frac{1}{e(x_i)}$$

respectively.

To obtain an estimates of $\mu(0)$ and $\mu(1)$, we consider

$$\widetilde{\mu}_{\text{AOR}}(0) = \frac{1}{n} \sum_{i=1}^{n} \left(\mu(x_i, 0) + \widehat{\phi}_0 \frac{1}{1 - e(x_i)} \right).$$

and

$$\widetilde{\mu}_{\text{AOR}}(1) = \frac{1}{n} \sum_{i=1}^{n} \left(\mu(x_i, 1) + \widehat{\phi}_1 \frac{1}{e(x_i)} \right).$$

AIPW via regression

Plugging in the estimates of the ϕs, we obtain

$$\widetilde{\mu}_{\text{AOR}}(0) = \frac{1}{n}\sum_{i=1}^{n}\mu(x_i, 0) + \frac{\sum_{i=1}^{n}\frac{1}{1-e(x_i)}}{\sum_{i=1}^{n} r_{0i}^2}\frac{1}{n}\sum_{i=1}^{n} r_{0i}(y_i - \mu(x_i, 0))$$

and

$$\widetilde{\mu}_{\text{AOR}}(1) = \frac{1}{n}\sum_{i=1}^{n}\mu(x_i, 1) + \frac{\sum_{i=1}^{n}\frac{1}{e(x_i)}}{\sum_{i=1}^{n} r_{1i}^2}\frac{1}{n}\sum_{i=1}^{n} r_{1i}(y_i - \mu(x_i, 1)).$$

AIPW via regression

Now

$$\frac{\displaystyle\sum_{i=1}^{n}\frac{1}{e(X_i)}}{\displaystyle\sum_{i=1}^{n}R_{1i}^2}=\frac{\displaystyle\frac{1}{n}\sum_{i=1}^{n}\frac{1}{e(X_i)}}{\displaystyle\frac{1}{n}\sum_{i=1}^{n}R_{1i}^2}\xrightarrow{p}\frac{\mathbb{E}_X^{\mathcal{O}}\left[\dfrac{1}{e(X)}\right]}{\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\dfrac{Z^2}{\{e(X)\}^2}\right]}=1.$$

as

$$\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\frac{Z^2}{\{e(X)\}^2}\right]\equiv\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\frac{Z}{\{e(X)\}^2}\right]=\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)}\right]$$

AIPW via regression

Therefore

$$\widetilde{\mu}_{\text{AOR}}(1) = \frac{1}{n}\sum_{i=1}^{n}\mu(X_i, 1) + \frac{1}{n}\sum_{i=1}^{n}R_{1i}(Y_i - \mu(X_i, 1)) + o_p(1)$$

$$= \widetilde{\mu}_{\text{AIPW}}(1) + o_p(1).$$

Similarly

$$\widetilde{\mu}_{\text{AOR}}(0) = \widetilde{\mu}_{\text{AIPW}}(0) + o_p(1)$$

This approach to IPW estimation is known as *augmented outcome regression* (AOR).

AIPW via regression

To estimate the ATE

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

we can also use augmented outcome regression based on the mean model

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

and take the difference between the fitted values to obtain the estimate, $\widehat{\delta}_{\text{AOR}}$.

AIPW via regression

Note

The variance of the estimator $\tilde{\mu}(1)$ is given by

$$\frac{1}{n}\mathrm{Var}^{\mathcal{O}}_{X,Y,Z}\left[\frac{ZY}{e(X)}\right]$$

and under the correct specification of $e(x)$, we have

$$\mathrm{Var}^{\mathcal{O}}_{X,Y,Z}\left[\frac{ZY}{e(X)}\right] = \mathbb{E}^{\mathcal{O}}_{X,Y,Z}\left[\frac{Z^2Y^2}{\{e(X)\}^2}\right] - \{\mu(1)\}^2$$

AIPW via regression

Note

Now

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}\left[\frac{Z^2Y^2}{\{e(X)\}^2}\right] = \mathbb{E}^{\mathcal{O}}_{X,Z}\left[\frac{Zv(X,Z)}{\{e(X)\}^2}\right] = \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{v(X,1)}{e(X)}\right]$$

where

$$v(x,z) = \mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y^2|X = x, Z = z].$$

Note

Thus the magnitude of the variance depends on

$$\mathbb{E}_X^{\mathcal{O}}\left[\frac{v(X,1)}{e(X)}\right]$$

and if $v(x,z) \equiv v$, a constant, then this equals

$$v\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)}\right].$$

In general, although we have assumed positivity ($e(x) > 0$ for all x) we have no guarantee that the expectation in this expression is *finite*; even if it is finite, it may be large due to the reciprocation of $e(X)$.

AIPW via regression

Note

This feature can affect all IPW estimators.

- it is sometimes assumed that $e(x)$ must be bounded away from zero;
- alternatively, it is common to *truncate* the propensity score values such that either data for which, for some $\epsilon > 0$

  $$e(x_i) < \epsilon$$

  are omitted, or to use

  $$e_\epsilon(x_i) = \max\{e(x_i), \epsilon\}.$$

AIPW via regression

Note

Note that

$$\begin{aligned}\widetilde{\delta}_{\text{IPW}} \equiv \widetilde{\mu}_{\text{IPW}}(1) - \widetilde{\mu}_{\text{IPW}}(0) &= \frac{1}{n}\sum_{i=1}^{n}\frac{Z_i Y_i}{e(X_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-Z_i)Y_i}{1-e(X_i)} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i}{e(X_i)} - \frac{(1-Z_i)}{1-e(X_i)}\right)Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i - e(X_i)}{e(X_i)(1-e(X_i))}\right)Y_i .\end{aligned}$$

AIPW via regression

Note

Note also that

$$e(X_i)(1 - e(X_i)) \equiv \mathrm{Var}_{Z_i|X_i}[Z_i|X_i]$$

so in fact

$$\widetilde{\mu}_{\text{IPW}}(1) - \widetilde{\mu}_{\text{IPW}}(0) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Z_i - e(X_i)}{\mathrm{Var}_{Z_i|X_i}[Z_i|X_i]}\right) Y_i$$

which resembles the earlier formulae for the randomized experimental case.

AIPW via regression

Note

Finally, note the variance of $\widetilde{\delta} = \widetilde{\mu}(1) - \widetilde{\mu}(0)$ is

$$\frac{1}{n}\text{Var}^{\mathcal{O}}_{X,Y,Z}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)Y\right].$$

Now, in this expression, the variance term can be written

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)^2 Y^2\right] - \delta^2$$

as $\widetilde{\delta}$ is unbiased for δ.

AIPW via regression

Note

Using the previous notation, we have that the first term is

$$\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)^2 v(X,Z)\right]$$

and if $v(X,Z) \equiv v$, a constant, this reduces to

$$v\,\mathbb{E}_{X,Z}^{\mathcal{O}}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)^2\right].$$

AIPW via regression

Note

Finally, the expectation simplifies

$$
\begin{aligned}
\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)^2\right] = \mathbb{E}^{\mathcal{O}}_{X,Z}&\left[\left(\frac{Z}{e(X)}\right)^2\right] \\
&+ \mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{(1-Z)}{1-e(X)}\right)^2\right] \\
&- 2\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\frac{Z(1-Z)}{e(X)(1-e(X))}\right]
\end{aligned}
$$

However, $Z(1-Z) = 0$ w.p. 1, so the third term is zero.

AIPW via regression

Note

Hence as $Z^2 = Z$ and $(1-Z)^2 = (1-Z)$ w. p. 1,

$$
\begin{aligned}
\mathbb{E}^{\mathcal{O}}_{X,Z}\left[\left(\frac{Z}{e(X)} - \frac{(1-Z)}{1-e(X)}\right)^2\right] &= \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{e(X)}\right] + \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{1-e(X)}\right] \\
&= \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{e(X)(1-e(X)}\right] \\
&= \mathbb{E}^{\mathcal{O}}_{X}\left[\frac{1}{\mathrm{Var}^{\mathcal{O}}_{Z|X}[Z|X]}\right]
\end{aligned}
$$

AIPW via regression

Note

Therefore, the variance of $\widetilde{\delta}_{\text{IPW}}$ is

$$\frac{v}{n}\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{\mathrm{Var}_{Z|X}^{\mathcal{O}}[Z|X]}\right] - \frac{\delta^2}{n}$$

G-estimation

Earlier we saw the idea of *propensity score regression*, where we construct a model of the form

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

which is potentially useful as

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

G-estimation

In the binary treatment, linear model case, we saw that if the data generating model is

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[Y|X=x, Z=z] = \mathbf{x}_0\beta_{\text{TRUE}} + z\,\mathbf{x}_2\psi = \mu(x, z; \beta_{\text{TRUE}}, \psi)$$

then the propensity score regression model

$$m(x, z; \beta, \psi, \phi) = \mathbf{x}_1\beta + z\mathbf{x}_2\psi + e(x)\mathbf{x}_2\phi$$

will block the confounding paths and return a consistent estimator of ψ even if the *treatment-free mean model* $\mathbf{x}_1\beta$ is mis-specified.

Consider the OLS estimation of (β, ψ, ϕ): we solve

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{i1}^\top \\ z_i \mathbf{x}_{i2}^\top \\ e(x_i)\mathbf{x}_{i2}^\top \end{pmatrix} (y_i - \mathbf{x}_{i1}\beta - z_i\, \mathbf{x}_{i2}\psi - e(x_i)\mathbf{x}_{i2}\phi) = \mathbf{0}$$

analytically using the usual approaches.

However, note that subtracting the third component from the second, we obtain the equivalent system

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{i1}^{\top} \\ (z_i - e(x_i))\mathbf{x}_{i2}^{\top} \\ e(x_i)\mathbf{x}_{i2}^{\top} \end{pmatrix} (y_i - \mathbf{x}_{i1}\beta - z_i\,\mathbf{x}_{i2}\psi - e(x_i)\mathbf{x}_{i2}\phi) = \mathbf{0}$$

which has an identical solution.

The second component takes the form

$$\sum_{i=1}^{n} (z_i - e(x_i))\mathbf{x}_{i2}^{\top}(y_i - \mathbf{x}_{i1}\beta - z_i\,\mathbf{x}_{i2}\psi - e(x_i)\mathbf{x}_{i2}\phi) = \mathbf{0}$$

Notice first that if the mean model is *correctly specified*

$$m(x, z; \beta, \psi, \phi) = \mathbf{x}_1\beta + z\mathbf{x}_2\psi + e(x)\mathbf{x}_2\phi$$

with $\phi = \mathbf{0}$, that is, the true model is nested inside the fitted model, then β and ψ will be consistently estimated, and we will observe

$$\widehat{\phi} \xrightarrow{p} 0$$

as $n \longrightarrow \infty$; indeed, for finite n, the expected value of $\widehat{\phi}$ is zero.

G-estimation

Now suppose the mean model is *mis-specified*, but that

(i) the propensity score model $e(x)$ is correctly specified;

(ii) the random quantity

$$\varepsilon_i = (Y_i - \mathbf{X}_{i1}\beta - Z_i\,\mathbf{X}_{i2}\psi - e(X_i)\mathbf{X}_{i2}\phi)$$

is *functionally independent* of Z_i, that is, the dependence of the mean model on Z_i is correctly specified, and the effect of Z_i is captured via

$$Z_i\,\mathbf{X}_{i2}\psi.$$

Then we have that

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))\mathbf{X}_2^\top (Y - \mathbf{X}_1\beta - Z\,\mathbf{X}_2\psi - e(X)\mathbf{X}_2\phi)] = \mathbf{0}$$

as, using iterated expectation, we have first that

$$\mathbb{E}^{\mathcal{O}}_{Y|X,Z}[(Y - \mathbf{X}_1\beta - Z\,\mathbf{X}_2\psi - e(X)\mathbf{X}_2\phi)|X, Z] = h(\mathbf{X}; \beta, \psi, \phi)$$

where

$$\begin{aligned} h(\mathbf{x}; \beta, \psi, \phi) &= (\mathbf{x}_0\beta_{\text{TRUE}} + z\,\mathbf{x}_2\psi) - (\mathbf{x}_1\beta + z\,\mathbf{x}_2\psi + e(x)\mathbf{x}_2\phi)) \\ &= \mathbf{x}_0\beta_{\text{TRUE}} - (\mathbf{x}_1\beta + e(x)\mathbf{x}_2\phi)). \end{aligned}$$

That is, $h(\mathbf{X};\beta,\psi,\phi)$ is *functionally independent* of Z. Then

$$\begin{aligned}\mathbb{E}^{\mathcal{O}}_{Z|X}[(Z-e(X))\mathbf{X}_2^\top h(X;\beta,\psi,\phi)|X] &= \mathbf{X}_2^\top h(X;\beta,\psi,\phi)\mathbb{E}^{\mathcal{O}}_{Z|X}[(Z-e(X))|X] \\ &= \mathbf{0}\end{aligned}$$

by the correct specification of $e(X)$, so the overall expectation is zero.

Thus, this is an *unbiased* estimating equation and therefore the solutions to the resulting equation are *consistent* for the true values.

G-estimation

This is another form of *double robustness*; inference for ψ is correct if *either*

- the mean model, *or*
- the propensity score model

(or both) is correctly specified, *provided* the expectation

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

does not depend on Z.

Note

Under correct specification of the propensity score, the G-estimation procedure is robust to mis-specification of the treatment-free mean model

$$\mathbf{x}_1 \beta$$

so in fact we may re-write the G-estimating equation by combining the two terms that do not depend on Z, and omitting the nuisance parameter ϕ from the procedure.

Note

That is, consider the reduced form

$$\sum_{i=1}^{n} \begin{pmatrix} \mathbf{x}_{i1}^{\top} \\ (z_i - e(x_i))\mathbf{x}_{i2}^{\top} \end{pmatrix} (y_i - \mathbf{x}_{i1}\beta - z_i\, \mathbf{x}_{i2}\psi) = \mathbf{0}.$$

This form still leads to double robustness by identical arguments.

The most basic form of the G-estimating equation arises from the model that omits the treatment-free component:

$$\sum_{i=1}^{n}(z_i - e(x_i))\mathbf{x}_{i2}^{\top}(y_i - z_i\,\mathbf{x}_{i2}\psi) = \mathbf{0}$$

and in the simplest case with ψ one-dimensional

$$\sum_{i=1}^{n}(z_i - e(x_i))(y_i - z_i\psi_0) = 0$$

say.

G-estimation

The estimating equation invokes the moment requirement

$$\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))(Y - Z\psi_0)] = 0$$

which is a form of *orthogonality* statement, that is

$$(Z - e(X)) \text{ is } \textit{uncorrelated} \text{ with } (Y - Z\psi_0).$$

G-estimation

In this case we can solve explicitly to obtain

$$\widehat{\psi}_0 = \frac{\displaystyle\sum_{i=1}^{n} (z_i - e(x_i))y_i}{\displaystyle\sum_{i=1}^{n} z_i(z_i - e(x_i))}$$

with corresponding estimator

$$\frac{\displaystyle\sum_{i=1}^{n} (Z_i - e(X_i))Y_i}{\displaystyle\sum_{i=1}^{n} Z_i(Z_i - e(X_i))}.$$

G-estimation

Using standard arguments, we have that as $n \longrightarrow \infty$

$$\frac{\sum_{i=1}^{n} (Z_i - e(X_i))Y_i}{\sum_{i=1}^{n} Z_i(Z_i - e(X_i))} \quad \xrightarrow{p} \quad \frac{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y]}{\mathbb{E}^{\mathcal{O}}_{X,Z}[Z(Z - e(X))]}.$$

and note that in the denominator, by iterated expectation

$$\mathbb{E}^{\mathcal{O}}_{X,Z}[Z(Z - e(X))] = \mathbb{E}^{\mathcal{O}}_{X}\left[\mathbb{E}^{\mathcal{O}}_{Z|X}[Z(Z - e(X))|X]\right].$$

Then, as $Z^2 = Z$ w.p. 1, we have

$$\mathbb{E}^{\mathcal{O}}_{Z|X}[Z(Z - e(X))|X] = \mathbb{E}^{\mathcal{O}}_{Z|X}[Z^2 - Ze(X)|X] = e(X)(1 - e(X))$$

Thus

$$\mathbb{E}^{\mathcal{O}}_{X,Z}[Z(Z - e(X)] = \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))]$$
$$\equiv \mathbb{E}^{\mathcal{O}}_{X}[\mathrm{Var}^{\mathcal{O}}_{Z|X}[Z|X]]$$

where the second line follows as

$$Z|X \sim \mathit{Bernoulli}(e(X)).$$

G-estimation

In the numerator

$$
\begin{aligned}
\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y] &= \mathbb{E}^{\mathcal{O}}_{X,Z}[(Z - e(X))\mu(X, Z)] \\
&= \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))\mu(X, 1) - (1 - e(X))e(X)\mu(X, 0)] \\
&= \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))(\mu(X, 1) - \mu(X, 0))] \\
&= \psi_0 \mathbb{E}^{\mathcal{O}}_{X}[e(X)(1 - e(X))]
\end{aligned}
$$

as, here

$$
\mu(X, 1) - \mu(X, 0) = \psi_0
$$

with probability 1.

Therefore

$$\frac{\sum_{i=1}^{n}(Z_i - e(X_i))Y_i}{\sum_{i=1}^{n} Z_i(Z_i - e(X_i))} \quad \xrightarrow{p} \quad \frac{\psi_0 \mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]} = \psi_0.$$

and we have consistent estimation.

For the variance, note first that by the previous result

$$\widehat{\psi}_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{(Z_i - e(X_i))}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]} Y_i + \mathrm{o}_p(1)$$

so we may compute the large sample variance by computing the variance of the statistic on the right hand side; this variance is

$$\frac{1}{n\{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X))]\}^2} \mathrm{Var}_{X,Y,Z}^{\mathcal{O}}[(Z-e(X))Y].$$

We have that

$$
\begin{aligned}
&\mathrm{Var}^{\mathcal{O}}_{X,Y,Z}[(Z-e(X))Y] \\
&= \mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z-e(X))^2Y^2] - \{\mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z-e(X))Y]\}^2 \\
&= \mathbb{E}^{\mathcal{O}}_{X,Y,Z}[(Z-e(X))^2Y^2] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_{X}[e(X)(1-e(X)]\}^2 \\
&= \mathbb{E}^{\mathcal{O}}_{X,Z}[(Z-e(X))^2v(X,Z)] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_{X}[e(X)(1-e(X)]\}^2
\end{aligned}
$$

If $v(x,z) = v$ is a constant, then this becomes

$$v\mathbb{E}^{\mathcal{O}}_{X,Z}[(Z - e(X))^2] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_X[e(X)(1 - e(X)]\}^2$$

but by iterated expectation

$$\mathbb{E}^{\mathcal{O}}_{X,Z}[(Z - e(X))^2] = \mathbb{E}^{\mathcal{O}}_X[e(X)(1 - e(X)].$$

Thus

$$\begin{aligned}&\mathrm{Var}^{\mathcal{O}}_{X,Y,Z}[(Z - e(X))Y] \\ &\qquad = v\mathbb{E}^{\mathcal{O}}_X[e(X)(1 - e(X)] - \psi_0^2\{\mathbb{E}^{\mathcal{O}}_X[e(X)(1 - e(X)]\}^2.\end{aligned}$$

G-estimation

Therefore, combining all the elements, we conclude that the variance of $\widehat{\phi}_0$, obtained by G-estimation, satisfies

$$n\mathrm{Var}^{\mathcal{O}}[\widehat{\psi}_0] \longrightarrow v\frac{1}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X)]} - \psi_0^2.$$

Recall that in this model

$$\psi_0 = \mathbb{E}_X^{\mathcal{O}}[\mu(X,1) - \mu(X,0)] = \mu(\textcolor{red}{1}) - \mu(\textcolor{red}{0})$$

so ψ_0 is the ATE.

We contrast this with the variance of the IPW estimator of the ATE obtained earlier: we had that as $\delta = \psi_0$,

$$n\mathrm{Var}^{\mathcal{O}}[\widetilde{\delta}_{\text{IPW}}] = v\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)(1-e(X))}\right] - \psi_0^2$$

G-estimation

Now by Jensen's inequality

$$\mathbb{E}_X^{\mathcal{O}}\left[\frac{1}{e(X)(1-e(X))}\right] \geqslant \frac{1}{\mathbb{E}_X^{\mathcal{O}}[e(X)(1-e(X)]}$$

and so it is evident that for n large enough

$$\mathrm{Var}^{\mathcal{O}}[\widetilde{\delta}_{\text{IPW}}] > \mathrm{Var}^{\mathcal{O}}[\widehat{\psi}_0].$$

Recall, however, that the two methods make different assumptions: specifically, G-estimation requires the *correct specification* of the treatment effect model.

Note

These results extend to more complicated settings: for example, the *doubly robust* G-estimator takes the form

$$\frac{\displaystyle\sum_{i=1}^{n}(Z_i - e(X_i))(Y_i - \mathbf{X}_{i1}\widehat{\beta})}{\displaystyle\sum_{i=1}^{n} Z_i(Z_i - e(X_i))}.$$

and we can achieve similar comparisons with AIPW estimators.

G-estimation

Note

Throughout, we have assumed $e(X)$ is known precisely. More typically, we will propose a parametric model $e(x) \equiv e(x;\alpha)$, and then estimate α using a further estimation procedure.

For example, using *logistic regression*, we could solve

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

where $\mathbf{x}_i$ is a row vector of the same dimension as α.

Having obtained $\widehat{\alpha}$, we then proceed with $e(x_i;\widehat{\alpha})$ in place of $e(x_i)$ in the earlier formulae, using a *plug-in* strategy.

G-estimation

Note

The plug-in approach will work provided the estimator of α is consistent; but

- Should we 'pay a penalty' for estimating α, that is, will the variance of the ATE estimators *increase* ?
- Do we need to account for the estimation of α ?