Hypothesis Tests and Confidence Intervals in Multiple Regression (SW Ch. 7)
Part 2

Dragos Ailoae
dailoae@gradcenter.cuny.edu

Advanced Economics and Business Statistics
ECON-4400w

Brooklyn College

Outline

1. Hypothesis tests and confidence intervals for one coefficient
2. Joint hypothesis tests on multiple coefficients
3. Other types of hypotheses involving multiple coefficients
4. **Model specification: how to decide which variables to include in a regression model**

Model specification: How to decide what variables to include in a regression (Section 7.5) (1 of 2)

1. Identify the variable of interest
2. Think of the omitted causal effects that could result in omitted variable bias
3. Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables.
   - The control variables are effective if the conditional mean independence assumption plausibly holds, that is, if *u* is uncorrelated with *STR* once the control variables are included.
   - This results in a "base" or "benchmark" model.

Model specification: How to decide what variables to include in a regression (Section 7.5) (2 of 2)

4. Also specify a range of plausible alternative models, which include additional candidate variables.

5. Estimate your base model and plausible alternative specifications (“sensitivity checks”).
   - Does a candidate variable change the coefficient of interest (β_1)?
   - Is a candidate variable statistically significant?
   - Use judgment, not a mechanical recipe…
   - Don’t just try to maximize R^2!

Digression about measures of fit…

It is easy to fall into the trap of maximizing the R^2 and $\bar{R}^2$, but this loses sight of our real objective, an unbiased estimator of the class size effect.

- A high R^2 (or $\bar{R}^2$) means that the regressors explain the variation in Y.
- A high R^2 (or $\bar{R}^2$) does *not* mean that you have eliminated omitted variable bias.
- A high R^2 (or $\bar{R}^2$) does *not* mean that you have an unbiased estimator of a causal effect (β_1).
- A high R^2 (or $\bar{R}^2$) does *not* mean that the included variables are statistically significant – this must be determined using hypotheses tests.

Analysis of the Test Score Data Set (SW Section 7.6) (1 of 3)

1. Identify the variable of interest:
   STR

2. Think of the omitted causal effects that could result in omitted variable bias (*Whether the students know English; outside learning opportunities; parental involvement; teacher quality and if teacher salary is correlated with district wealth – there is a long list!*)

Variables we would like to see in the data set

School characteristics:

- student-teacher ratio
- teacher quality
- computers (non-teaching resources) per student
- measures of curriculum design…

Student characteristics:

- English proficiency
- availability of extracurricular enrichment
- home learning environment
- parent's education level…

Variables actually in the data set

- student-teacher ratio (STR)
- percent English learners in the district (PctEL)
- percent eligible for subsidized/free lunch
- percent on public income assistance
- average district income
- expenditures per pupil
- number of computers

https://rdrr.io/cran/AER/man/CASchools.html

Analysis of the Test Score Data Set (SW Section 7.6) (2 of 3)

3. Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables. The control variables are effective if the conditional mean independence assumption plausibly holds (if *u* is uncorrelated with *STR* once the control variables are included). This results in a "base" or "benchmark" model.

   Many of the omitted causal variables are hard to measure, so we need to find control variables. These include PctEL (both a control variable and an omitted causal factor) and measures of district wealth.

Analysis of the Test Score Data Set (SW Section 7.6) (3 of 3)

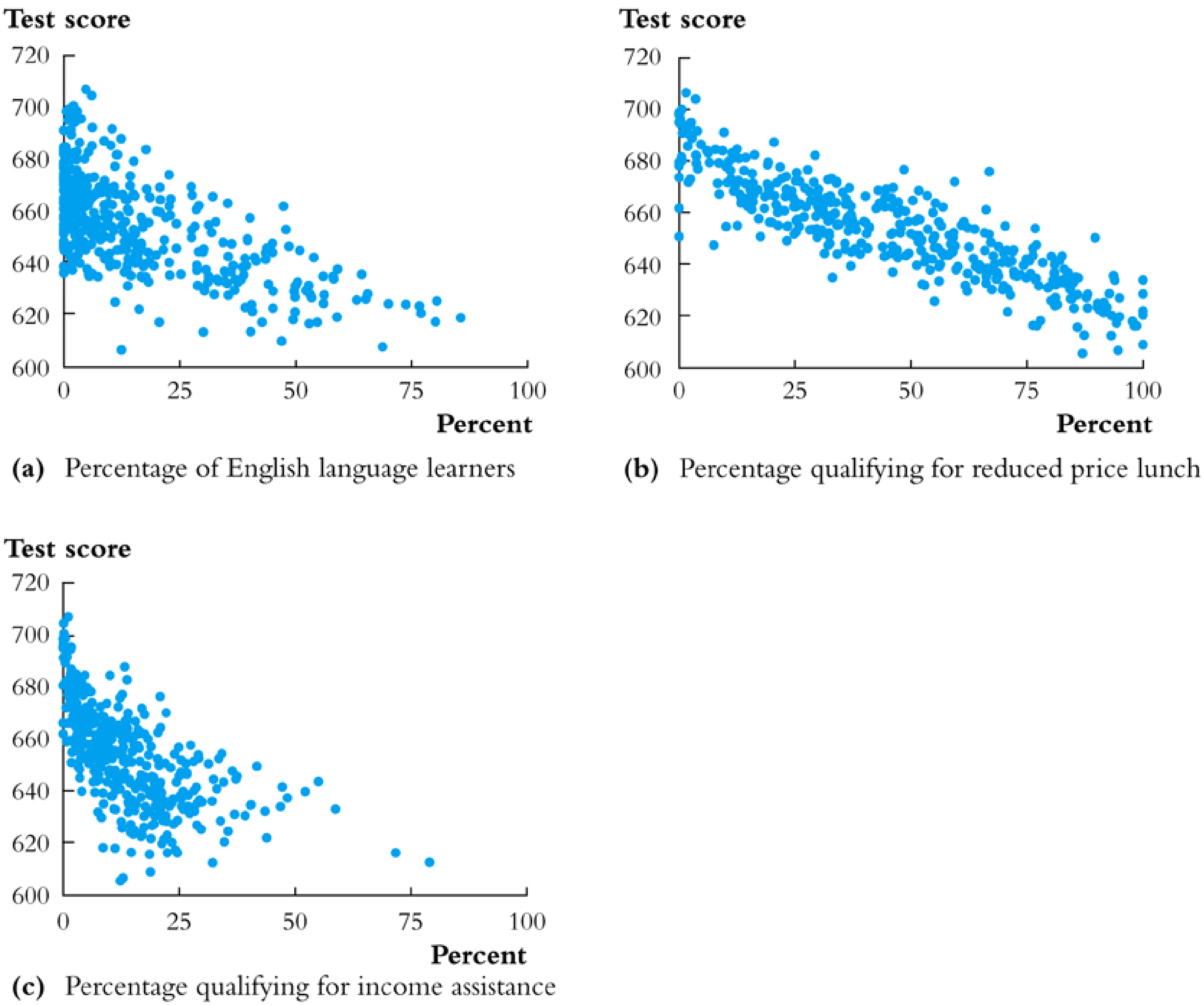
4. Also specify a range of plausible alternative models, which include additional candidate variables.

   It isn't clear which of the income-related variables will best control for the many omitted causal factors such as outside learning opportunities, so the alternative specifications include regressions with different income variables. The alternative specifications considered here are just a starting point, not the final word!

5. Estimate your base model and plausible alternative specifications ("sensitivity checks").

Test scores and California socioeconomic data…

FIGURE 7.2 Scatterplots of Test Scores vs. Three Student Characteristics



(a) Percentage of English language learners

(b) Percentage qualifying for reduced price lunch

(c) Percentage qualifying for income assistance

Presentation of regression results (1 of 2)

- We have a number of regressions and we want to report them. It is awkward and difficult to read regressions written out in equation form, so instead it is conventional to report them in a table.
- A table of regression results should include:
  - estimated regression coefficients
  - standard errors
  - measures of fit
  - number of observations
  - relevant *F*-statistics, if any
  - any other pertinent information, such as confidence intervals for the causal effect of interest
- Find this information in the following table!

Presentation of regression results (2 of 2)

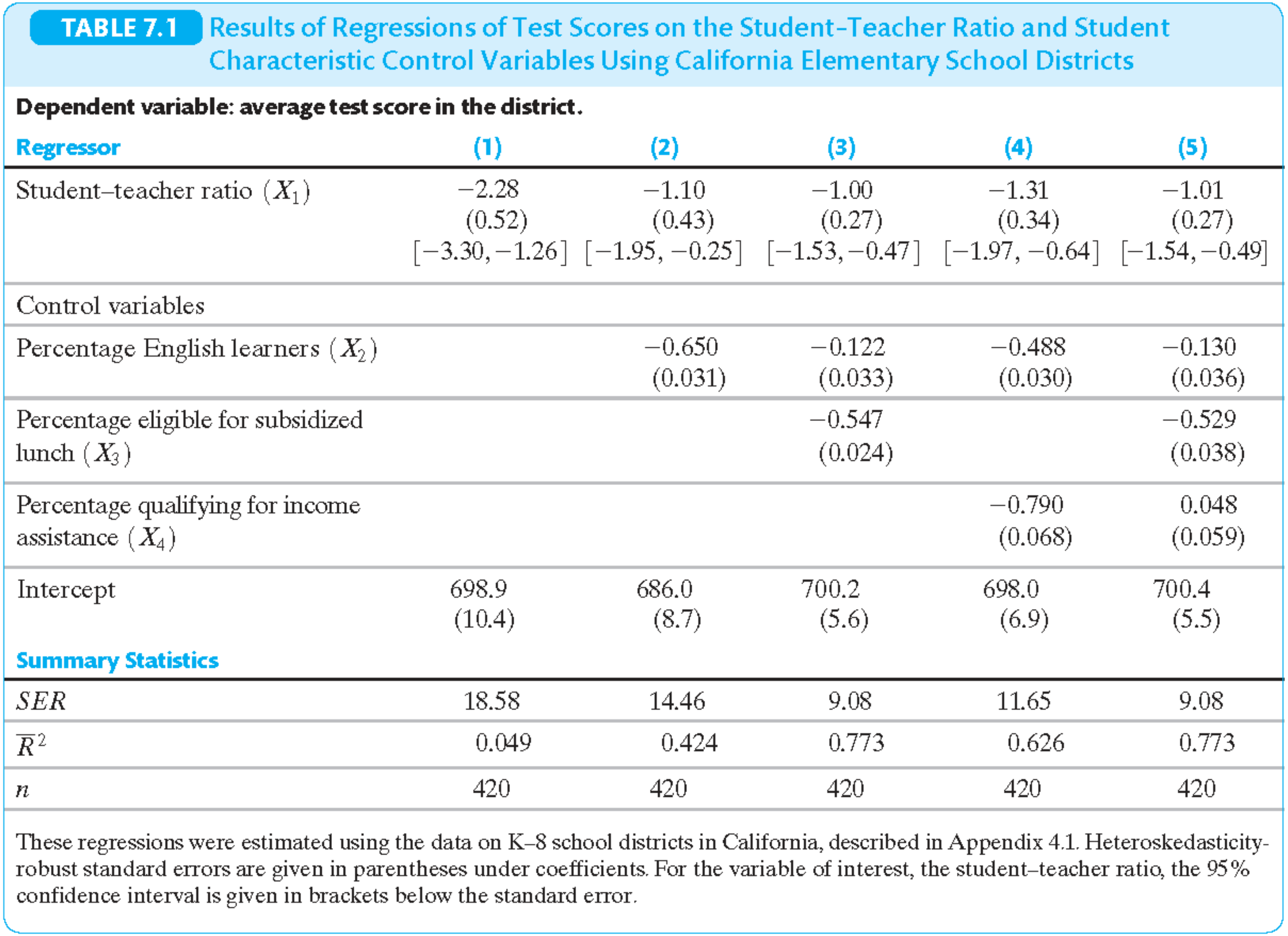


TABLE 7.1 Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28 (0.52) [−3.30, −1.26]	−1.10 (0.43) [−1.95, −0.25]	−1.00 (0.27) [−1.53, −0.47]	−1.31 (0.34) [−1.97, −0.64]	−1.01 (0.27) [−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650 (0.031)	−0.122 (0.033)	−0.488 (0.030)	−0.130 (0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547 (0.024)		−0.529 (0.038)
Percentage qualifying for income assistance (X_4)				−0.790 (0.068)	0.048 (0.059)
Intercept	698.9 (10.4)	686.0 (8.7)	700.2 (5.6)	698.0 (6.9)	700.4 (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

Summary: Multiple Regression

- Multiple regression allows you to estimate the effect on Y of a change in X_1, holding other included variables constant.
- If you can measure a variable, you can avoid omitted variable bias from that variable by including it.
- If you can't measure the omitted variable, you still might be able to control for its effect by including a control variable.
- There is no simple recipe for deciding which variables belong in a regression – you must exercise judgment.
- One approach is to specify a base model – relying on *a-priori* reasoning – then explore the sensitivity of the key estimate(s) in alternative specifications.