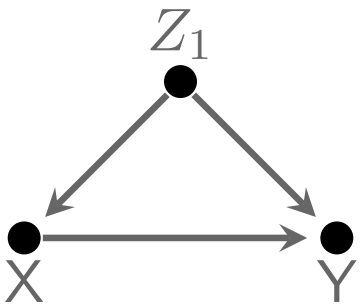
Faculty of Health Sciences

SASA '23 Short Course on Causal Inference

Lecture 2 - Estimation of Causal Effects

November 2023

Case 2



Estimation

Let Z_1 be normal (0) versus low (1) BMI at the time of vaccination. Here, $P(Z_1 = 1) = 0.2$

$E\{Y|X = 1, Z_1 = 1\} = 2$

$E\{Y|X = 0, Z_1 = 1\} = 1.8$

$E\{Y|X = 1, Z_1 = 0\} = 3$

$E\{Y|X = 0, Z_1 = 0\} = 2$

$E\{Y(1)\} = 2 * 0.2 + 3 * 0.8 = 2.8$

$E\{Y(0)\} = 1.8 * 0.2 + 2 * 0.8 = 1.96$

The average causal effect is $2.8 - 1.96 = 0.84$. So, on average, the HPV vaccination causes 0.84 more children to be born.

G-formula

What we just did can also be written as
$E(Y(x)) = \sum_z E(Y|X = x, Z_1 = z)P(Z_1 = z)$

This is called the G-formula.

We can also write it for the Probability as for a binary outcome W:
$P(W(x) = 1) = \sum_z P(W = 1|X = x, Z_1 = z)P(Z_1 = z)$

Standardization and G-computation

In practice, how do we do this? $\hat{E}(Y(x)) = \frac{1}{n}\sum_i \hat{Y}(x|z_{1i})$
Where we set $X = x$, and use the observed values of Z_1.

But how do we get this $\hat{Y}(x|z_{1i})$?

Outcome modeling!

Linear Regression Example

We fit the linear regression:
$E\{Y|X, Z_1\} = \beta_0 + \beta_1 X + \beta_2 Z_1$

$\hat{Y}_i(1) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 * z_{1i}$

$\hat{Y}_i(0) = \hat{\beta}_0 + \hat{\beta}_2 * z_{1i}$

And our estimate is $\frac{1}{n}[\sum_i \hat{Y}_i(1) - \sum_i \hat{Y}_i(0)]$
Here β_1 is also an estimate of the average casual effect because linear regression is collapsible.

Collapsible

This is actually a pretty good definition of collapsible!

If $\frac{1}{n}[\sum_i \hat{Y}_i(1) - \sum_i \hat{Y}_i(0)]$ is the same as the β associated with X in your model (without an interaction) then it is collapsible.

Misspecificifaction

Misspecified model - The true generating mechanism is not contained in the mechanisms that are possible under the selected model.

Correctly Specified - A model is correctly specified if it is not misspecified.

Correctly specified for confounding - A correctly specified model that contains a sufficient set of confounders.

Example Correctly Specified

A model can be Correctly Specified, but not correctly specified for confounding.

For example, a logistic regression model with two binary variables and their interaction is always correctly specified, but if this model does not contain all confounders, it is not correctly specified for confounding.

Back to linear regression example

If $E\{Y|X, Z_1\} = \beta_0 + \beta_1 X + \beta_2 Z_1$ is correctly specified for confounding, then β_1 is a consistent estimate of the average causal effect of vaccination on the number of births.
If it is not, there can be two reasons for the bias:

1. our DAG was wrong and for example, Z_2 is also a confounder, then there is not conditional exchangeability **unmeasured confounding**
2. the model should have contained an interaction between X and Z_1 to properly model the relationship between Z_1 and Y, thus the model is misspecified, **Residual confounding**

We will never know that these are true, which is the risk of attempting causal inference in observational studies.

We can attempt to do something about the misspecification, e.g. double robustness.

Interaction model

We correct the model
$E\{Y|X, Z_1\} = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 X * Z_1$

$\beta_1 \neq E\{Y(1)\} - E\{Y(0)\}$

G-computation
$\hat{Y}_i(1) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 * z_{1i} + \beta_3 * z_{1i}$

$\hat{Y}_i(0) = \hat{\beta}_0 + \hat{\beta}_2 * z_{1i}$

And our estimate is again $\frac{1}{n}[\sum_i \hat{Y}_i(1) - \sum_i \hat{Y}_i(0)]$

This model is still collapsible, but we have spread out the effect of X over levels of Z_1 so we need to sum them up.

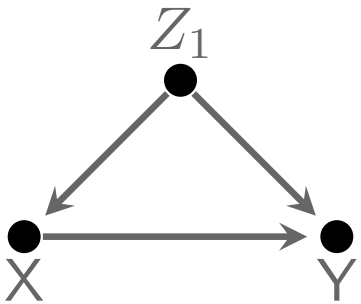
In practice

Your outcome model will generally be much more complicated. This doesn't matter you just need to be able to get the predictions from the model.

$q(E\{Y|X, Z_1, \ldots Z_4\}) = \beta_0 + \beta_1 * X + m(X, Z_1, Z_2, Z_3, Z_4, \boldsymbol{\gamma})$
Where $m(X, Z_1, Z_2, Z_3, Z_4, \boldsymbol{\gamma})$ can be as complicated a function of the variables as you wish.

Provided you can get $\hat{Y}(X = 1; z_{1i}, \ldots z_{4i})$, it doesn't matter how complicated!!

Removing Arrows



So far we have used estimation to remove the arrow from Z_1 to Y, when estimating the effect of X on Y. But could we not remove the arrow from Z_1 to X?

Propensity score

If we can model $P(X = 1|Z_1)$, and it is correctly specified and contains all confounders, then we can use that to estimate $W_i = \frac{X_i}{P(X=1|z_{1i})} + \frac{1-X_i}{1-P(X=1|z_{1i})}$ to make an unconfounded data set. Denote $\hat{w}_i$ the estimated version of W_i

Propensity score matching is one way of doing this (although it generally gets you the causal effect among the exposed not the average causal effect)

We can also simply weight the mean.
$\hat{E}\{Y(1)\} = \frac{1}{\sum_i x_i \hat{w}_i} \sum_i y_i * \hat{w}_i * x_i$
$\hat{E}\{Y(0)\} = \frac{1}{\sum_i (1-x_i) \hat{w}_i} \sum_i Y_i * \hat{w}_i * (1 - x_i)$ Then again
$\hat{E}\{Y(1)\} - \hat{E}\{Y(0)\}$

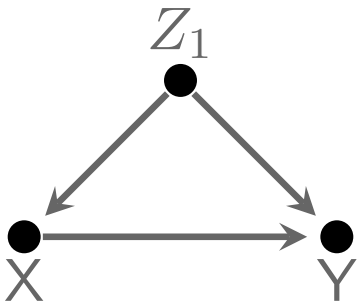
Note

Since $E\{\sum_i x_i \hat{w}_i\} = n$ we can also use

$$\hat{E}\{Y(1)\} = \frac{1}{n}\sum_i Y_i * \hat{w}_i * x_i$$
$$\hat{E}\{Y(0)\} = \frac{1}{n}\sum_i Y_i * \hat{w}_i * (1 - x_i)$$

Then again
$\hat{E}\{Y(1)\} - \hat{E}\{Y(0)\}$

Make a randomized trial setting



By weighting we are removing the arrow from Z_1 to X in the DAG, when estimating the effect of X on Y, thus we do not need to further adjust for anything when estimating. We are back in the randomized trial setting if the propensity score is correctly specified for confounding.

Modeling $P(X|Z)$

Generally, in practice, we are going to use a logistic regression model.
$q(p(X = 1|Z_1)) = \alpha_0 + \alpha_1 Z_1$
When q is the logit function or $\log(\frac{p}{1-p})$, so our
$\widehat{P(X|z_{1i})} = q^{-1}(\hat{\alpha}_0 + \hat{\alpha}_1 z_{1i}) = \text{Expit}(\hat{\alpha}_0 + \hat{\alpha}_1 z_{1i})$

You can get this directly from R using the predict function and setting the type option to "response"

In practice, you are going to have many more variables in your model, and likely some will be continuous. We will use this in the exercise.