**Econometrics**
Chapter 9: Assessing Studies based on Regression

*Internal and External Validity*

Internal - are the inferences about causal effects valid for the population studied

External - can the results be generalized to populations/settings outside from where they were concluded

*Threats to Internal*

**Omitted variable bias:** fixes
1. add the missing
2. control variables
*3. panel data approaches - view overtime so can hold 'missing' constant
*4. instrumental variable

**Functional form mispecification**

Specify linear when really quadratic - bias partial effects

Solution - add nonlinear components

**Measurement Error - Errors in Variables Bias**

Problem when X's measured with error

Maybe coding wrong, people lie, scale was broken etc.

So say X is mismeasured by $\tilde{X}$, then $Y = \beta_0 + \beta_1 X + U$ becomes

$Y = \beta_0 + \beta_1 \tilde{X} + [\beta_1(X - \tilde{X}) + u] = \beta_0 + \beta_1 \tilde{X} + v$

Think about if this case meets the problems associated with the OVB issue.

Classical Measure error - purely random error in measurement

$\tilde{X} = X + w$ where w is purely random (but note it is correlated with $\tilde{X}$)

this implies $\hat{\beta} \to \frac{\sigma_x^2}{\sigma_x^2 + \sigma_w^2}\beta_1$

So as the size of the error becomes large it essentially drowns out any signal from the true data and the estimate is biased to zero.

Error in Y is no problem, just our variances of our $\beta$'s will be larger than they would have been - not as precise in measuring the effects.

Solution: if cannot get good measure the other solution is instrumental variables

*Example:* Intergenerational Elasticity of Income

Want a measure of the effect of parental lifetime earnings (a measure of their status) on children's lifetime earnings.

But of course will not likely have data on lifetime earnings. And we can view any year (once an adult) as an error ridden measure $\tilde{y}_p = y_p + w$.

So it can be shown that estimates of the IGE (which are log-log regressions) will be downward biased but less so with more years averaged for a measure of parental earnings.

A paper with Professor Kiho Jeong using Korean Data:
1 year average: IGE = 0.166 (t-stat = 5.06);
3 yr average: IGE = 0.214 (t-stat = 4.77);
5 yr average: IGE = 0.256 (t-stat = 5.01)

And these are big differences, using 5 vs 1 year gives a measure of IGE that is 54% larger.

**Missing Data and Sample Selection**

If the reason the data is missing is related to Y.

For example want to know the effect of education on women's wages. But only have data on Y for women who chose to work.

Solution beyond scope of this book/class

Survivorship and mutual funds? - only see the ones that survived so a random, or average one will not do as well

**Simultaneous Causality**

Say X causes Y, but Y also causes X.
1. Crime and police force size
2. Test scores and class size (with government programs to expand teachers in low performing areas)

$Y = \beta_0 + \beta_1 X + u$
$X = \gamma_0 + \gamma_1 Y + v$

So u and X are correlated and the OLS results will be biased and inconsistent (u affects Y which affects X so u affects X)

Fix: instrumental variables or experiments

**Inconsistent SE**

Even if OLS estimates are good, need our SE's to be good so that we have reliable tests.

Be sure to heter robust SE's

Also in cases of serial correlation - panel or times series or clustering (geography) use robust SE's

*Internal vs. External and Forecasting*

Up to now focused on causal interpretations, but if only care about forecasting this may not really be a concern. That is internal validity may be ignored.

What if instead of caring about relation between class size and test scores cause want to maybe spend money as an Ed Corridinator you care because you are a parent and want to choose a good school. Even though the basic regression may not be causal, it still predicts the average test score which may be all the parents care about.

External Validity - harder to ask and answer. Ex: do results for Cali extend to all states in US?