Lecture Notes #7
(Chapter 9)
Economics 120B
Econometrics

Prof. Dahl

UC San Diego

Assessing Studies Based on Multiple Regression (SW Chapter 9)

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 – but what are the pitfalls of multiple regression?
- When we put all this together, what have we learned about the effect on test scores of class size reduction?

Is there a systematic way to assess regression studies?

Multiple regression has some key virtues:

- It provides an estimate of the effect on Y of arbitrary changes ΔX.
- It resolves the problem of omitted variable bias, if an omitted variable can be measured and included.
- It can handle nonlinear relations (effects that vary with the X's)

Still, OLS might yield a *biased* estimator of the true *causal* effect – it might not yield "valid" inferences…

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

How far can we generalize class size results from California school districts?

- Differences in populations
  - California in 2005?
  - Massachusetts in 2005?
  - Mexico in 2005?
- Differences in settings
  - different legal requirements concerning 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:

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 $E(u_i|X_{1i},\ldots,X_{ki}) \neq 0$ – in which case OLS is biased and inconsistent.

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*, but OV bias will arise when there are multiple *X*'s as well, if the omitted variable satisfies conditions (i) and (ii) above.

Potential solutions to omitted variable bias

1. If the variable can be measured, include it as an additional regressor in multiple regression;
2. Possibly, use *panel data* in which each entity (individual) is observed more than once;
3. If the variable cannot be measured, use *instrumental variables regression*;
4. 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

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

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

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 problems (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?)

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

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

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

But $\tilde{X}_i$ typically is correlated with $\tilde{u}_i$ so $\hat{\beta}_1$ is biased:

$$\begin{aligned} \text{cov}(\tilde{X}_i, \tilde{u}_i) &= \text{cov}(\tilde{X}_i, \beta_1(X_i - \tilde{X}_i) + u_i) \\ &= \beta_1 \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) + \text{cov}(\tilde{X}_i, u_i) \\ &= \beta_1[\text{cov}(\tilde{X}_i, X_i) - \text{var}(\tilde{X}_i)] + 0 \neq 0 \end{aligned}$$

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

“Errors-in-variables” bias, ctd.

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

- 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.

Potential solutions to errors-in-variables bias

1. Obtain better data.
2. Develop a specific model of the measurement error process.
3. 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.)
4. 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.

Example #1: 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?

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.

Example #2: 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?

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).

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.
- *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

Potential solutions to simultaneous causality bias

1. 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 Forecasting
(SW Section 9.3)

- 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
  - More on forecasting when we take up time series data

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

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	

FIGURE 9.1 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.

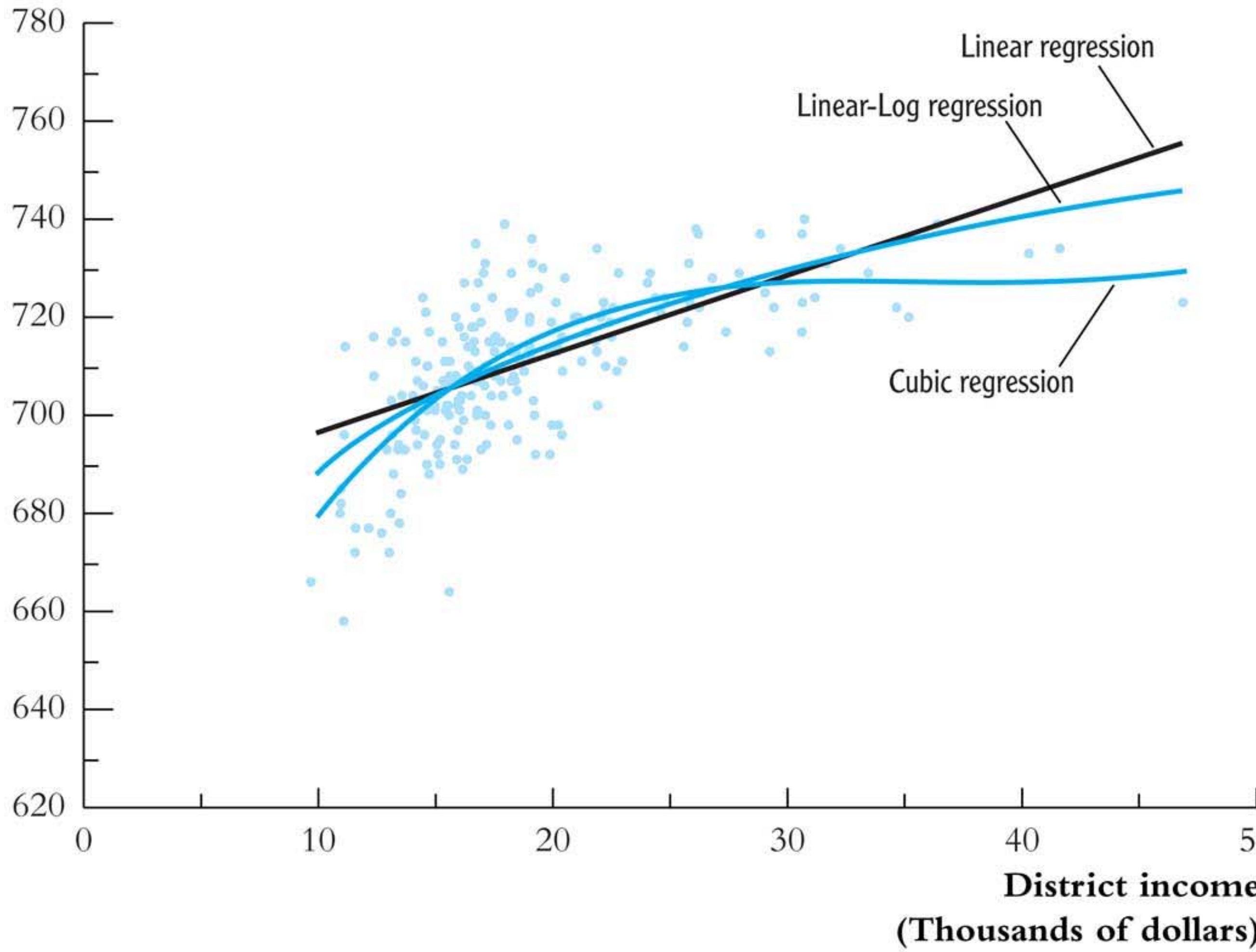

TABLE 9.2 **Multiple Regression Estimates of the Student–Teacher Ratio and Test Scores: Data from Massachusetts**

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)

(Table 9.2 continued)

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

	(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
$\bar{R}^2$	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:

$$\widehat{TestScore} = \underset{(21.3)}{744.0} - \underset{(0.27)}{0.64} * STR - \underset{(0.303)}{0.437} * PctEL$$

$$-\underset{(0.097)}{0.582} * LunchPct - \underset{(2.35)}{3.07} * Income$$

$$+\underset{(0.085)}{0.164} * Income^2 - \underset{(0.0010)}{0.0022} * Income^3$$

- Estimated effect = -0.64×(-2) = 1.28
- Standard error = 2×0.27 = 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 ± 1.96×0.54 = (0.22, 2.34)

Computing predicted effects in nonlinear models

Use the "before" and "after" method:

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

Estimated reduction from 20 students to 18:

$$\Delta\widehat{TestScore} = [12.4 * 20 - 0.680 * 20^2 + 0.0115 * 20^3]$$
$$-[12.4 * 18 - 0.680 * 18^2 + 0.0115 * 18^3]$$
$$= 1.98$$

- 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 $\hat{\beta}_{STR}$	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?

Omitted variable bias?

This analysis controls for:

- district demographics (income)
- some student characteristics (English speaking)

What is missing?

- Additional student characteristics, for example native ability (but is this correlated with *STR*?)
- Access to outside learning opportunities
- Teacher quality (perhaps better teachers are attracted to 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.

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?**
   - *STR* is a district-wide measure
   - Presumably there is some measurement error – students who take the test might not have experienced the measured *STR* for the district
   - Ideally we would like data on individual students, by grade level.

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.

Additional example for class discussion

Does appearing on *America's Most Wanted* TV show increase your chance of being caught?

> ***reference***: Thomas Miles (2005), "Estimating the Effect of *America's Most Wanted*: A Duration Analysis of Wanted Fugitives," *Journal of Law and Economics*, 281-306.

- Observational unit: Fugitive criminals
- Sampling scheme: 1200 male fugitives, from FBI, NYCPD, LAPD, PhilaPD, USPS, Federal Marshalls Web sites (*all data were downloaded from the Web!*)
- Dependent variable: length of spell (years until capture)
- Regressors:
  - Appearance on *America's Most Wanted* (175 of the 1200) (Fox, Saturdays, 9pm)
  - type of offence, personal characteristics

America's Most Wanted: Threats to Internal and External Validity

External validity: what would you want to extrapolate the results to – having the show air longer? putting on a second show of the same type? Be precise….

Internal validity: how important are these threats?

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

Anything else?