Assessing Studies Based on Multiple Regression

Outline

1. Internal and External Validity
2. Threats to Internal Validity
   a. Omitted variable bias
   b. Functional form misspecification
   c. Errors-in-variables bias
   d. Missing data and sample selection bias
   e. Simultaneous causality bias
3. Application to Test Scores

A Framework for Assessing Statistical Studies: 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 studied to other populations
  - California in 2011? Massachusetts in 2011? Mexico in 2011?
  - Hard to say anything very specific...

Threats to Internal Validity of Multiple Regression Analysis

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 $E(u_i|X_{1i},\ldots,X_{ki}) \neq 0$ (or that conditional mean independence fails) – in which case OLS is biased and inconsistent. (Actually more to worry about -- what other assumptions might be violated?)

1. Omitted variable bias

Omitted variable bias arises if an omitted variable is ***both*:**

(i) a determinant of Y and

(ii) correlated with at least one included regressor.

- We first discussed omitted variable bias in regression with a single X. OV bias arises in multiple regression if the omitted variable satisfies conditions (i) and (ii) above.
- If the multiple regression includes control variables, then we need to ask whether there are omitted factors that are not adequately controlled for, that is, whether the error term is correlated with the variable of interest even after we have included the control variables.

Solutions to omitted variable bias

1. If you have data on one or more controls and they are adequate (in the sense of conditional mean independence plausibly holding for the causal variable of interest) then include the control variables;

---

2. Possibly, use *panel data* in which each entity (individual) is observed more than once;
3. If the omitted variable(s) cannot be measured, use *instrumental variables regression*;
4. Run a randomized controlled experiment.

2. Functional form misspecification bias

Solutions to functional form misspecification bias

1. Continuous dependent variable: use the "appropriate" nonlinear specifications in X (interactions, etc.)

   ---

2. Discrete dependent variable: need an extension of multiple regression methods ("probit" or "logit" analysis for binary dependent variables).

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?)

Solutions to errors-in-variables bias

1. Obtain better data (often easier said than done).
2. Develop a specific model of the measurement error process. This is only possible if a lot is known about the nature of the measurement error – for example a subsample of the data are cross-checked using administrative records and the discrepancies are analyzed and modeled. (Very specialized; we won't pursue this here.)

---

3. Instrumental variables regression.

Details of Errors-in-variables bias

Leads to correlation between the measured variable and the regression error. Consider the single-regressor model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

and suppose $E(u_i|X_i) = 0$). Let

X_i = unmeasured true value of X

$\tilde{X}_i$ = mis-measured version of X (the observed data)

Then

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1(X_i - \tilde{X}_i) + u_i]$$

So the regression you run is,

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i, \text{ where } \tilde{u}_i = \beta_1(X_i - \tilde{X}_i) + u_i$$

Classical measurement error:

$$\tilde{X}_i = X_i + v_i,$$

where v_i is mean-zero random noise with corr(X_i, v_i) = 0 and corr(u_i, v_i) = 0.

Under the classical measurement error model, $\hat{\beta}_1$ is biased towards zero. (Why? Imagine the variance of v approaching infinity.)

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 $\hat{\beta}_1$ is unbiased. (Actually 2 also can cause problems. Why? Think about non-linear models...)

Case 3 introduces "sample selection" bias.

Missing data: Case 1

Data are missing 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. This is equivalent to your having taken a simple random sample of 80 workers (think about it).

This inflates variance but doesn't cause bias.

Missing data: Case 2

Data are missing based on a value of one of the X's

In the test score/class size application, suppose you restrict your analysis to the subset of school districts with $STR < 20$. This is equivalent to having missing data, where the data are missing if $STR > 20$.

This inflates variance but doesn't cause bias.

Missing data: 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:

(i) influences the availability of data and

(ii) is related to the dependent variable.

Example #1: Height of undergraduates

Your stats prof asks you to estimate the mean height of undergraduate males. 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?

Example #3: returns to education

- What is the return to an additional year of education?
- Empirical strategy:
  - Sampling scheme: simple random sample of employed people (employed, so we have wage data)
  - Data: earnings and years of education
  - Estimator: regress ln(*earnings*) on *years_education*

What is the selection bias?

The Basic Point: ***Sample selection bias induces correlation between a regressor and the error term***.

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Cases with large X values are more likely to have received positive shocks (taller people, surviving funds, highly-educated people, ...)

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 unemployed as well as employed
- Randomized controlled experiment.
- Construct a model of the sample selection 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?

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*?

Simultaneous causality bias in equations

(a) Causal effect on Y of X: $Y_i = \beta_0 + \beta_1 X_i + u_i$

(b) Causal effect on X of Y: $X_i = \gamma_0 + \gamma_1 Y_i + v_i$

- Large u_i means large Y_i, *which implies* large X_i (if $\gamma_1 > 0$)
- Thus $\text{corr}(X_i, u_i) \neq 0$
- Thus $\hat{\beta}_1$ is biased and inconsistent.

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 bi-directional causality. *This is 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).

Brief Intro to Instrumental Variables Regression (Single *X* and *Z*)

- A valid instrument Z must satisfy two conditions:
  (1) *relevance*: $\text{corr}(Z_i, X_i) \neq 0$
  (2) *exogeneity*: $\text{corr}(Z_i, u_i) = 0$
- TSLS proceeds by first regressing X on Z to get $\hat{X}$, then regressing Y on $\hat{X}$
- The key idea is that the first stage isolates part of the variation in X that is uncorrelated with u
- If the instrument is valid, then the large-sample sampling distribution of the TSLS estimator is normal, so inference proceeds as usual

IV Regression: Test scores and class size

- The California test score/class size regressions still could have OV bias (e.g. parental involvement).
- In principle, this bias can be eliminated by IV regression (TSLS).
- IV regression requires a valid instrument, that is, an instrument that is:
  (1) relevant: corr(Z_i,STR_i) $\neq$ 0
  (2) exogenous: corr(Z_i,u_i) = 0

IV Regression: Test scores and class size, ctd.

Here is a (hypothetical) instrument:

- some districts, randomly hit by an earthquake, "double up" classrooms:

  $Z_i = Quake_i$ = 1 if hit by quake, = 0 otherwise

- *Do the two conditions for a valid instrument hold*?
- The earthquake makes it *as if* the districts were in a random assignment experiment. Thus, the variation in *STR* arising from the earthquake is exogenous.
- The first stage of TSLS regresses *STR* against *Quake*, thereby isolating the part of *STR* that is exogenous (the part that is "as if" randomly assigned)

Applying External and Internal Validity: Test Scores and Class Size

Objective: Assess the threats to the internal and external validity of the empirical analysis of the California test score data.

- External validity
  - Compare results for California and Massachusetts
  - Think hard…
- Internal validity
  - Go through the list of five potential threats to internal validity and think hard…

Check of external validity

We will compare the California study to one using Massachusetts data

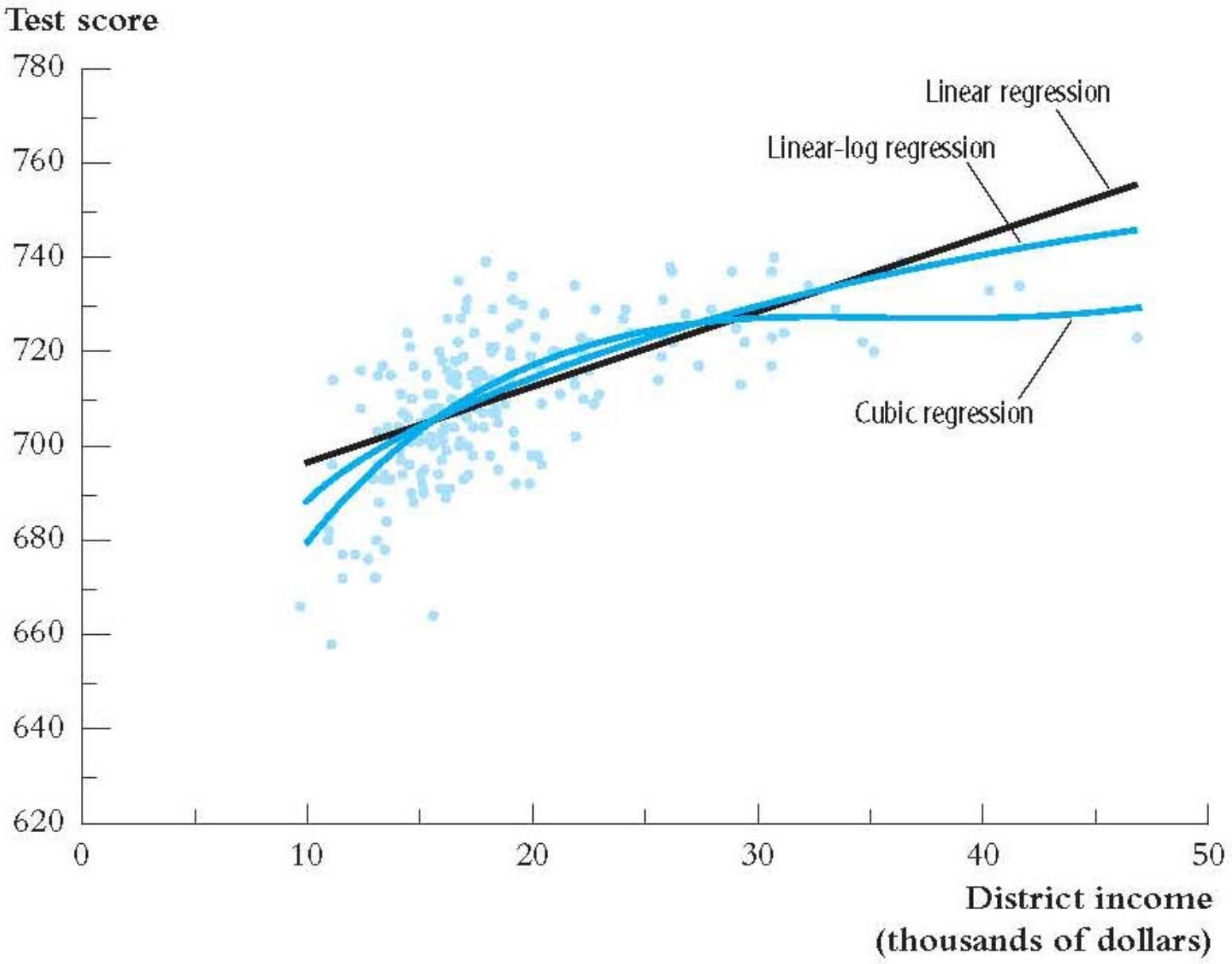
The Massachusetts data set

- 220 elementary school districts
- Test: 1998 MCAS test – fourth grade total (Math + English + Science)
- Variables: *STR*, *TestScore*, *PctEL*, *LunchPct*, *Income*

The Massachusetts data: summary statistics

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	



Test scores vs. Income & regression lines: Massachusetts data

TABLE 9.2 Multiple Regression Estimates of the Student–Teacher Ratio and Test Scores: Data from Massachusetts

Dependent variable: average combined English, math, and science test score in the school district, fourth grade; 220 observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
Student–teacher ratio (STR)	−1.72** (0.50)	−0.69* (0.27)	−0.64* (0.27)	12.4 (14.0)	−1.02** (0.37)	−0.67* (0.27)
STR^2				−0.680 (0.737)		
STR^3				0.011 (0.013)		
% English learners		−0.411 (0.306)	−0.437 (0.303)	−0.434 (0.300)		
% English learners > median? (Binary, $HiEL$)					−12.6 (9.8)	
$HiEL \times STR$					0.80 (0.56)	
% Eligible for free lunch		−0.521** (0.077)	−0.582** (0.097)	−0.587** (0.104)	−0.709** (0.091)	−0.653** (0.72)
District income (logarithm)		16.53** (3.15)				
District income			−3.07 (2.35)	−3.38 (2.49)	−3.87* (2.49)	−3.22 (2.31)
District income^2			0.164 (0.085)	0.174 (0.089)	0.184* (0.090)	0.165 (0.085)
District income^3			−0.0022* (0.0010)	−0.0023* (0.0010)	−0.0023* (0.0010)	−0.0022* (0.0010)
Intercept	739.6** (8.6)	682.4** (11.5)	744.0** (21.3)	665.5** (81.3)	759.9** (23.2)	747.4** (20.3)

(*Table 9.2 continued*)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

Regressor	(1)	(2)	(3)	(4)	(5)	(6)
All *STR* variables and interactions = 0				2.86 (0.038)	4.01 (0.020)	
$STR^2, STR^3 = 0$				0.45 (0.641)		
$Income^2, Income^3$			7.74 (< 0.001)	7.75 (< 0.001)	5.85 (0.003)	6.55 (0.002)
$HiEL, HiEL \times STR$					1.58 (0.208)	
SER	14.64	8.69	8.61	8.63	8.62	8.64
$\overline{R}^2$	0.063	0.670	0.676	0.675	0.675	0.674

These regressions were estimated using the data on Massachusetts elementary school districts described in Appendix 9.1. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. Individual coefficients are statistically significant at the *5% level or **1% level.

How do the Mass and California results compare?

- Logarithmic v. cubic function for *STR*?
- Evidence of nonlinearity in *TestScore-STR* relation?
- Is there a significant *HiEL*×*STR* interaction?

Predicted effects for a class size reduction of 2

Linear specification for Mass:

$$\widehat{TestScore} = 744.0 - 0.64STR - 0.437PctEL - 0.582LunchPct$$
$$(21.3)\ \ (0.27)\qquad (0.303)\qquad (0.097)$$

$$- 3.07Income + 0.164Income^2 - 0.0022Income^3$$
$$(2.35)\qquad (0.085)\qquad (0.0010)$$

- Estimated effect = -0.64×(-2) = 1.28

Computing predicted effects in nonlinear models

$$\widehat{TestScore} = 655.5 + 12.4STR - 0.680STR^2 + 0.0115STR^3$$
$$- 0.434PctEL - 0.587LunchPct$$
$$- 3.48Income + 0.174Income^2 - 0.0023Income^3$$

Estimated reduction from 20 students to 18:

$$\Delta\widehat{TestScore} = [12.4\times20 - 0.680\times20^2 + 0.0115\times20^3]$$
$$- [12.4\times18 - 0.680\times18^2 + 0.0115\times18^3] = 1.98$$

- compare with estimate from linear model of 1.28

Summary of Findings for Massachusetts

- Coefficient on *STR* falls from –1.72 to –0.69 when control variables for student and district characteristics are included – an indication that the original estimate contained omitted variable bias.
- The class size effect is statistically significant at the 1% significance level, after controlling for student and district characteristics
- No statistical evidence of nonlinearities in the *TestScore* – *STR* relation
- No statistical evidence of *STR* – *PctEL* interaction

Comparison of estimated class size effects: CA vs. MA

TABLE 9.3 Student–Teacher Ratios and Test Scores: Comparing the Estimates from California and Massachusetts

	OLS Estimate $\hat{\beta}_{STR}$	Standard Deviation of Test Scores Across Districts	Estimated Effect of Two Fewer Students per Teacher, In Units of: Points on the Test	Standard Deviations
California				
Linear: Table 8.3(2)	−0.73 (0.26)	19.1	1.46 (0.52)	0.076 (0.027)
Cubic: Table 8.3(7) *Reduce STR from 20 to 18*	—	19.1	2.93 (0.70)	0.153 (0.037)
Cubic: Table 8.3(7) *Reduce STR from 22 to 20*	—	19.1	1.90 (0.69)	0.099 (0.036)
Massachusetts				
Linear: Table 9.2(3)	−0.64 (0.27)	15.1	1.28 (0.54)	0.085 (0.036)

Standard errors are given in parentheses.

Summary: Comparison of California and Massachusetts Regression Analyses

- 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.

Step back: what are the remaining threats to internal validity in the test score/class size example?

1. Omitted variable bias?

What causal factors might be missing?

- 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

Omitted variable bias, ctd.

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

- Answering this requires using judgment.
- There is some evidence that the control variables are 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 need to be similar in the two data sets

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's 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.

4. Sample selection bias?

- Sample is all elementary public school districts (in California and in Mass.) – there are no missing data
- No reason to think that selection is a problem if we are interested only in public school students.

5. Simultaneous causality bias?

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