Confidence Sets for Multiple Coefficients (SW Section 7.4)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1,\ldots,n$$

What is a *joint* confidence set for β_1 and β_2?

A 95% ***joint confidence set*** is:

- A set-valued function of the data that contains the true parameter(s) in 95% of hypothetical repeated samples.
- The set of parameter values that cannot be rejected at the 5% significance level.
- You can find a 95% confidence set as the set of (β_1, β_2) that cannot be rejected at the 5% level using an *F*-test (*why not just combine the two 95% confidence intervals?*).

Joint confidence sets ctd.

Let $F(\beta_{1,0},\beta_{2,0})$ be the F-statistic testing the hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$:

95% confidence set = $\{\beta_{1,0}, \beta_{2,0}\colon\ F(\beta_{1,0}, \beta_{2,0}) < 3.00\}$

- 3.00 is the 5% critical value of the $F_{2,\infty}$ distribution
- This set has coverage rate 95% because the test on which it is based (the test it "inverts") has size of 5%

  5% of the time, the test incorrectly rejects the null when the null is true, so 95% of the time it does not; therefore the confidence set constructed as the nonrejected values contains the true value 95% of the time (in 95% of all samples).

The confidence set based on the F-statistic is an ellipse

$$\{\beta_1, \beta_2: \ F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) \le 3.00\}$$

Now

$$F = \frac{1}{2(1-\hat{\rho}_{t_1,t_2}^2)} \times \left[t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2\right]$$

$$= \frac{1}{2(1-\hat{\rho}_{t_1,t_2}^2)} \times$$

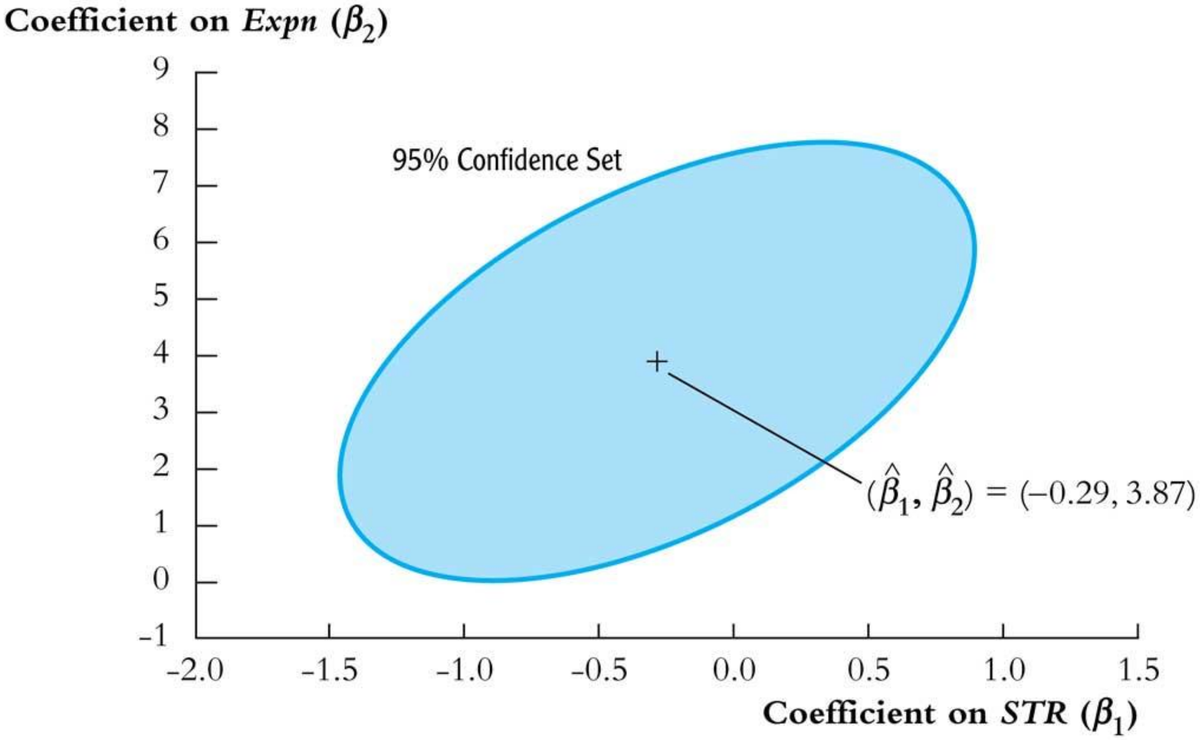
$$\left[\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)^2 + \left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)^2 + 2\hat{\rho}_{t_1,t_2}\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)\right]$$

This is a quadratic form in $\beta_{1,0}$ and $\beta_{2,0}$ – thus the boundary of the set F = 3.00 is an ellipse.

Confidence set based on inverting the F-statistic

FIGURE 7.1 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)

The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (β_2) is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the *F*-statistic at the 5% significance level.



An example of a multiple regression analysis – and how to decide which variables to include in a regression…

A Closer Look at the Test Score Data

(SW Sections 7.5 and 7.6)

We want to get an unbiased estimate of the effect on test scores of changing class size, holding constant student and school characteristics (but not necessarily holding constant the budget (*why*?)).

To do this we need to think about what variables to include and what regressions to run – and we should do this before we actually sit down at the computer. This entails thinking beforehand about your *model specification*.

A general approach to variable selection and "*model specification*"

- Specify a "base" or "benchmark" model.
- Specify a range of plausible alternative models, which include additional candidate variables.
- 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…

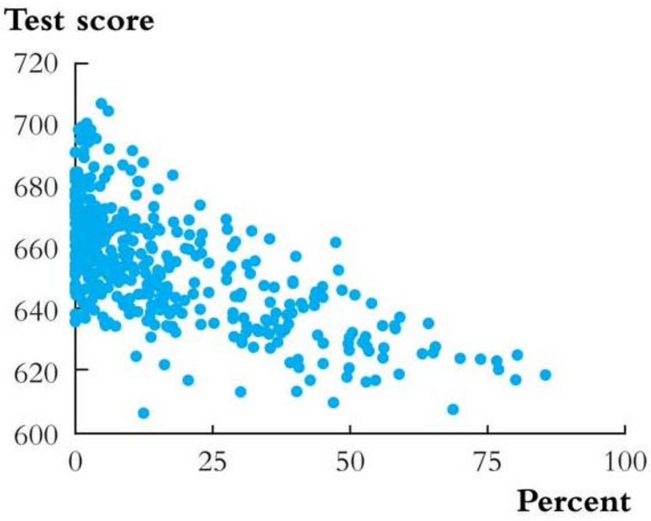
It is easy to fall into the trap of maximizing the R^2 and $\bar{R}^2$ – but this loses sight of our real objective, an unbiased estimator of the class size effect.

- 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.

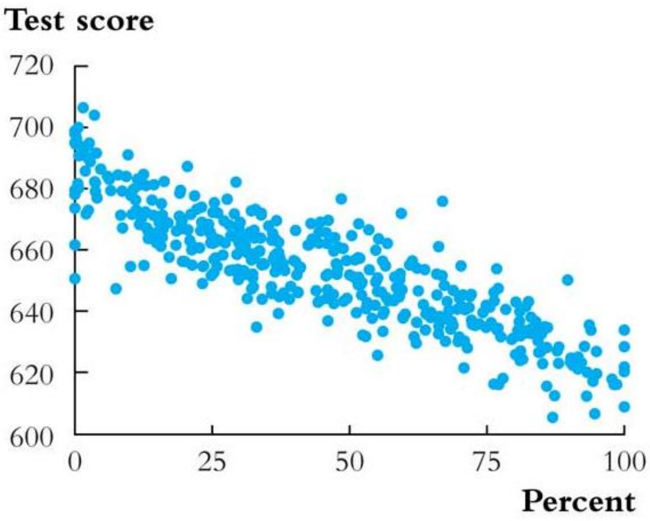
Back to the test score application:

- *What variables would you want – ideally – to estimate the effect on test scores of STR using school district data?*
- *Variables actually in the California class size data set*:
  - student-teacher ratio (*STR*)
  - percent English learners in the district (*PctEL*)
  - school expenditures per pupil
  - name of the district (so we could look up average rainfall, for example)
  - percent eligible for subsidized/free lunch
  - percent on public income assistance
  - average district income
- *Which of these variables would you want to include?*

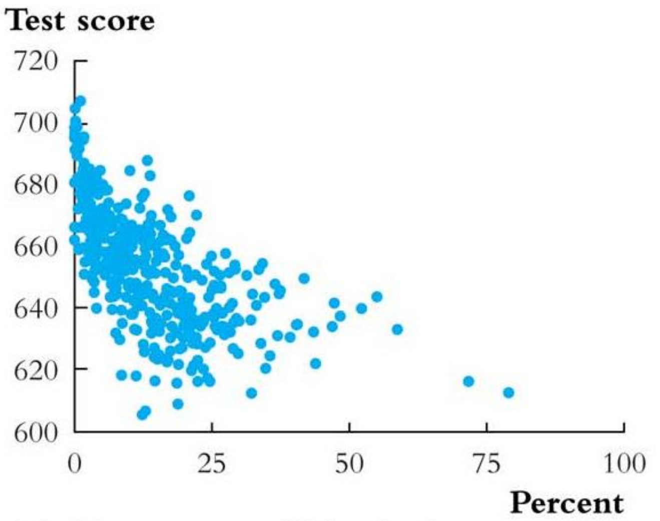
More California data…



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



(c) Percentage qualifying for income assistance

Digression on presentation of regression results

- 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.

Find this information in the following table:

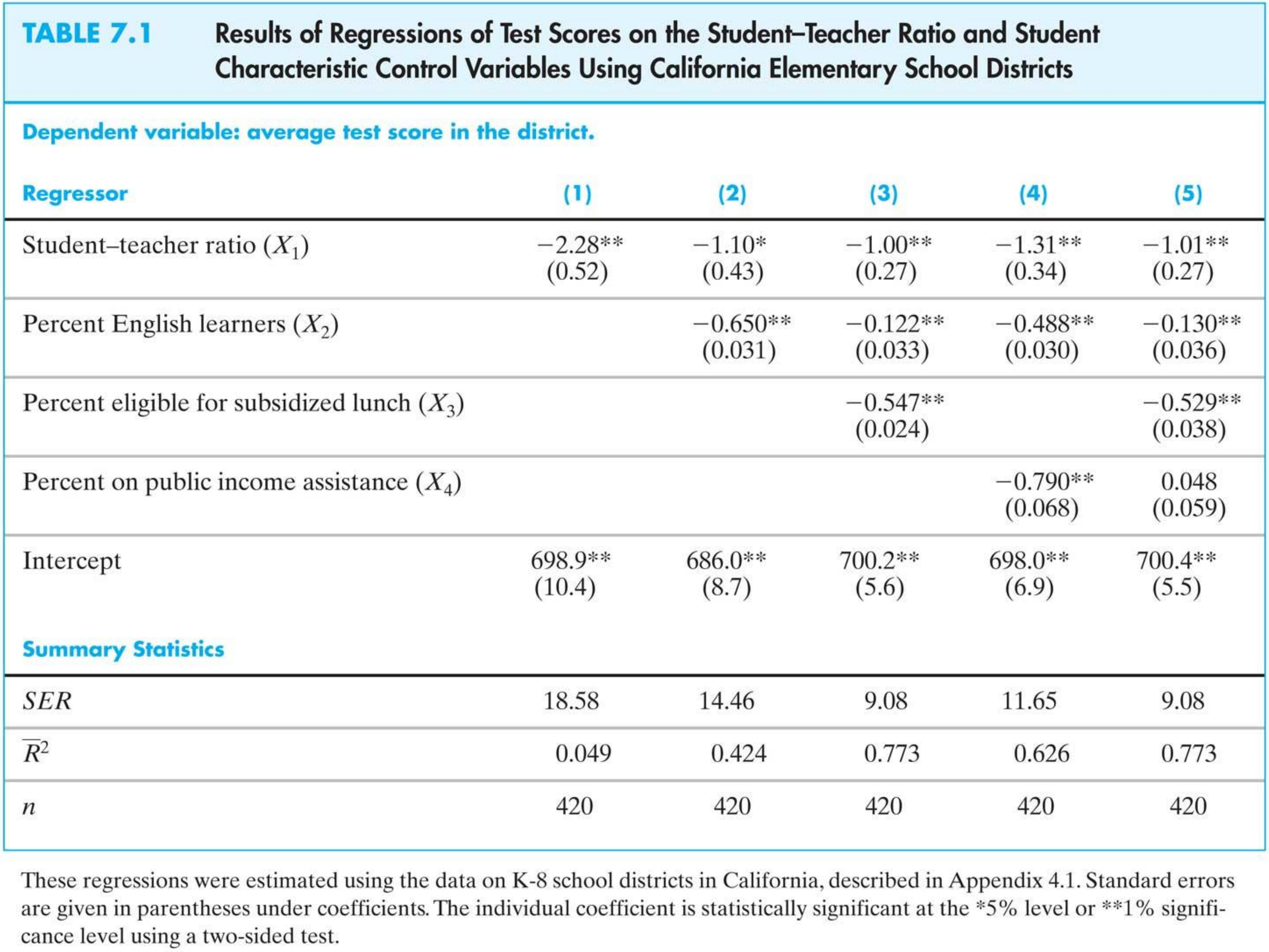


TABLE 7.1 **Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts**

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public 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
$\overline{R}^2$	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. Standard errors are given in parentheses under coefficients. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

Summary: Multiple Regression

- Multiple regression allows you to estimate the effect on Y of a change in X_1, holding X_2 constant.
- If you can measure a variable, you can avoid omitted variable bias from that variable by including it.
- 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.