What is critical replication?

Subject economic and econometric results to tests for accuracy, robustness, and validity

1. Replicate the original result; then

2. Extend
   
   • By applying the same model to new data (different years; different country; different countries; different or placebo policy; different frequency, aggregation, source, or quality of data); or
   
   • By applying a different model to the same data, e.g., political economy (race/class/gender, democracy, environment) functional form, non-linearity, categorization, definitions, outliers, estimation method.
More on Critical Replication

• Why replicate?
  ◦ Test for simple errors
  ◦ Understand the procedure

• Why extend?
  ◦ Test internal or external validity
  ◦ Test the causal model
  ◦ Reassess with a new paradigm

• Some examples in Econ 753
  ◦ Update of Feldstein-Horioka savings-investment model
  ◦ Reassessment of Levine-Zervos
Data for Critical Replication

Alternative strategies

1. Assemble your own from the author’s published instructions (hard!)
2. Use raw data and author’s programs
3. Acquire final dataset from author
Assessing Studies Based on Multiple Regression

Stock and Watson, Chapter 7

- **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?

- “Do countries with lower policy-induced barriers to international trade grow faster, once other relevant country characteristics are controlled for?”
- “Large empirical literature providing an affirmative answer”
Rod-Rod

- Begins with a small learning-by-doing model of ambiguous effect of trade policy
- Prima facie case is weak (Figs. I.1 and I.2)
- Definition of trade barrier
- Focus on trade policies rather than volume of trade
- Data quality
- Relevant outcome (GDP v. welfare)
- Publication bias
Weak relationship between tariffs and growth
Figure I.2: Partial Association between Growth and Non-Tariff Barriers

Weak relationship between NTB’s and growth
Rod-Rod

- Example: Dollar (1992)
  - DISTORTION: index of real exchange rate distortion (based on LOP)
  - VARIABILITY: index of real exchange rate variability
    - Do not necessarily measure trade policy. Role of Exchange-Rate Policy
    - Devaluation strategy incorrectly associated with low distortion
  - Does not include initial income, education, regional dummies (standard variables in cross-sectional growth literature)
  - Not robust to data revisions (PWT 5.6 instead of 4.0)
  - Outlier analysis: Distortion effect driven by Ghana and Uganda
The relationship between trade barriers and economic growth receives weak support from the evidence. Results are contingent on dubious definitions, specific data, outliers.

- Explore contingent relationships (different effects of trade policy in: low- versus high-income countries, mfg versus primary economies, dependent on world economy)
- Disaggregate policy and explore more carefully (tariff, capital controls, export-processing zones, etc.)
- Plant-level datasets, especially does export cause efficiency or efficiency cause export
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.”

\[
\hat{\beta}_1 \xrightarrow{p} \frac{\sigma^2_X}{\sigma^2_X + \sigma^2_w} \beta_1
\]

\( \hat{\beta}_1 \) 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.
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
Internal and External Validity

- External Validity
- Internal Validity
  - Omitted Variable Bias
  - Misspecified Functional Form
  - Errors-in-Variables
  - Sample Selection, InnerChange example
  - Simultaneous (Reverse) Causality
  - Inconsistent Standard Errors
Sample Selection

- Availability of the data is influenced by a selection process that is related to the value of the dependent variable.
- In all 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
**InnerChange**

Goal: evaluation of InnerChange (IFI), a religion-based rehabilitation program for prisoners
Outcome variables: (1) percent re-arrested within two years; (2) percent incarcerated after two years.

\[
H_0 : p_{\text{IFI}} = p_{\text{Control}} \\
H_1 : p_{\text{IFI}} \neq p_{\text{Control}}
\]
Two-sample test of proportion

| Variable | Mean | Std. Err. | z   | P>|z| | 95% Conf. Interval |
|----------|------|-----------|-----|-----|-------------------|
| x        | .243 | .0322377  | .1798152 | .3061848 |
| y        | .203 | .0096042  | .1841761 | .2218239 |
| diff     | .04  | .033638  | -.0259292 | .1059292 |
|          | under Ho: | .031934 | 1.25 | 0.210 |

Ho: proportion(x) - proportion(y) = diff = 0

<table>
<thead>
<tr>
<th>Ha:</th>
<th>diff &lt; 0</th>
<th>Ha: diff != 0</th>
<th>Ha: diff &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>z</td>
<td>1.253</td>
<td>1.253</td>
<td>1.253</td>
</tr>
<tr>
<td>P &lt; z</td>
<td>0.8948</td>
<td>P &gt;</td>
<td>z</td>
</tr>
</tbody>
</table>
InnerChange: Sample Selection Bias?

- Analysis above is based on Intention-To-Treat
- The program evaluation of InnerChange also reported
  - The average outcomes of IFI “graduates,” who completed the 16-month program (in prison and after release) and found employment were better than the average outcomes of the comparison groups.
  - Consider the outcomes of IFI graduates as Treatment-On-Treated (TOT).
- Conditioning the outcome on graduation is selecting the sample on factors that are highly correlated with the outcome variable.
- In this case, the selection of IFI graduates selects winners among the treatment group.
InnerChange: Sample Selection Bias?

- Why isn’t TOT interesting in this case?
  - Because the untreated within the treatment group (IFI participants who did not graduate) did substantially worse than the control group. Two possibilities:
    1. IFI actually harmed some participants while helping others
    2. The program had no effect but selecting graduates selected for good outcomes.
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
Test Scores and STR: External Validity

Massachusetts and California

- Although distinct, both tests are broad measures of student knowledge and academic skill
- Differences in elementary school funding and curriculum
- District-level test scores (on different tests)
- Higher average STR in California (19.6 in CA vs. 17.3 in MA)
- Higher average district income in MA
- Wider spread in income in CA
- More poor students and English learners in CA
## External Validity: Basic Characteristics

<table>
<thead>
<tr>
<th></th>
<th>California</th>
<th>Massachusetts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average</strong></td>
<td><strong>Standard Deviation</strong></td>
<td><strong>Average</strong></td>
</tr>
<tr>
<td>Test scores</td>
<td>654.1</td>
<td>19.1</td>
</tr>
<tr>
<td>Student-teacher ratio</td>
<td>19.6</td>
<td>1.9</td>
</tr>
<tr>
<td>% English learners</td>
<td>15.8%</td>
<td>18.3%</td>
</tr>
<tr>
<td>% Receiving lunch subsidy</td>
<td>44.7%</td>
<td>27.1%</td>
</tr>
<tr>
<td>Average district income ($)</td>
<td>$15,317</td>
<td>$7,226</td>
</tr>
<tr>
<td>Number of observations</td>
<td>420</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>1999</td>
<td></td>
</tr>
</tbody>
</table>
# External Validity: MA Regression

## Table 7.2

<table>
<thead>
<tr>
<th>Dependent Variable: Average Combined English, Math, and Science Test Score in the School District, Fourth Grade; 220 Observations.</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regressor</td>
<td>Student-teacher ratio</td>
<td>-1.72**</td>
<td>-0.69*</td>
<td>-0.64*</td>
<td>12.4</td>
<td>-1.02**</td>
</tr>
<tr>
<td></td>
<td>(STR)</td>
<td>(0.50)</td>
<td>(0.27)</td>
<td>(0.27)</td>
<td>(14.0)</td>
<td>(0.37)</td>
</tr>
<tr>
<td></td>
<td>STR²</td>
<td>-0.689</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.737)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STR³</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>% English learners</td>
<td>-0.411</td>
<td>-0.437</td>
<td>-0.434</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.306)</td>
<td>(0.303)</td>
<td>(0.309)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>% English learners &gt; median? (Binary, Hiel)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-12.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(9.8)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hiel × STR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.56)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>% Eligible for free lunch</td>
<td>-0.521**</td>
<td>-0.582**</td>
<td>-0.587**</td>
<td>-0.709**</td>
<td>-0.653**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.077)</td>
<td>(0.097)</td>
<td>(0.104)</td>
<td>(0.091)</td>
<td>(0.72)</td>
</tr>
<tr>
<td></td>
<td>District income (logarithm)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>16.53***</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3.15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>District income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-3.07</td>
<td>-3.38</td>
<td>-3.87*</td>
<td>-3.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(2.35)</td>
<td>(2.49)</td>
<td>(2.49)</td>
<td>(2.31)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>District income²</td>
<td>0.164</td>
<td>0.174</td>
<td>0.184*</td>
<td>0.165</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.085)</td>
<td>(0.089)</td>
<td>(0.090)</td>
<td>(0.085)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>District income³</td>
<td>-0.0022*</td>
<td>-0.0023*</td>
<td>-0.0023*</td>
<td>-0.0022*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0010)</td>
<td>(0.0010)</td>
<td>(0.0010)</td>
<td>(0.0010)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intercept</td>
<td>739.6**</td>
<td>682.4**</td>
<td>744.0**</td>
<td>665.5**</td>
<td>759.9**</td>
</tr>
<tr>
<td></td>
<td>(8.6)</td>
<td>(11.5)</td>
<td>(21.3)</td>
<td>(81.3)</td>
<td>(23.2)</td>
<td>(20.3)</td>
</tr>
</tbody>
</table>

Table 7.2 continued
External Validity: First case

- California
  - Adding variables that control of student background characteristics reduced the coefficient from $-2.28$ to $-0.73$, a reduction of 68 percent.
  - The null hypothesis ($H_0 : \beta_{\text{STR}} = 0$) was rejected at the 1 percent significance level.

- Massachusetts
  - Adding variables that control of student background characteristics reduced the coefficient from $-1.72$ to $-0.69$, a reduction of 60 percent.
  - The null hypothesis ($H_0 : \beta_{\text{STR}} = 0$) was rejected only at the 5 percent significance level (but there are more observations in CA, which tends to drive down the SE).
External Validity: Truly comparable?

- Different tests, different scores, different differences in scores
  - One test point does not mean the same thing in MA and CA.
  - Standardize the results by dividing the change in test score by the standard deviation of test scores.
  - Expresses the effect on test scores in terms of the observed overall spread in test scores.
### TABLE 7.3  Student-Teacher Ratios and Test Scores: Comparing the Estimates from California and Massachusetts

<table>
<thead>
<tr>
<th></th>
<th>OLS Estimate $\hat{\beta}_{STR}$</th>
<th>Standard Deviation of Test Scores Across Districts</th>
<th>Estimated Effect of 2 Fewer Students Per Teacher, In Units of:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Points on the Test</td>
</tr>
<tr>
<td>California</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear: Table 6.2(2)</td>
<td>$-0.73$ (0.26)</td>
<td>19.1</td>
<td>1.46 (0.52)</td>
</tr>
<tr>
<td>Cubic: Table 6.2(7)</td>
<td>$-$</td>
<td>19.1</td>
<td>2.93 (0.70)</td>
</tr>
<tr>
<td>Reduce STR from 20 to 18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cubic: Table 6.2(7)</td>
<td>$-$</td>
<td>19.1</td>
<td>1.90 (0.69)</td>
</tr>
<tr>
<td>Reduce STR from 22 to 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Massachusetts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear: Table 7.2(3)</td>
<td>$-0.64$ (0.27)</td>
<td>15.1</td>
<td>1.28 (0.54)</td>
</tr>
</tbody>
</table>

Standard errors are given in parentheses.
Test Scores and STR: Internal Validity

- Omitted variable bias: we’ve done our best (English learners, socioeconomic background). Other important possibilities: teacher quality; extracurriculars; family commitment to learning. Next stop: experiment

- Functional form (Chapter 6, not this course): Basic result is that linear is ok.

- Errors-in-Variables: student moves may lead to mismeasurement of district STR; income data are from 1990 Census while other data pertain to late 1990’s.

- Selection: all districts included. Absenteeism?

- Simultaneous causality: for example, allocation of resources based on test performance (either compensatory or sanctioning). Neither California nor Massachusetts had (late 1990’s) significant allocation based on performance.
Test Scores and STR: Internal Validity, cont’d

- Heteroskedasticity-consistent standard errors
- Universe of observations is not the same as a random sample (the ideal under OLS Assumption #2), but probably ok.
Test Scores and STR: Finally done!

- Effect of STR: cutting STR by two students increases test scores by approximately 0.08 standard deviations.
  - Effect is significant but small
- Result appears generalizable in U.S. elementary school systems.
- Consider other omitted variables (listed above)
- Basis for policy decisions, estimating benefits in a cost-benefit analysis of changing STR versus other possible uses of education resources.
- Did we measure the right outcome? Alternative outcomes?