ECON 7310 Elements of Econometrics
Week 7: Internal and External Validity

David Du[1]

[1] University of Queensland

Draft

Outline

Assessing studies based on multiple regression (SW Chapter 9)

- Internal and External Validity
- Threats to Internal Validity
- Application to the California Test Score data set

Assessing Studies Based on Multiple Regression: Internal and External Validity SW Section 9.1

- We will study the most common reasons that multiple regression estimates can result in biased estimates of the causal effect of interest
- In the test score application, we address these threats as best we can.
- **Internal Validity**: the statistical inferences about causal effects are valid for the population being studied.
- **External Validity**: the statistical inferences can be generalized from the population and setting studied to other populations and settings.
  - Here, "setting" refers to the institutional, legal, social, and economic environment, e.g., tomatoes in the lab $\rightarrow$ tomatoes in the field?

Threats to External Validity

Potential threats arise from differences between the population and setting studied and the population and setting of interest.

- **Differences in populations:**
  - True causal effect can be different in the population studied and in the population of interest.
- **Differences in settings:**
  - a study of effect on college binge drinking of an anti-drinking campaign might not generalize to another identical group of college students if legal penalties for drinking at the two colleges are different.
- We find that *TestScore* $\uparrow$ as *STR* $\downarrow$ from CA (elementary) school districts. Can we generalize this result to
  - elementary schools in MA? (Probably Yes..)
  - high schools in CA? (Well... not sure)
  - universities in CA? (Probably, no)

Threats to Internal Validity of Regression Analysis SW section 9.2

- Studies based on regression analysis are internally valid
  1. if the estimated coefficients are unbiased and consistent and
  2. if statistical inference is valid.
- Five threats:
  - Omitted variable bias
  - Wrong functional form
  - Errors-in-variables bias
  - Sample selection bias
  - Simultaneous causality bias
- All of these imply $E[u_i|X_{i1}, \ldots, X_{ik}] \neq 0$.
  Then, OLS is biased and inconsistent.

Threats to Internal Validity: 1. Omitted variable bias

- Recall that if the omitted variable Z satisfies two conditions,
  1. Z is a determinant of Y (i.e. Z is part of u); and
  2. Z is correlated with the regressor X (i.e. $corr(Z, X) \neq 0$),

  OLS estimators are biased and inconsistent.
- **If there is a set of control variables,** include adequate control variables to address the problem of omitted variable bias.
- In practice, however, adding a variable has both costs and benefits;
  - adding an adequate variable reduces omitted variable bias.
  - adding an irrelevant variable reduces precision of the estimator

  So, there is a trade-off b/w bias and variance of the coefficient of interest

Threats to Internal Validity: 1. Omitted variable bias

Five steps to make a decision whether to add a (set of) variable(s);

1. Be specific about the coefficient(s) of interest.
2. Identify the most likely sources of important omitted variable bias, using economic theory and or other *a priori* knowledge, and set up a base specification and a list of additional candidates for control variables
   - "additional resources for kids" directly affects *TestScore* and also indirectly does so via *STR*. List "income" and "lunch support" as a control variable.
3. Augment the base specification with the additional control variables and examine
   - whether the additional coefficients are statistically significant, and/or
   - whether the estimates of the coefficient of interest substantially change

   If so, include those variables in regression.
4. Present an accurate summary of your results in tabular form.
   This "full disclosure" allows for readers to draw their own conclusions.

Threats to Internal Validity: 1. Omitted variable bias

- **If there is no control variable**, you can use either
  1. panel data approach (Chapter 10), or
  2. instrumental variables regressions (Chapter 12), or
  3. data generated from a randomized controlled experiment (Chapter 13)

Threats to Internal Validity: 2. Misspecification of Functional Form

- A misspecification bias arises if the functional form is incorrect.
  For example, an interaction term is incorrectly omitted; then inferences on causal effects will be biased.
- **Solution:**
  - Include higher order terms and/or interaction terms (Chapter 8)

Threats to Internal Validity:
3. Measurement error and errors-in-variables bias

- There are many possible sources of measurement error.
  - typos, coding errors, imprecise/wrong answers to survey questions (intentionally?), etc.
- Suppose we observe $\widetilde{X}_i = X_i + w_i$ (instead of X_i) and estimate

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_i + v_i,$$

  instead of $Y_i = \beta_0 + \beta_1 X_i + u_i$ (correct one). Then, $v_i = -\beta_1 w_i + u_i$.
- As long as w_i is correlated with $\widetilde{X}_i$, we have $E[v_i|\widetilde{X}_i] \neq 0$.
- Suppose w_i is a pure random error with $E[w_i] = 0$, $V(w_i) = \sigma_w^2$, $Cov(w_i, X_i) = 0$, and $Cov(w_i, u_i) = 0$.
- **Classical measurement error model:** regress Y on $\widetilde{X}$. Then

$$\widehat{\beta}_1 \xrightarrow{p} \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \right) \beta_1 \text{. That is, } \widehat{\beta}_1 \text{ will be biased toward 0.}$$

- **Measurement error in Y:** Let $\widetilde{Y}_i = Y_i + w_i$. Suppose we regress $\widetilde{Y}_i$ on X_i. Then, we still have $E[\widehat{\beta}_1] = \beta_1$ and $\widehat{\beta}_1 \xrightarrow{p} \beta_1$, but $V(\widehat{\beta}_1)$ will be larger.
- **Solution:** instrumental variables regression (Chapter 12)

Threats to Internal Validity: 4. Missing Data and Sample Selection

- Data are often missing. We consider three cases:
  1. Data are missing at random
  2. Data are missing based on the value of one or more Xs
  3. Data are missing based in part on the value of Y (or u)
- Cases 1 and 2 do not introduce bias but make standard errors larger
  1. For some reason, you lost half of your sample randomly
     $\rightarrow$ Now, your sample is only smaller: so $\widehat{\beta}$ unbiased but $SE(\widehat{\beta})$ larger.
  2. Suppose we only observe districts with $STR_i > 20$. We can still unbiasedly estimate the effect of class size for districts with $STR > 20$.
- Case 3 introduces "sample selection" bias.
  3. If we estimate the wage equation only using individuals with annual wage larger than \$300k, the estimates will be clearly biased.
- **Solution:** the methods to correct the sample selection bias is beyond the scope of the course, but they are based on techniques in Section 11.3.

Threats to Internal Validity: 5. Simultaneous causality bias

- So far we have assumed that X causes Y. What if Y causes X, too?
- Example: Class size effect
  - Low *STR* results in better test scores
  - But suppose districts with low test scores are given extra resources: as a result of a political process they also have low *STR*
  - What does this mean for a regression of *TestScore* on *STR*?

Threats to Internal Validity: 5. Simultaneous causality bias

- Causal effect on Y of X: $Y_i = \beta_0 + \beta_1 X_i + u_i$
  Causal effect on X of Y: $X_i = \gamma_0 + \gamma_1 Y_i + v_i$
  1. Large u_i means large Y_i, which implies large X_i (if $\gamma_1 > 0$)
  2. Thus, u_i and X_i are correlated, $E[u_i|X_i] \neq 0$.
  3. So, $\widehat{\beta}_1$ is biased and inconsistent.
- **Solutions:**
  1. Estimate instrumental variables regression; Chapter 12 (focus on $X \rightarrow Y$)
  2. Run a randomized controlled experiment; Chapter 13 (block $Y \rightarrow X$)
  3. Develop and estimate a complete model of both directions of causality, i.e., $X \rightarrow Y$ and $Y \rightarrow X$. Ex: large macro models.

Additional threats to Internal Validity: Inconsistency of $SE(\widehat{\beta})$

- Even when OLS estimator is consistent, inconsistent SEs will lead to invalid inference (hypothesis testing & confidence intervals).
- Source of inconsistency of SE:
  - **Heteroskedasticity:** SEs would be inconsistent if we do not use (heteroskedasticity) robust standard errors. In Stata, use the option `robust`.
  - **Correlation of u_i across i:** do not happen if sampling is random. But, in practice, sampling can be only partially random:
    - time series data (serial correlation)
    - panel data (individual i is observed over different time points, serial correlation)
    - some individuals are clustered geographically
    - In many cases, this problem can be fixed by using an alternative formula for SE

Internal and External Validity When the Regression is Used for Forecasting SW Section 9.3

- Forecasting and estimation of causal effects are quite different objectives.
- For forecasting,
  - $\overline{R}^2$ matters (a lot!)
  - Omitted variable bias is not a problem! Don't need consistent $\widehat{\beta}$
  - Interpreting coefficients in forecasting models is not important – the important thing is a good fit and a model you can "trust" to work in your application
  - External validity is paramount: the model estimated using historical data must hold into the (near) future
  - More on forecasting when we take up time series data

Example: Test Scores and Class Size SW Section 9.4

External Validity

- Compare results for California and Massachusetts
- Both CA and MA tests are broad measures of student knowledge and analytic skills
- Elementary schools in US are similar in organization of classroom instructions, although different states have different funding and curriculums
- So, if we find similar results from MA data, that would be evidence of external validity of the findings from CA data.

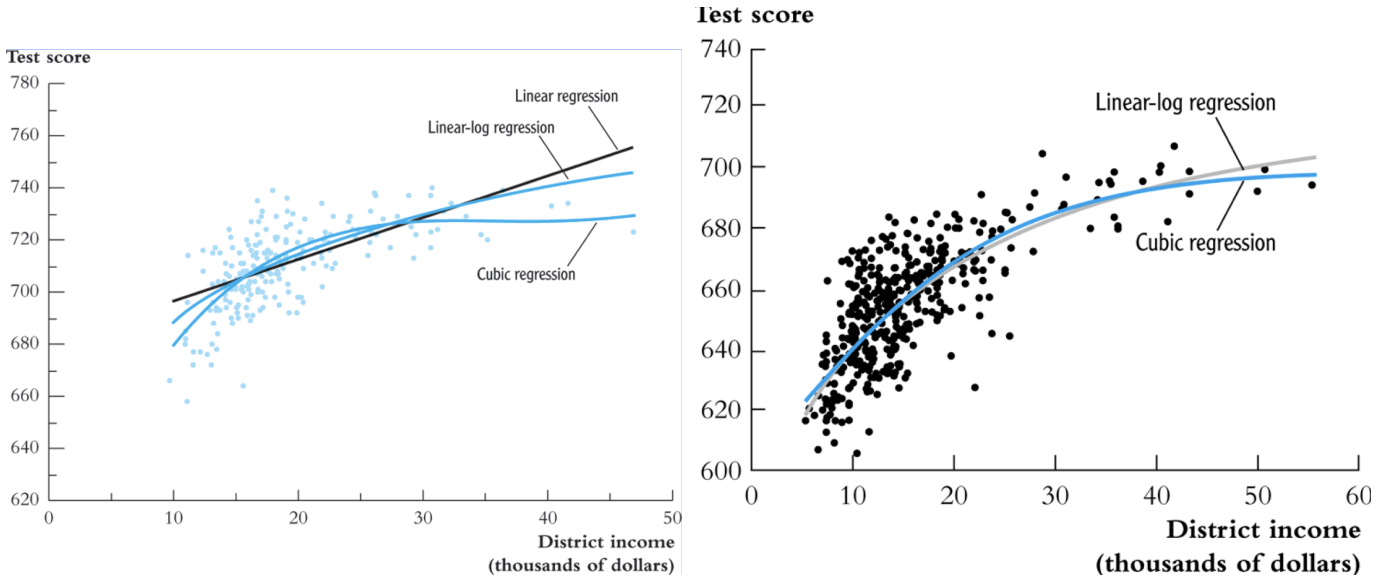
Example: Test Scores and Class Size SW Section 9.4

TABLE 9.1 Summary Statistics for California and Massachusetts Test Score Data Sets

	California		Massachusetts	
	Average	Standard Deviation	Average	Standard Deviation
Test scores	654.1	19.1	709.8	15.1
Student–teacher ratio	19.6	1.9	17.3	2.3
% English learners	15.8%	18.3%	1.1%	2.9%
% Receiving lunch subsidy	44.7%	27.1%	15.3%	15.1%
Average district income ($)	$15,317	$7226	$18,747	$5808
Number of observations	420		220	
Year	1999		1998	

- Average scores cannot be compared directly, as they are different tests
- Average *STR* higher in CA
- Average income lower but more widely spread in CA
- More English Learners and more kids with lunch support in CA

Example: Test Scores and Class Size SW Section 9.4



- (Left, Right) = (MA, CA) for *TestScore* vs *Income*
- They are similar: the relationship is steeper for low value of income
- Best functional forms differ (Cubic for MA, Linear-log for CA)

To assess the external validity of the regression analysis using CA data, run similar multiple regressions using MA data (results presented in SW Section 9.4) and compare two sets of estimation results. We see...

Example: Test Scores and Class Size SW Section 9.4

- Class size effect falls in both CA, MA data when student and district control variables are added.
- Class size effect is statistically significant in both CA, MA data.
- Estimated effect of a 2-student reduction in STR is quantitatively similar for CA, MA.
- Neither data set shows evidence of *STR* – *PctEL* interaction.
- Some evidence of STR nonlinearities in CA data, but not in MA data.
- Overall, this analysis of MA data suggests hat the CA results are externally valid

Example: Test Scores and Class Size SW Section 9.4

Internal Validity

1. Omitted variable bias:

- What causal factors are missing?
  - Student characteristics such as native ability
  - Access to outside learning opportunities
  - Other district quality measures such as teacher quality
- The regressions attempt to control for these omitted factors using control variables that are not necessarily causal but are correlated with the omitted causal variables:
  - district demographics (income, % free lunch eligible)
  - Fraction of English learners

Example: Test Scores and Class Size SW Section 9.4

Are the control variables effective? That is, after including the control variables, is the error term uncorrelated with STR?

$$E[u|X, W] = E[u|W]$$

- Answering this requires using judgment.
- There is some evidence that the control variables might be effective:
  - The STR coefficient doesn't change much when the control variables specifications change
  - The results for California and Massachusetts are similar – so if there is OV bias remaining, that OV bias would be similar in the two data sets
- What additional control variables might you want to use – and what would they be controlling for?

Example: Test Scores and Class Size SW Section 9.4

2. Wrong functional form:

- We have tried quite a few different functional forms, in both the California and Mass. data
- Nonlinear effects are modest
- Plausibly, this is not a major threat at this point.

3. Errors-in-variables bias:

- The data are administrative so it is unlikely that there are substantial reporting/typo type errors.
- STR is a district-wide measure, so students who take the test might not have experienced the measured STR for the district – a complicated type of measurement error
- Ideally we would like data on individual students, by grade level.

Example: Test Scores and Class Size SW Section 9.4

4. Sample selection bias:

- Sample is all elementary public school districts (in CA and in MA.) – there are no missing data
- No reason to think that selection is a problem.

5. Simultaneous causality bias:

- School funding equalization based on test scores could cause simultaneous causality.
- This was not in place in CA or MA. during these samples, so simultaneous causality bias is arguably not important.