

Assessing Studies Based on Multiple Regression (SW Ch. 7)

Lecture 4 EC 471

Multiple regression has some key features:

- It provides an estimate of the change in Y for a unit change in X .
- It resolves the problem of omitted variables: an omitted variable can be measured and included in the model.
- It can handle nonlinear relationships between Y and the X 's.

Still, OLS might yield a *biased* estimate of a *causal* effect.

7-1

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 and setting studied to other populations and settings, where the “setting” refers to the legal, policy, and physical environment and related salient features.

7-3

Threats to External Validity

How far can we generalize class results to other California school districts?

- Differences in populations
 - California in 2005?
 - Massachusetts in 2005?
 - Mexico in 2005?
- Differences in settings
 - different legal requirements for education
 - different treatment of bilingual education
 - differences in teacher characteristics

Threats to Internal Validity of Multiple Regression Analysis (SW Section 7.2)

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

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 $E(u_i|X_{1i}, \dots, X_{ki}) \neq 0$.

7-5

2. Wrong functional form

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

Potential 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).

7-7

1. Omitted variable bias

Arises if an omitted variable *both* (i) affects Y and (ii) is correlated with at least one regressor.

Potential solutions to omitted variable bias

- If the variable can be measured, include it as a regressor in multiple regression.
- Possibly, use *panel data* in which the variable (individual) is observed more than once.
- If the variable cannot be measured, use *instrumental variables regression*;
- Run a randomized controlled trial.

3. Errors-in-variables bias

So far we have assumed that X_i is measured without error. In reality, economic data often have measurement errors.

- Data entry errors in administrative data.
- Recollection errors in survey data (e.g., “What is your current job?”)
- Ambiguous questions/problems (e.g., “What is your income last year?”)
- Intentionally false responses (e.g., “What is the current value of your car?” “How often do you drink and drive?”)

In general, measurement error in a regressor results in “errors-in-variables” bias.

Illustration: suppose

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

is “correct” in the sense that the three least squares assumptions hold (in particular $E(u_i|X_i) = 0$).

Let

X_i = unmeasured true value of X

\tilde{X}_i = imprecisely measured version of X

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

- If X_i is measured with error, \tilde{X}_i is in general correlated with \tilde{u}_i , so $\hat{\beta}_1$ is biased and inconsistent.
- It is possible to derive formulas for this bias, but they require making specific mathematical assumptions about the measurement error process (for example, that \tilde{u}_i and X_i are uncorrelated). Those formulas are special and particular, but the observation that measurement error in X results in bias is general.

Then

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

or

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

If \tilde{X}_i is correlated with \tilde{u}_i then $\hat{\beta}_1$

$$\begin{aligned} \text{cov}(\tilde{X}_i, \tilde{u}_i) &= \text{cov}(\tilde{X}_i, \beta_1(X_i - \tilde{X}_i) + u_i) \\ &= \beta_1 \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) \\ &= \beta_1 [\text{cov}(\tilde{X}_i, X_i) - \text{cov}(\tilde{X}_i, \tilde{X}_i)] \end{aligned}$$

because in general $\text{cov}(\tilde{X}_i, X_i) \neq \text{cov}(\tilde{X}_i, \tilde{X}_i)$

Potential solutions to errors-

- Obtain better data.
- Develop a specific model of the measurement error process.
- This is only possible if a lot of information about the nature of the measurement error process is available. For example, if administrative records and other data on the measurement error process are available, they can be analyzed and modeled. (We will pursue this here.)
- Instrumental variables regression

4. Sample selection bias

So far we have assumed simple random sampling of the population. In some cases, simple random sampling is thwarted because the sample, in effect, “selects itself.”

Sample selection bias arises when a selection process (i) influences the availability of data and (ii) that process is related to the dependent variable.

7-13

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

Mutual fund example:

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

Being a managed fund in the sample ($managed_fund_i = 1$) means that your return was better than failed managed funds, which are not in the sample – so $corr(managed_fund_i, u_i) \neq 0$.

7-15

Example #1: Mutual funds

- Do actively managed mutual “the-market” funds?
- Empirical strategy:
 - Sampling scheme: simple mutual funds available to date.
 - Data: returns for the prec
 - Estimator: average ten-ye mutual funds, minus ten-y
 - Is there sample selection b

Example #2: returns to education

- What is the return to an addit
- Empirical strategy:
 - Sampling scheme: simple **workers**
 - Data: earnings and years o
 - Estimator: regress $\ln(earn$
 - Ignore issues of omitted v measurement error – is th bias?

Potential solutions to sample selection bias

- Collect the sample in a way that avoids sample selection.
 - *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).

7-17

Simultaneous causality bias in equations

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

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

- 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.
- *Ex*: A district with particularly bad test scores given the *STR* (negative u_i) receives extra resources, thereby lowering its *STR*; so STR_i and u_i are correlated

7-19

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 *STR* have extra resources: as a result of this, they also have low *STR*
- What does this mean for a regression estimate of the effect of *STR*?

Potential solutions to simultaneous causality bias

- Randomized controlled experiment (participants chosen at random by the experimenter, no feedback from the outcome variable, perfect compliance).
- Develop and estimate a complete model of both directions of causality. This is often done in large macro models (e.g. Federal Reserve models). *This is extremely difficult in practice.*
- Use instrumental variables regression to estimate the causal effect of interest (effect of X on Y , or effect of Y on X).

Applying this Framework: Test Scores and Class Size (SW Chapter 7.3)

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...

7-21

The Massachusetts data: summary statistics

TABLE 7.1 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	\$7,226	\$18,747	\$5,808
Number of observations	420		220	
Year	1999		1998	

7-23

Check of external validity

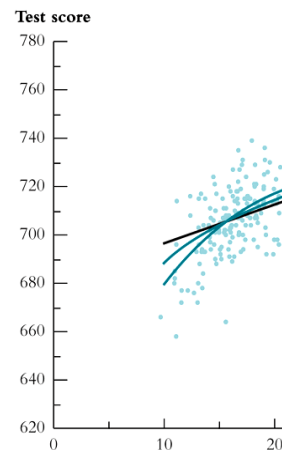
compare the California study
Massachusetts data

The Massachusetts data set

- 220 elementary school districts
- Test: 1998 MCAS test – four subjects (Math, Reading, English + Science)
- Variables: *STR*, *TestScore*, *P*

FIGURE 7.1 Test Scores vs. Income for Massachusetts Data

The estimated linear regression function does not capture the nonlinear relation between income and test scores in the Massachusetts data. The estimated linear-log and cubic regression functions are similar for district incomes between \$13,000 and \$30,000, the region containing most of the observations.



7-23

TABLE 7.2 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 × 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 7.2 continued)

(Table 7.2 continued)			
F-statistics and p-values Testing Exclusion of Groups of Variables			
	(1)	(2)	(3)
all STR variables and interactions = 0			
STR ² , STR ³ = 0			
Income ² , Income ³			7.74 (< 0.001)
HiEL, HiEL × STR			
SER	14.64	8.69	8.61
R ²	0.063	0.670	0.67

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

- Logarithmic v. cubic function
- Evidence of nonlinearity in STR
- Is there a significant $HiEL \times STR$ interaction?

Predicted effects for a class size reduction of 2

Linear specification for Mass:

$$\widehat{TestScore} = 744.0 - 0.64STR - 0.437PctEL - 0.582LunchPct - 3.07Income + 0.164Income^2 - 0.0022Income^3$$

(21.3) (0.27) (0.303) (0.097) (2.35) (0.085) (0.0010)

Estimated effect = $-0.64 \times (-2) = 1.28$

Standard error = $2 \times 0.27 = 0.54$

NOTE: $\text{var}(aY) = a^2\text{var}(Y)$; $SE(a\hat{\beta}_1) = |a|SE(\hat{\beta}_1)$

95% CI = $1.28 \pm 1.96 \times 0.54 = (0.22, 2.34)$

Computing predicted effects in nonlinear specification:

Use the “before” and “after” values:

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

Estimated reduction from 20 students to 18 students:

$$\Delta \widehat{TestScore} = [12.4 \times 20 - 0.680 \times 20 - 0.434 \times 20 - 0.587 \times 20 - 3.48 \times 20 + 0.174 \times 20^2 - 0.0023 \times 20^3] - [12.4 \times 18 - 0.680 \times 18 - 0.434 \times 18 - 0.587 \times 18 - 3.48 \times 18 + 0.174 \times 18^2 - 0.0023 \times 18^3]$$

- compare with estimate from linear specification
- SE of this estimated effect: use the “transform the regression” (“transform the regression”)

Summary of Findings for Massachusetts

1. 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.
2. The class size effect is statistically significant at the 1% significance level, after controlling for student and district characteristics
3. No statistical evidence on nonlinearities in the *TestScore* – *STR* relation
4. No statistical evidence of *STR* – *PctEL* interaction

7-29

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.

7-31

Comparison of estimated class

TABLE 7.3 Student-Teacher Ratios and Test Scores: Comparison of California and Massachusetts

	OLS Estimate $\hat{\beta}_{STR}$	Standard Deviation of Test Scores Across Districts
California		
Linear: Table 6.2(2)	-0.73 (0.26)	19.1
Cubic: Table 6.2(7) Reduce <i>STR</i> from 20 to 18	—	19.1
Cubic: Table 6.2(7) Reduce <i>STR</i> from 22 to 20	—	19.1
Massachusetts		
Linear: Table 7.2(3)	-0.64 (0.27)	15.1

Standard errors are given in parentheses.

Remaining threats to internal

What the CA v. MA comparison

1. Omitted variable bias

This analysis controls for:

- district demographics (income, race, etc.)
- some student characteristics (ability, etc.)

What is missing?

- Additional student characteristics (ability (but is this correlated with district demographics?))
- Access to outside learning opportunities
- Teacher quality (perhaps better in schools with lower *STR*)

Omitted variable bias, ctd.

- We have controlled for many relevant omitted factors;
- The nature of this omitted variable bias would need to be similar in California and Massachusetts to be consistent with these results;
- In this application we will be able to compare these estimates based on observational data with estimates based on experimental data – a check of this multiple regression methodology.

7-33

4. Selection

- Sample is all elementary public school districts (in California; in Mass.)
- no reason that selection should be a problem.

5. Simultaneous Causality

- 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.

7-35

2. Wrong functional form

- We have tried quite a few different functional forms in both the California and Massachusetts regressions
- Nonlinear effects are modest
- Plausibly, this is not a major source of bias

3. Errors-in-variables bias

- *STR* is a district-wide measure of student achievement
- Presumably there is some measurement error in the students who take the test might not be the measured *STR* for the district
- Ideally we would like data on individual student achievement at the grade level.

Summary

- Framework for evaluating regression analysis
 - Internal validity
 - External validity
- Five threats to internal validity
 1. Omitted variable bias
 2. Wrong functional form
 3. Errors-in-variables bias
 4. Sample selection bias
 5. Simultaneous causality bias
- Rest of course focuses on econometrics and addressing these threats.