Causality, OVB, And Endogenity

LR, Interpretation & More Testing

- Interpretation of coefficients under different specifications
  - No transform, log(y), log(x), log(x) and log(y)
- Other specifications
  - Indicator
  - No intercept
  - Dummy variables
- Hypothesis test in MLR
  - For the more adventuresome, where SOS = Sums Of Squares –
  - SOS partitioning: https://en.wikipedia.org/wiki/Partition_of_sums_of_squares
  - Extra-SOS testing: https://online.stat.psu.edu/stat501/lesson/6/6.3
  - Alternative SOS testing and approaches, sequential and partial : https://towardsdatascience.com/anovas-three-types-of-estimating-sums-of-squares-don-t-make-the-wrong-choice-91107c77a27a

"Causality" The Most Misused Word in Statistics

In order for a researcher to claim that the evidence generated by their study has some kind of causal interpretation requires a lot of justification beyond $E(u|X)=0$

For example, does anyone really think that the graduating from your junior year of high school "causes" your wages to increase by β_1 in the same way that a drug like Lipitor reduces your cholesterol? No of course not, so unless you can justify a model as having a causal interpretation in terms of things like the potential outcome framework, confounding, and selection bias, please stick to associative (descriptive) interpretations.

Regression is a statistical analysis technique, the purpose is to estimate a target variable y according to some other variables x1, x2, ..., namely, if you passively "see" that X=x, what the value of Y will be? Regression cares about “linear” correlation relationships. You can always get a regression formula between Y and X, even when they by no means entail any causal relationships.

However, correlation does not necessary mean causation. An example is that we can build a regression model between "quality of an instructor" and "amount of alcohol the consume" but there shouldn't exist actual causal relationship between them.

Causal analysis tries to estimate the effect of intervention. that is, if you actively "make" X=x, what will Y be? For better understanding, I really recommend Judea Pearl very classic Causality book "Causality: Models, Reasoning and Inference".

Causal Thinking

- Ceteris Paribus – All other things remaining constant or the same
- If wish to establish the existence of a casual relationship between an outcome and a treatment, need control for all other variables
  - Randomize across by randomly assigned samples to treatment
  - Randomly select within blocks variables
  - Include in the model
- Failure to do so will lead to potential lack of sufficient control and selection bias

Omitted Variable Bias

- Omitted Variable Bias (OVB): “The regression version of selection bias generated by inadequate control” – Master Metrics
- Two models
  - $Y_i = \alpha^c + \beta^c X_i + \gamma R_i + \varepsilon^c_i$
  - $Y_i = \alpha^r + \beta^r X_i + \varepsilon^r_i$
- In the second specification we have left out R_i; do not have it or think it unimportant
- This not limited to causal models, but let’s assume that Y is the response or outcome variable, X is the treatment and R is a control variable

OVB Math

- Mathematically: $\beta^r = cov(Y_i, X_i)/V(X_i)$
- Substituting the Y_i from the c model we have
- $Cov(\alpha^c + \beta^c X_i + \gamma R_i + \varepsilon^c_i, X_i)/V(X_i) = \beta^c + \gamma\, cov(R_i, X_i)/V(X_i) = \beta^c + \gamma\, \pi_{R,X}$
- $\pi_{R,X}$ is the regression coefficient of regressing R on X
- Recall γ is the coefficient for regressing Y upon R
- $\gamma\, \pi_{R,X}$ is the OVB
- Interpretation for the OVB to be non-zero: there is both a relationship between 1) the treatment variable and control: $\pi_{R,X} \neq 0$ and 2) the control and the outcome: $\gamma \neq 0$.

$$\beta^r = \beta^c + \gamma \pi_{R,X} \Rightarrow \beta^c = \beta^r - \gamma \pi_{R,X}$$

Linear Regression, Endogeneity, Omitted Variable Bias

- $y(i) = \beta(0) + \beta(1)x(1,i) + \beta(2)\, x(2,i) + \ldots + \beta(p)x(p,i) + \varepsilon(i)$
- Assumptions
  - Linearity
  - IID
  - No perfect multicollinearity
  - Zero expection conditional mean
- Endogenity – x's caused by y (simultaneity problem) or correlation between your x variable and the error term in your model.
  - $E[\varepsilon|X] \neq 0$ for atleast one X ➔ not meeting Zero expection conditional mean assumption
  - Weaker form $E[\varepsilon X] = 0$ yields consistent but biased estimates
- Effects of Endogenity: Omitted variable bias
  - If model used (associative model) departs from true (causal model) and omitted variable is correlated with other predictors and Y, variation from omitted variable will be in the error term and error term will be correlated with predictors
  - Under this circumstance coefficient of correlated and present predictor will be "off" by an additive function of the coefficient of the omitted variable
- To complete – Exogenity: X's are not depend on Y ➔ ε is uncorrelated to X
  - Strict ≡ past, present or future instances of X