Course notes

- Last time
  - Multiple regression
  - Alternative specifications
  - Sign, Size and, Significance
- Time for some Stata questions
- Today
  - Fast recap of Chapter 6
  - Assessing Studies Based on Multiple Regression
What’s in Chapter 6?

Non-linear regression functions

- Curved or bent regression functions (curved relationship between $X$ and $Y$)
  - Does the effect of STR “tail off” at some point?
- Interactions between independent variables
  - Is the effect of distance from college different for men and women (already done in the homework)?
  - Does STR matter more for classrooms with many English learners?

Techniques are fairly easy to implement but beyond the scope of this course. See me for more detail when you do workshop and capstone.

Internal and External Validity

- Focus: causal effect of independent variable on dependent variable, basis of policy.
- Assessing threats to validity for improvement and critique
- Internal Validity: statistical inferences about causal effects are valid for the population being studied.
  - What would happen to California elementary school test scores if every district reduced the STR by two students?
- External Validity: statistical inferences about causal effects can be generalized from the population and setting being studied to other populations and settings.
  - What would happen to California HS test scores if every district reduced the STR by two students?
  - What would happen to Iowa elementary school test scores if every district reduced the STR by two students?
  - What would happen to Japanese elementary school test scores if every district reduced the STR by two students?
Internal Validity

- Want $\hat{\beta}$ to be unbiased and consistent estimator of $\beta$
- Hypothesis tests should have the intended significance level and CI's should have the desired confidence level.
  - Depends on SE's being accurately estimated
- Example of a threat to internal validity: omitted variable bias
- Solution: include omitted variables

External Validity

Example
- Laboratory animal toxicity studies to study, predict, and regulate human exposure and health effects

What can go wrong?
- Differences in Populations between population studied and the population of interest (geography, time [e.g., RAND HIE])
- Differences in Settings (legal, institutional, and physical environments)

Test scores and STR
- ES scores in the U.S. more likely to be an externally valid application than HS scores in U.S. or ES scores in Japan.
- Except: High-stakes testing (for students, teachers, etc.)
Assessing External Validity

Requires

• Specific knowledge of the population and setting studied and the population and setting of interest; or
• Studies on several populations and settings that generate similar results.

Threats to Internal Validity, and Solutions

• Omitted Variable Bias
• Misspecification of the Functional Form
• Imprecise Measurement of the Independent Variables (“Errors-in-Variables”)
• Sample Selection
• Simultaneous Causality

Each of these is an instance of correlation between the regressor ($X$) and the error term ($u$), which violates the first least squares assumption.
Omitted Variable Bias

- Already discussed the problem at length
- Solution when Omitted Variable is Observed
  - Identify the key coefficient(s) of interest
  - Consider which control variables to include based on expert judgment
  - Estimate alternative specifications, keep additional variables that
    - are themselves statistically significant; or
    - affect the sign, size, or significance of the coefficients on key variables.
  - Full disclosure of the specifications tested
- Solution when Omitted Variable is Unobserved
  - Compare a unit to itself (over time or within super-unit)
  - Experimental or quasi-experimental design

Misspecification of the Functional Form

- Curved and changing relationships (Chapter 6)—beyond the scope of this course. Use scatterplots to identify non-linear relationships.
- Discrete outcome variables (Chapter 9)
Errors-in-Variables

• Only a problem with imprecise measurement of the *independent* variables

• The imprecise measurement is not biased up or down, simply $X$ is true signal but $w$ is noise (imprecise measurement of $X$) is added: $\tilde{X} = X + w$.

• Leads to “Attenuation Bias.”

\[
\hat{\beta}_1 \rightarrow \sigma_{\tilde{X}}^2 \bar{\sigma}_{X}^2 + \sigma_{w}^2 \beta_1
\]

$\hat{\beta}_1$ is always an underestimate of $\beta_1$ (estimated effect is closer to zero, smaller than true effect).

• Imprecise measurement of the dependent variable is not a problem: imprecise measurement is simply one of the “other factors” $u$ that affect $Y$.

Errors-in-Variables, cont’d

• Solutions
  
  ○ Multiple independent measures of $X$ (even if all are imprecise)
  
  ○ Adjust estimates for attenuation bias based on estimated size of the imprecision
Sample Selection

- Availability of the data is influenced by a selection process that is related to the value of the dependent variable.
- In all these cases, “other factors” $u$ may be correlated with $X$.
  - 1936 Presidential poll limited to car and telephone owners
  - People who apply for job-training programs likely have barriers to employment.
  - People with jobs may have high earning potential (controlling for their characteristics).
  - InnerChange program evaluation, attrition in general
- Solutions: various and complex; create an explicit model of the selection process.

InnerChange
Simultaneous, or Reverse, Causality

- Government may hire additional teachers in low-performing districts (or now government may penalize low-performing districts).

\[ Y_i = \beta_0 + \beta_1 X_i + u_i \]
\[ X_i = \gamma_0 + \gamma_1 Y_i + v_i \]

- Induces correlation between \( u \) and \( X \).
  - Consider case where \( u_i \) is low, hence \( Y_i \) is low.
  - If \( Y_i \) is low, then (assuming \( \gamma_1 \) positive) \( X_i \) is low.
  - But this means that \( u_i \) and \( X_i \) are low together—correlated!

- Solutions: randomized controlled experiments (Chapter 11) and econometric quasi-experimental methods (beyond the scope of this course).

Summary

- Note that every problem discussed so far involved a violation of OLS assumption #1: the conditional distribution of \( u_i \) given \( X_i \) has mean zero.

- General language for discussing problems with causal models (within and beyond econometrics)
**Inconsistency in OLS Standard Errors**

- The OLS estimates of $\beta$ remain consistent and unbiased; but
- Inference (CI’s, hypothesis tests) will be wrong because the SE’s are wrong.
- Heteroskedasticity: use robust standard errors
- Correlation of the error term across observations

$$
Y_i = \beta_0 + \beta_1 X_i + u_i \\
Y_j = \beta_0 + \beta_1 X_j + u_j
$$

$u_i$ and $u_j$ should not be related.
- Repeated sampling of the same unit over time, “serial correlation”
- Sampling within the same household or geographical unit
- Less (fewer observations) than meets the eye