1 Bias due to missing predictors.

- Suppose the true model is

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{1}$$

- and we fit the model

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \tag{2}$$

- The least squares estimates are

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y} \\ &= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon})\end{aligned}$$

- which has mean $\boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2$.
- This means that the estimate of $\boldsymbol{\beta}_1$ is biased unless $\boldsymbol{X}_1^T\boldsymbol{X}_2 = [0]$, that is, unless the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal.
- In practice this means that unless you include all of the right predictors in your model, then your estimates will be biased.

2 Inclusion of unrelated predictors

- Suppose that (2) is the correct model, and we fit (1), meaning we've included unrelated predictors $\boldsymbol{X}_2$. The parameter estimates in $\hat{\boldsymbol{\beta}}_1$ will still be unbiased, but their estimated variances will typically be incorrect because

$$Var(\hat{\boldsymbol{\beta}}_1) = \sigma^2[\boldsymbol{X}_1^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{X}_1]^{-1}$$

rather than

$$Var(\hat{\boldsymbol{\beta}}_1) = \sigma^2[\boldsymbol{X}_1^T\boldsymbol{X}_1]^{-1}$$

3 Confounder bias

Suppose that we are interested in assessing the relationship between X and Y, specifically does X cause Y. Also, assume that there is no direct relationship. However, suppose that there is a 3rd variable Z and for which large values of Z result in large values of both X and Y. If Z is not controlled for, then it will look as though X causes Y, whereas their relationship is driven solely by Z. Z is referred to as a confounder.

An example: Does increased coffee consumption increase the risk of cardiovascular disease.

- Coffee drinking is the predictor of interest (X).
- Cardiovascular disease (Y) is the outcome variable.
- Cigarette smoking (Z) correlates with drinking coffee and triggers heart disease on its own. Heavy smokers tend to drink lots of coffee, and have increased incidence of cardiovascular disease.
- In older or poorly adjusted observational studies, researchers noticed that people who drank large amounts of coffee had higher rates of cardiovascular disease. However, this apparent risk was heavily distorted by confounder bias. Because a large proportion of heavy coffee drinkers were also smokers, the data falsely attributed the cardiovascular damage of smoking to coffee consumption.
- To remove the bias, look at the relation between X and Y after controlling for Z. Use "lm(Y +X)" rather than "lm(Y +Z)". The Anova table for the former looks at the effect of X after the effect of Z has been removed.
- To do - draw a picture illustrating Berkson's paradox.

4 Collider bias

(Notes in process)

Suppose that X and Y both cause Z. In the terminology of causal graphs, Z is termed a collider.

If Z is controlled for, either by including in the regression $Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon$, or by stratifying the analysis (say by regressing Y on X for fixed values of Z), then typically the analysis will be biased. For example, if Z is fixed, and

X changes, then Y will change to compensate. To see this, consider the pair of causal linear models (without error, for simplicity)

$Z = \beta_0 + \beta_1 X$ and $Z = \gamma_0 + \gamma_1 Y$. This means that

$Y = \alpha_0 + \alpha_1 X$, with $\alpha_1 = \beta_1 / \gamma_1$ and $\alpha_0 = (\beta_0 - \gamma_0)/\gamma_1$.

When we condition, or stratify, on Z, we introduce a relationship between X and Y, whereas there may be no direct association between X and Y. This type of bias is referred to as collider bias - the bias caused by controlling for a common effect of an exposure and an outcome. A classic example concerned the association between locomotor disease and respiratory illness.