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

Internal and External Validity

- Let's step back and take a broader look at regression. Is there a systematic way to assess (critique) regression studies? We know the strengths of multiple regression – but what are the pitfalls?
- We will list the most common reasons that multiple regression estimates, based on observational data, can result in biased estimates of the causal effect of interest.
- In the test score application, we'll try to address these threats as best we can – and assess what threats remain. After all this work, what have we learned about the effect on test scores of class size reduction?

A Framework for Assessing Statistical Studies: Internal and External Validity (SW Section 9.1)

- ***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 of Multiple Regression Studies

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 2019?
  - Massachusetts in 2019?
  - Mexico in 2019?
- Differences in settings
  - different legal requirements (e.g. special education)
  - different treatment of bilingual education
- Differences in teacher characteristics

Threats to Internal Validity of Multiple Regression Analysis (SW Section 9.2)

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

Five threats to the internal validity of regression studies:

- Omitted variable bias
- Wrong functional form
- Errors-in-variables bias
- Sample selection bias
- 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.

1. Omitted variable bias

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

1. a determinant of *Y* and
2. 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 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 (that is, if conditional mean independence plausibly holds), then include the control variables;
3. Possibly, use *panel data* in which each entity (individual) is observed more than once;
4. If the omitted variable(s) cannot be measured or adequately controlled for, use *instrumental variables regression*;
5. 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 $E(u|X = x) = 0$.

2. Wrong functional form (functional form misspecification)

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

Solutions to functional form misspecification

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

3. Errors-in-variables bias (1 of 4)

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?)

3. Errors-in-variables bias (2 of 4)

In general, measurement error in a regressor results in ***"errors-in-variables" bias***.

A bit of math shows that errors-in-variables typically 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)

3. Errors-in-variables bias (3 of 4)

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,

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i, \text{ where } \tilde{u}_1 = \beta_1 (X_i - \tilde{X}_i) + u_i$$

With measurement error, typically $\tilde{X}_i$ is correlated with $\tilde{u}_i$ so $\hat{\beta}_1$ is biased:

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

It is often plausible that $\text{cov}(\tilde{X}_i, u_i) = 0$ (if $E(u_i | X_i) = 0$ then $\text{cov}(\tilde{X}_i, u_i) = 0$ if the measurement error in $\tilde{X}_i$ is uncorrelated with u_i). But typically $\text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) \neq 0$….

3. Errors-in-variables bias (4 of 4)

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

$$\begin{aligned} \text{so} \quad \text{cov}(\tilde{X}_i, \tilde{u}_i) &= \text{cov}(\tilde{X}_i, \beta_1 (X_i - \tilde{X}_i) + u_i) \\ &= \beta_1 \, \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) + \text{cov}(, u_i) \\ &= \beta_1 \, \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) \quad \text{if} \quad \text{cov}(\tilde{X}_i, u_i) = 0 \end{aligned}$$

To get some intuition for the problem, consider two special cases:

A. Classical measurement error
B. "Best guess" measurement error

A. Classical measurement error

The classical measurement error model assumes that

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

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. Here's the idea: Suppose you take the true variable then add a huge amount of random noise – random numbers generated by the computer. In the limit of "all noise," $\tilde{X}_i$ will be unrelated to Y_i (and to everything else), so the regression coefficient will have expectation zero. If $\tilde{X}_i$ has some noise but isn't "all noise" then the relation between $\tilde{X}_i$ and Y_i will be attenuated, so $\hat{\beta}_1$ is biased towards zero.

Classical measurement error: the math

$$\tilde{X}_i = X_i + v_i, \text{ where } \text{corr}(X_i, v_i) = 0 \text{ and } \text{corr}(u_i, v_i) = 0.$$

Then $\text{var}(\tilde{X}_i) = \sigma_X^2 + \sigma_v^2$

$$\text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) = \text{cov}(X_i + v_i, -v_i) = -\sigma_v^2$$

so $\text{cov}(\tilde{X}_i, \tilde{u}_i) = -\beta_1 \sigma_v^2$

so $$\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \frac{\sigma_v^2}{\sigma_{\tilde{X}}^2} = \left(1 - \frac{\sigma_v^2}{\sigma_{\tilde{X}}^2}\right)\beta_1$$

$$= \left(\frac{\sigma_{\tilde{X}}^2 - \sigma_v^2}{\sigma_{\tilde{X}}^2}\right)\beta_1 = \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_v^2}\right)\beta_1$$

So $\hat{\beta}_1$ is biased towards zero.

The classical measurement error model is special because it assumes corr(X_i, v_i) = 0.

B. "Best Guess" measurement error (1 of 2)

Suppose the respondent doesn't remember X_i, but uses another variable W to make a best guess of the form $\tilde{X}_i = E(X_i|W_i)$, where $E(u_i|W_i) = 0$. Then,

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

- $\text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) = 0$ because $\tilde{X}_i = E(X_i|W_i)$ (because $\tilde{X}_i$ is the best guess, the error $X_i - \tilde{X}_i$ is uncorrelated with $\tilde{X}_i$).
- $\text{cov}(\tilde{X}_i, u_i) = 0$ because $E(u_i|W_i) = 0$ ($\tilde{X}_i$ is a function of W_i and by assumption, $E(u_i|W_i) = 0$).
- Thus $\text{cov}(\tilde{X}_i, \tilde{u}_i) = 0$, so $\hat{\beta}_1$ is unbiased.

B. "Best Guess" measurement error (2 of 2)

- Under the "Best Guess" model, you still have measurement error – you don't observe the true value of X_i – but there this measurement error doesn't introduce bias into $\hat{\beta}_1$!
- The "best guess" model is extreme – it isn't enough to make a good guess, you need the "best" guess $\tilde{X}_i = E(X_i|W_i)$, that is, the conditional expectation of X given W, where $E(u_i|W_i) = 0$, and moreover W_i must be uncorrelated with u_i

Lessons from the classical and best-guess models:

- The amount of bias in $\hat{\beta}_1$ depends on the nature of the measurement error – these models are two special cases.
- If there is pure noise added to X_i, then $\hat{\beta}_1$ is biased towards zero.
- The "best guess" model is extreme. In general, if you think there is measurement error, you should worry about measurement error bias.
- The potential importance of measurement error bias depends on how the data are collected.
  - Some administrative data (e.g. number of teachers in a school district) are often quite accurate.
  - Survey data on sensitive questions (how much do you earn?) often have considerable measurement error.

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.

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.

Case 3 introduces "sample selection" bias.

Missing data: Case 1

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), so your dog didn't introduce any bias.

Missing data: Case 2

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 $STR < 20$. 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.

Missing data: Case 3

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?

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($managed\_\ fund_i, u_i$) ≠ 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?
- 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_education*
  - Ignore issues of omitted variable bias and measurement error – is there sample selection bias?
  - How does this relate to the basketball player example?

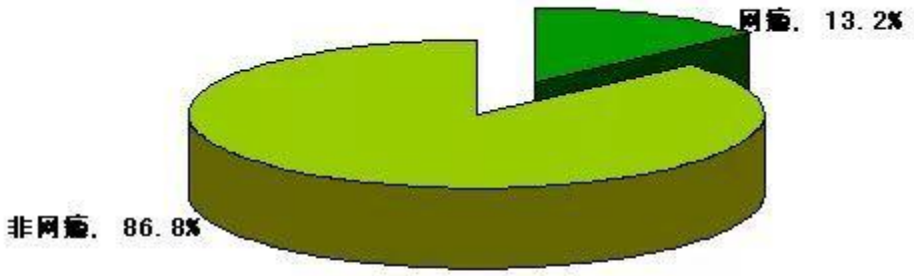
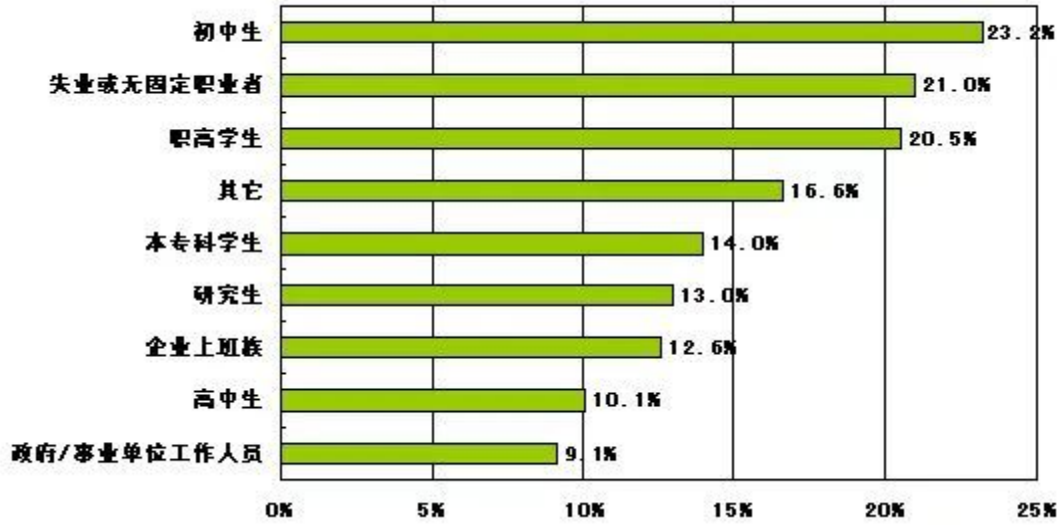


图 2-1：青少年网民中网瘾的比例

另外，根据本次网络调查的结果，这一比例还要更高，达到16.6%。



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?

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 corr(X_i, u_i) $\neq 0$

- Thus $\hat{\beta}_1$ is biased and inconsistent.
- *Example*: 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

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 the Regression is Used for Prediction (SW Section 9.3)

- Prediction and estimation of causal effects are quite different objectives.
- For prediction:
  - The data used to estimate the prediction model must be from the same distribution as the out-of-sample observation for which the prediction is made. This is an external validity requirement for a prediction model.
  - The predictors should be ones that substantially contribute to explaining the variation in *Y*. They do not need to have direct causal interpretations, and the regression coefficients in general need not estimate causal effects.
  - The estimator must be one that produces reliable out-of-sample predictions. When the number of regressors (predictors) is small relative to the number of observations, OLS can be used. But when the number of predictors is large, there are better estimators than OLS – ones developed specially for the prediction problem.

Applying External and Internal Validity: Test Scores and Class Size (SW Section 9.4)

- 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

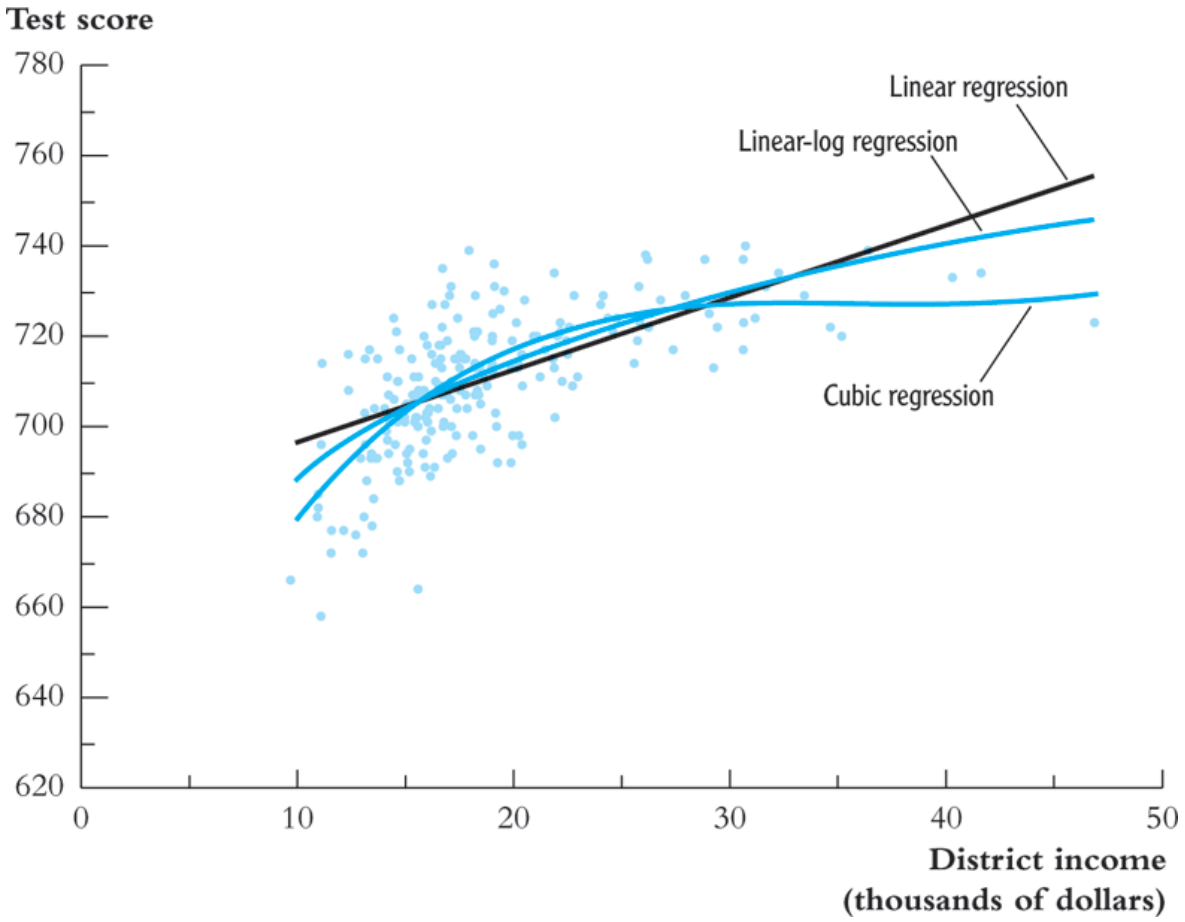
- 220 elementary school districts
- Test: 1998 MCAS test – fourth grade total (Math + English + Science)
- Variables: *STR*, *TestScore*, *PctEL*, *LunchPct*, *Income*

The Massachusetts data: summary statistics (1 of 5)

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	

The Massachusetts data: summary statistics (2 of 5)



Test scores vs. Income & regression lines: Massachusetts data

The Massachusetts data: summary statistics (3 of 5)

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) [−2.70, −0.73]	−0.69 (0.27) [−1.22, −0.16]	−0.64 (0.27) [−1.17, −0.11]	12.4 (14.0)	−1.02 (0.37)	−0.67 (0.27) [−1.21, −0.14]
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)

The Massachusetts data: summary statistics (4 of 5)

TABLE 9.2 (*Continued*)

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)

The Massachusetts data: summary statistics (5 of 5)

TABLE 9.2 (*Continued*)

***F*-Statistics and *p*-Values Testing Exclusion of Groups of Variables**						
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* × *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. All regressions include an intercept (not reported). Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. 95% confidence intervals for the coefficient on the student-teacher ratio are presented in brackets for regressions (1), (2), (3), and (6), but not for the regressions with nonlinear terms in *STR*.

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:

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

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

- Estimated effect = –0.64 × (–2) = 1.28
- Standard error = 2 × 0.27 = 0.54

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

- 95% CI = 1.28 ± 1.96 × 0.54 = (0.22, 2.34)

Computing predicted effects in nonlinear models *Use the "before" and "after" method:*

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

Estimated reduction from 20 students to 18:

$$\Delta TestScore = [12.4 \times 20 - 0.680 \times 20^2 + 0.0115 \times 20^3]$$
$$- [12.4 \times 18 - 0.680 \times 18^2 + 0.0115 \times 18^3] = 1.98$$

- compare with estimate from linear model of 1.28
- *SE* of this estimated effect: use the "rearrange the regression" ("transform the regressors") method

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 on 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 β_{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) [0.46, 2.48]	0.076 (0.027) [0.024, 0.130]
Cubic: Table 8.3(7) *Reduce STR from 20 to 18*	–	19.1	2.93 (0.70) [1.56, 4.30]	0.153 (0.037) [0.081, 0.226]
Cubic: Table 8.3(7) *Reduce STR from 22 to 20*	–	19.1	1.90 (0.69) [0.54, 3.26]	0.099 (0.036) [0.028, 0.171]
Massachusetts				
Linear: Table 9.2(3)	−0.64 (0.27)	15.1	1.28 (0.54) [0.22, 2.34]	0.085 (0.036) [0.015, 0.154]

Standard errors are given in parentheses. 95% confidence intervals for the effect of a two-student reduction are given in brackets.

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 of 4)

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

Step back: what are the remaining threats to internal validity in the test score/class size example? (2 of 4)

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 might be doing their job:
  - 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
- *What additional control variables might you want to use – and what would they be controlling for?*

Step back: what are the remaining threats to internal validity in the test score/class size example? (3 of 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'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.

Step back: what are the remaining threats to internal validity in the test score/class size example? (4 of 4)

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.

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.