Assessing Studies Based on Multiple Regression

SW Chapter 9

Internal and External Validity

Potential threats to internal validity

Learning objectives

- Understand principles of **internal validity** and **external validity**
- Identify threats to internal validity and possible solutions

Internal and External Validity

Internal and external validity

- Validity is when we buy into the results of our estimation
- Did we learn what we set out to learn?
  - Did we estimate the causal impact of class size on test scores?
  - Did we correctly measure the impact of marital status on wages?
- We will go through a framework for evaluating whether a statistical or econometric study is useful for answering a particular question of interest.
- What are the **threats to validity**?

Internal and external validity

- **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, where the "setting" refers to the legal, policy, and physical environment and related salient features.

Threats to external validity

Assessing threats to external validity requires detailed substantive knowledge and judgment on a case-by-case basis.

How far can we generalize class size results from California?

- Differences in populations
  - California in 2011?
  - Massachusetts in 2011?
  - Mexico in 2011?
- Differences in settings
  - different legal requirements (e.g. special education)
  - different treatment of bilingual education
- Differences in teacher characteristics

Internal validity

Internal validity: the statistical inferences about causal effects are valid for the population being studied.

Estimates are internally valid if

- Estimator is consistent (or unbiased)
- In at least 95% of data samples, estimated confidence interval for that data contains the true population parameter (correct coverage)

Potential threats to internal validity

Potential threats to internal validity

We categorize potential threats into three groups:

1. Those that affect consistency of the estimator (most serious)
2. Those that affect confidence intervals but not consistency (wrong inference)
3. Those that make the confidence intervals larger but do not affect consistency or correct coverage

Potential threats to internal validity

- You should place each potential threat into one of these categories
- In the case of 1 and 2, these are *actual* threats to validity
- In the case of 3, estimates become imprecise but not a threat to validity because standard errors properly measure imprecision

Potential threats to internal validity

Five threats to the internal validity of regression studies:

1. Omitted variable bias
2. Wrong functional form
3. Errors-in-variables bias
4. Sample selection bias
5. Simultaneous causality bias

All of these imply that $E(u_i|X_{1i}, , X_{ki}) \neq 0$ (conditional mean independence fails)
$\Rightarrow$ OLS is biased and inconsistent.

1. Omitted variable bias

We know this one!

Omitted variable bias arises if an omitted variable is both:

1. a determinant of Y and
2. correlated with at least one included regressor.

With control variables, are there still omitted factors that are not adequately controlled for? Is the error term is correlated with the variable of interest even after we have included the control variables.?

Solutions to OVB

1. If the omitted causal variable can be measured, include it as an additional regressor in multiple regression;
2. If you have data on one or more controls and they are adequate (in the sense of conditional mean independence plausibly holding) then include the control variables;
3. Run a randomized controlled experiment.
   - Why does this work? Remember – if X is randomly assigned, then X necessarily will be distributed independently of u; thus $E(u|X = x) = 0$.

Coming soon ...

- Possibly, use panel data in which each entity (individual) is observed more than once;
- If the omitted variable(s) cannot be measured, use instrumental variables regression;

2. Wrong functional form

Arises if the functional form is incorrect – for example, an interaction term is incorrectly omitted; then inference on causal effects will be biased.

Solutions to functional form misspecification

- Continuous dependent variable: use the “appropriate” nonlinear specifications in X (logarithms, interactions, etc.)
- Discrete (example: binary) dependent variable: need an extension of multiple regression methods (“probit” or “logit” analysis for binary dependent variables).

Note: This issue is rarely our biggest problem in regression analysis

3. Errors-in-variables bias

So far we have assumed that X is measured without error.

In reality, economic data often have measurement error

- Data entry errors in administrative data
- Recollection errors in surveys (when did you start your current job?)
- Ambiguous questions (what was your income last year?)
- Intentionally false response problems with surveys (What is the current value of your financial assets? How often do you drink and drive?)

Measurement error

- Where does measurement error occur?
  - Dependent variable
  - Independent variables
- Types of measurement error
  - Classical measurement error (random!)
    - Dependent variable: small problem
    - Independent variable: medium problem $\Rightarrow$ attenuation bias
  - Non-classical measurement error: always big problems!

Measurement error in a dependent variable

Mismeasured value = True value + measurement error:

$$y = y^* + e_0 \Rightarrow y^* = y - e_0$$

Population regression:

$$y^* = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

Estimated regression:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + (u + e_0)$$

Measurement error

- True model is $Y_i = \beta_0 + \beta_1 X_i + u_i$
- Let $X_i = X_i^* + w_i$
- Assume
  - $Cov(w, u) = 0$:
    - Measurement error is uncorrelated with other omitted variables
  - $Cov(w, X) = 0 \Leftarrow$ Classical errors-in-variables assumption
    - Measurement error is uncorrelated with the true value of X
  - $Cov(X, u) = 0$
    - No traditional omitted variable bias problems

Formal derivation of measurement error

Formal derivation of measurement error

$$
\begin{aligned}
\widehat{\beta_1} &= \frac{Cov(X_i, Y_i)}{Var(X_i)} \\
&= \frac{Cov(X_i^* + w_i, \beta_0 + \beta_1 X^* + u_i)}{Var(X_i^* + w_i)} \\
&= \frac{Cov(X_i^*, \beta_0) + Cov(X_i^*, \beta_1 X_i^*) + Cov(w_i, \beta_0) + Cov(w_i, \beta_1 X^*) + Cov(w_i, u_i)}{Var(X_i^*) + Var(w_i) + 2Cov(X_i^*, w_i)} \\
&= \frac{0 + Cov(X_i^*, \beta_1 X_i^*) + 0 + Cov(w_i, \beta_1 X^*) + 0}{Var(X_i^*) + Var(w_i) + 2Cov(X_i^*, w_i)} \\
&= \frac{\beta_1 Cov(X_i^*, X_i^*) + \beta_1 Cov(w_i, X^*)}{Var(X_i^*) + Var(w_i) + 2Cov(X_i^*, w_i)} \\
&= \frac{\beta_1 Cov(X_i^*, X_i^*) + 0}{Var(X_i^*) + Var(w_i) + 0} \\
&= \beta_1 \frac{Var(X_i^*)}{Var(X_i^*) + Var(w_i)}
\end{aligned}
$$

Attenuation bias

$$\widehat{\beta_1} = \beta_1 \frac{Var(X_i^*)}{Var(X_i^*) + Var(w_i)}$$

- Fraction on the right is between 0 and 1
- If increasing class size increases test score ($\beta_1 > 0$) then probability limit will be closer to 0
- If increasing class size decreases test score ($\beta_1 < 0$) then probability limit will be closer to 0
- In other words, the magnitude of our estimate will be biased towards 0
- Attenuation Bias: bias towards 0 gives economically smaller magnitude
- Our estimate would suggest that changing class size is less effective than it is

Measurement error in regressor: intuition

$$\widehat{\beta_1} = \beta_1 \frac{Var(X_i^*)}{Var(X_i^*) + Var(w_i)}$$

- $\tilde{X}$ has a higher variance than the true X
  - Measured X is moving more in the data
- When $\tilde{X}$ moves, Y does not respond as much because Y only responds to movements in true X
- Regression sees changes in X not related to Y
- Regression concludes that the effect of X on Y must be smaller in magnitude
- Implies less economic significance

What if we violate CEV?

- Sometimes, CEV unlikely to hold: Measurement error is correlated with unobserved characteristics:
  - If people with high incomes under-report and/or people with low incomes over-report $\leftarrow Cov(income, error) \neq 0$
  - If people are less accurate about their age as they get older $\leftarrow Cov(age, error) \neq 0$
- Measurement error biases OLS estimates
- Interpreting the impacts more complicated, beyond scope of EC200

Summary of measurement error

Dependent variable	OLS still BLUE larger variance	
	Classical	**Non-classical**
Independent variable	OLS no longer BLUE "attentuation bias"	OLS no longer blue bias

Classical errors-in-variables (CEV): Error is not correlated with any unobserved explanatory variables: $Cov(x^*, u) = 0$

How to go on in the face of measurement error

- Dependent variables: Don't worry too much
- Independent variables
  - With attenuation bias, we can still sign our magnitudes (ie, returns to education are at least 10
  - Use "instruments" (later)
- Better data can always help!
- For now – be careful!

4. Missing data and sample selection bias

Data are often missing. Sometimes missing data introduces bias, sometimes it doesn't. It is useful to consider three cases:

1. Data are missing at random.
2. Data are missing based on the value of one or more X's
3. Data are missing based in part on the value of Y or u

Cases 1 and 2 don't introduce bias: the standard errors are larger than they would be if the data weren't missing but $\widehat{\beta}$ is unbiased.

Case 3 introduces "sample selection"bias.

Case 1: missing data at random

Suppose you took a simple random sample of 100 workers and recorded the answers on paper – but your dog ate 20 of the response sheets (selected at random) before you could enter them into the computer.

Essentially, this is like you had taken a simple random sample of 80 workers - your dog didn't introduce any bias!

Case 2: Data are missing based on the value of one or more X's

In the test score/class size application, suppose you restrict your analysis to the subset of school districts with $STR < 20$. By only considering districts with small class sizes you won't be able to say anything about districts with large class sizes, but focusing on just the small-class districts doesn't introduce bias. This is equivalent to having missing data, where the data are missing if $STR > 20$.

More generally, if data are missing based only on values of X's, the fact that data are missing doesn't bias the OLS estimator.

Case 3: Data are missing based in part on the value of Y or u

In general this type of missing data does introduce bias into the OLS estimator. This type of bias is also called **sample selection bias**.

Sample selection bias arises when a selection process:

1. influences the availability of data and
2. is related to the dependent variable.

Example 1: Height of undergraduates

You want to estimate the mean height of undergraduates. You collect your data (obtain your sample) by standing outside the basketball team's locker room and recording the height of the undergraduates who enter.

- Is this a good design – will it yield an unbiased estimate of undergraduate height?
- Formally, you have sampled individuals in a way that is related to the outcome Y (height), which results in bias.

Example 2: Mutual funds

- Do actively managed mutual funds outperform “hold-the-market” funds?
- Empirical strategy:
  - Sampling scheme: simple random sampling of mutual funds available to the public on a given date.
  - Data: returns for the preceding 10 years.
  - Estimator: average ten-year return of the sample mutual funds, minus ten-year return on S&P500
  - Is there sample selection bias? (Equivalently, are data missing based in part on the value of Y or u?)
  - How is this example like the basketball player example?

Sample selection bias induces correlation between a regressor and the error term.

Mutual fund example:

$$return_i = \beta_0 + \beta_1 managedfund_i + u_i$$

- Being a managed fund in the sample ($managedfund_i = 1$) means that your return was better than failed managed funds, which are not in the sample – so $corr(managedfund_i, u_i) \neq 0$.
- The surviving mutual funds are the "basketball players" of mutual funds.

Example 3: Returns to education

- What is the return to an additional year of education among college graduates?
- Empirical strategy:
  - Sampling scheme: simple random sample of employed college grads (employed, so we have wage data)
  - Data: earnings and years of education
  - Estimator: regress ln(earnings) on years of education
  - Ignore issues of omitted variable bias and measurement error – is there sample selection bias?
  - How does this relate to the basketball player example?

Solutions to sample selection bias

- Collect the sample in a way that avoids sample selection.
  - *Basketball player example:* obtain a true random sample of undergraduates, e.g. select students at random from the enrollment administrative list.
  - *Mutual funds example:* change the sample population from those available at the end of the ten-year period, to those available at the beginning of the period (include failed funds)
  - *Returns to education example:* sample college graduates, not workers (include the unemployed)
- Randomized controlled experiment.
- Construct a model of the sample selection problem and estimate that model (we won't do this).

5. Simultaneous causality bias

So far we have assumed that X causes Y. What if Y causes X, too? **Simultaneity!**

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?

Solutions to simultaneous causality bias

1. Run a randomized controlled experiment. Because X_i is chosen at random by the experimenter, there is no feedback from the outcome variable to Y_i (assuming perfect compliance).
2. Develop and estimate a complete model of both directions of causality. This is the idea behind many large macro models (e.g. Federal Reserve Bank-US). *This is extremely difficult in practice.*
3. Use instrumental variables regression to estimate the causal effect of interest (effect of X on Y, ignoring effect of Y on X)

Internal and external validity when forecasting

- Forecasting and estimation of causal effects are quite different objectives.
- For forecasting,
  - $\bar{R}^2$ matters (a lot!)
  - Omitted variable bias isn't a problem!
  - 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

Conclusion

Internal and External Validity

Potential threats to internal validity