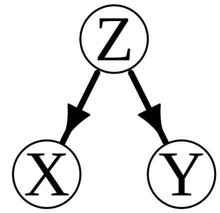
Another look at confounding

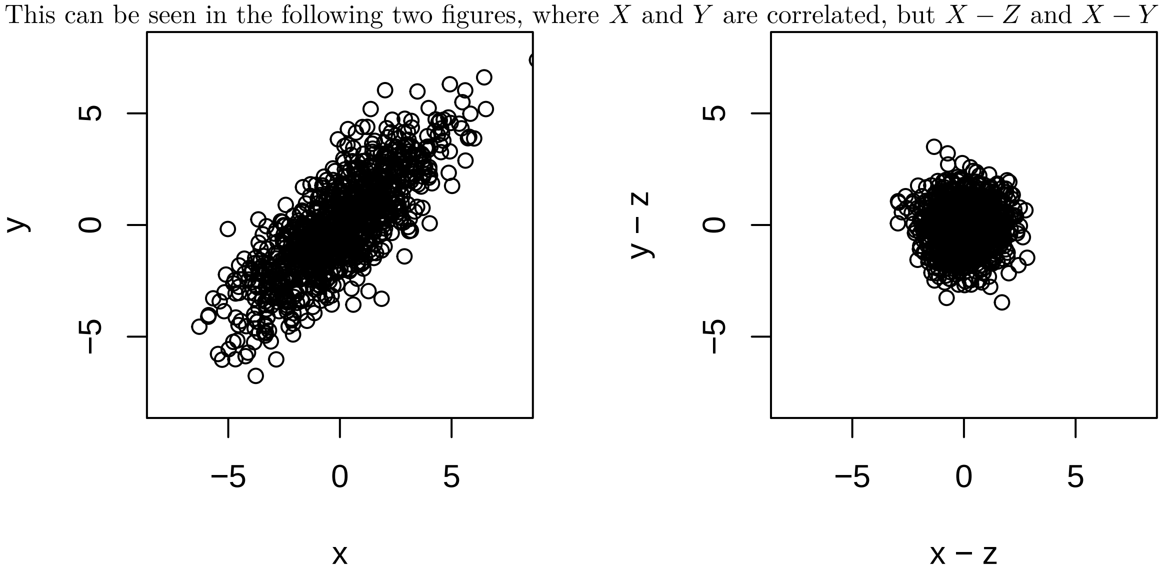
2017-03-29

Assume that one is interested in studying the relationship between two random variables X and Y. Simply put, a variable Z is said to be confonfounding if it explains some (or all) of the dependence between X and Y. For instance, if Z is such that $X \perp Y \mid Z$, then one says that Z is a confounding variable for the relationship between X and Y. This situation is illustrated in the following graphical model: 

For instance, with $X \mid Z \sim N(0,1)$, $Y \mid Z \sim N(0,1)$, and $Z \sim N(0,2)$, we have that $corr\,(X,Y) = 0.8$, namely X and Y are correlated. However, $corr\,(X,Z) = \sqrt{0.8}$ and $corr\,(Y,Z) = \sqrt{0.8}$, which imply that the there is no partial correlation between X and Y when controlling for the effect of Z, that is

$$corr\,(X,Y;Z) = \frac{corr\,(X,Y) - corr\,(X,Z)\,corr\,(Y,Z)}{\sqrt{1 - corr\,(X,Z)^2}\sqrt{1 - corr\,(X,Y)^2}} = 0.$$

This can be seen in the following two figures, where X and Y are correlated, but $X - Z$ and $X - Y$ are not:



As a next example, consider the conditional correlation between the random variables X and Y given X, which is defined by

$$corr\,(X,Y \mid Z) = \frac{E\left[\{X - E\,(X \mid Z)\}\,\{Y - E\,(Y \mid Z)\} \mid Z\right]}{\left(E\left[\{X - E\,(X \mid Z)\}^2 \mid Z\right] E\left[\{Y - E\,(Y \mid Z)\}^2 \mid Z\right]\right)^{1/2}}$$

For instance, when $X = A + BZ$ and $Y = C + DZ$ where A, B, C, D are independently distributed random variables, then

$$corr\left(X, Y \mid Z\right) = \frac{cov\left(A, C\right) + \left\{cov\left(A, D\right) + cov\left(B, C\right)\right\} Z + cov\left(B, D\right) Z^2}{\left[\left\{var\left(A\right) + 2\, cov\left(A, B\right) Z + var\left(B\right) Z^2\right\} \cdot \left\{var\left(C\right) + 2\, cov\left(C, D\right) Z + var\left(D\right) Z^2\right\}\right]^{1/2}}.$$

Furthermore, if A, B, C, D have equal variance and correlation ρ, we have that

$$corr\left(X, Y \mid Z\right) = \frac{\rho\left(1 + Z\right)^2}{\left(1 + 2\,\rho Z\ + Z^2\right)}.$$

In this example, it is clear that the conditional correlation depends on the value of the conditioning variable Z explicitly. Hence, it is in general not equal to the partial correlation.

To study how Z can act as a confounding variable in this context, it is interesting to start from another example. Consider $X, Y, Z \sim N(0, 1)$ marginally, but such that

$$corr\left(X, Y \mid Z\right) = \begin{cases} -0.5 & \text{if } Z \leq 0 \\ 0.5 & \text{otherwise} \end{cases}.$$

Then, because $P(X \leq 0) = 1/2$, we have that $corr\left(X, Y\right) = 0$. In other words, without looking at the effect of Z, one would wrongly conclude that X and Y are uncorrelated, as observed in the following figure: