Causal Inference and Regression

While adjusting for pre-treatment variables is advised, post treatment variables should not be treated in the same fashion.

Recall the idea of ignorability, where $y^0, y^i \perp z$, then under randomization there will be no differences, on average,

$$y = \tau z + \epsilon$$

Conditional ignorability also holds on a pretreatment variable, x such that $y^0, y^i \perp z|x$ then under randomization there will be no differences, on average, in the distribution of potential outcomes is the same across levels of the treatment after controlling for x.

However for a post-treatment variable q, we cannot, in general, state that $y^0, y^i \perp z|x, q$. The issue here is that q can be influenced by the treatment. The result is that

$$y = \tau^* z + x\beta^* + \delta q + \epsilon$$

The variable q is often referred to as an intermediate variables. The issue is when the intermediate variables is influenced by the treatment.

Hence, estimating the treatment effects and potential outcomes in y, conditional on q, would require accounting for both potential outcomes in q.

ROS states "randomized experiments are a black box approach to causal inference. We see what goes in (treatments) and see what comes out (outcomes), and we can make inferences about the relationships between these inputs and outputs."

Note that post treatment "mediating variables" induce challenges in interpreting the "causal paths" and require more thought than using regression for intermediate outcomes.

Observational studies and causal inference

We have looked at causal inference through the lens of randomized, designed experiments. Designed experiments, and ignorability in treatment assignment (based on potential outcomes), enabled estimates of average treatment effects.

Unfortunately, random treatment assignment is not always possible.

ROS describes an observational study to be the opposite of a designed (randomized) experiment.

With an observational study, under this definition, there may or may not be a direct manipulation of the treatment.

Generally, it is not reasonable to consider the treatment assignment as random across the groups.

Selection bias, where units receive a treatment or control based on some non-randomized mechanism, is a major issue for causal inference in observational studies. Often treatment assignment is confounded with other information, which presents challenges in estimating treatment effects.

If outcomes were compared, conditional on the confounding variable, then we *could* make causal claims.

Failing to account for lurking variables results in biased estimates of the treatment effect.

Consider the following "true" model:

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \epsilon_i$$

if x_i is not included in the model, but should be,

Then the original model can be rewritten as

$$y_i = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)z_i + \epsilon_i + \beta_2\nu_i$$

where $\beta_1^* = \beta_1 + \beta_2\gamma_1$.

In omitting the lurking variable, we hope to estimate β_1, but instead estimate β_1^*, which is biased unless: