Omitted variable bias

If we miss out an important variable it not only means our model is poorly specified it also means that any estimated parameters are likely to be biased. If the true model were

$$Y_t = \alpha_0 + \alpha_1 X_t + \alpha_2 Z_t + u$$

and we estimate

$$Y_t = \beta_0 + \beta_1 X_t + u$$

Then the omitted variable can be considered as a function of X in a conditional or auxiliary regression

$$Z_t = \gamma_0 + \gamma_1 X_t + w_t$$

So we have estimated

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2(\gamma_0 + \gamma_1 X_t + w_t) + u_t$$

or

$$Y_t = (\beta_0 + \beta_2\gamma_0) + (\beta_1 + \gamma_1\beta_2)X_t + (\beta_2 w_t + u_t)$$

$$Y_t = \delta_0 + \delta_1 X_t + \varepsilon_t$$

So unless $\beta_2 = 0$ $E(\hat{\beta}_1) \neq 0$

$$E(\hat{\beta}_1) = \beta_1 + \beta_2\left[\frac{\sum x_t z_t}{\sum x_t^2}\right]$$

so there will be a bias as the coefficient of X picks up the part of the influence of Z that was correlated with X.

- The coefficient estimate can have positive or negative bias
- Its standard error will also be biased positively
- The bias on the coefficient estimate can either cancel or reinforce the bias in the standard error when we do a t test.

  So it is not clear what the effect will be.
- Note that while incorrect omission of variables leads to biased estimates of the parameters that are included
- Incorrect inclusion only produces inefficient estimates, so don't have minimum variance
- So better to include the wrong variables rather than exclude the right ones.