Empirical problem: Class size and educational output

- Policy question: What is the effect of reducing class size by one student per class? by 8 students/class?
- What is the right output (performance) measure?
  - performance on standardized tests

The California Test Score Data Set

All K-6 and K-8 California school districts (n = 420)

Variables:

- 5th grade test scores (Stanford-9 achievement test, combined math and reading), district average
- Student-teacher ratio (STR) = no. of students in the district divided by no. full-time equivalent teachers
- Object of policy interest: $\frac{\Delta \text{Test score}}{\Delta STR}$

Do districts with smaller classes (lower STR) have higher test cores?

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: the sample correlation is –0.23.

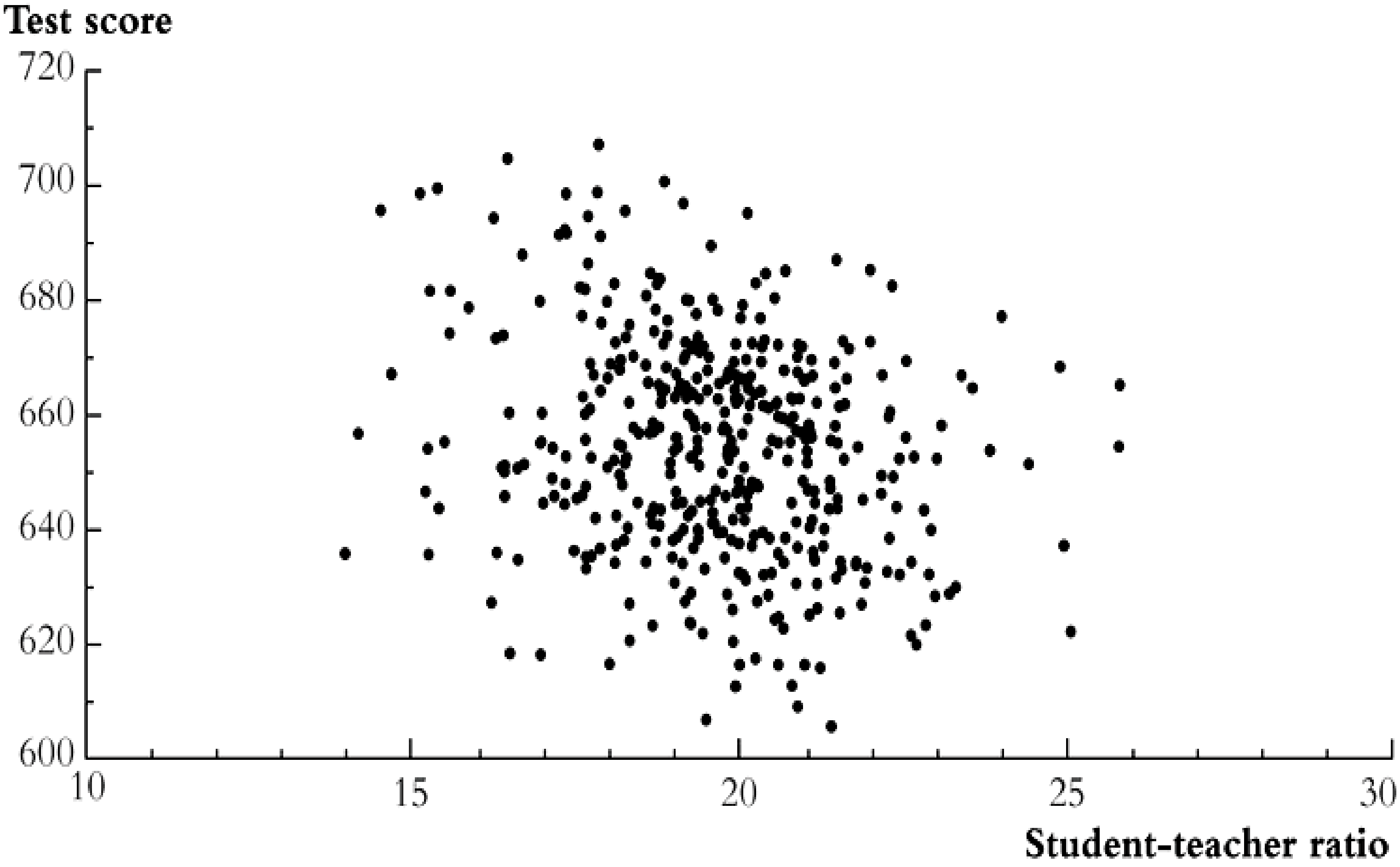
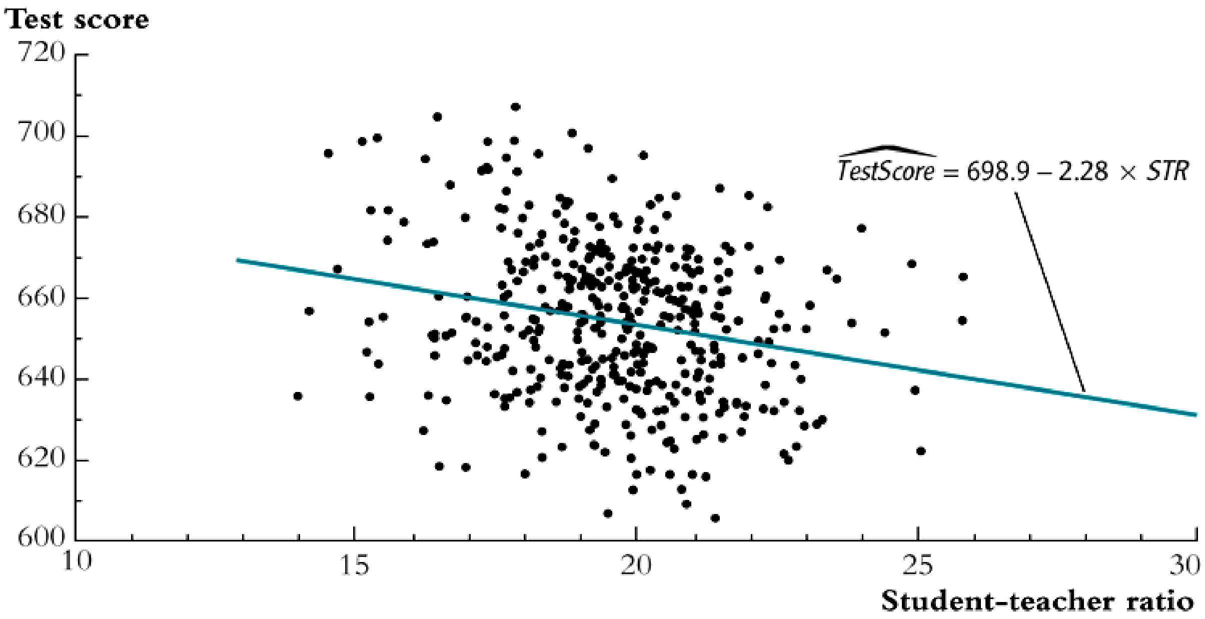


FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



OLS estimate of the *Test Score/STR* relation:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR, \quad R^2 = .05, \quad SER = 18.6$$

$$(10.4) \quad (0.52)$$

Is this a credible estimate of the causal effect on test scores of a change in the student-teacher ratio?

No: there are omitted confounding factors

Potential omitted factor- Percentage of English earners:

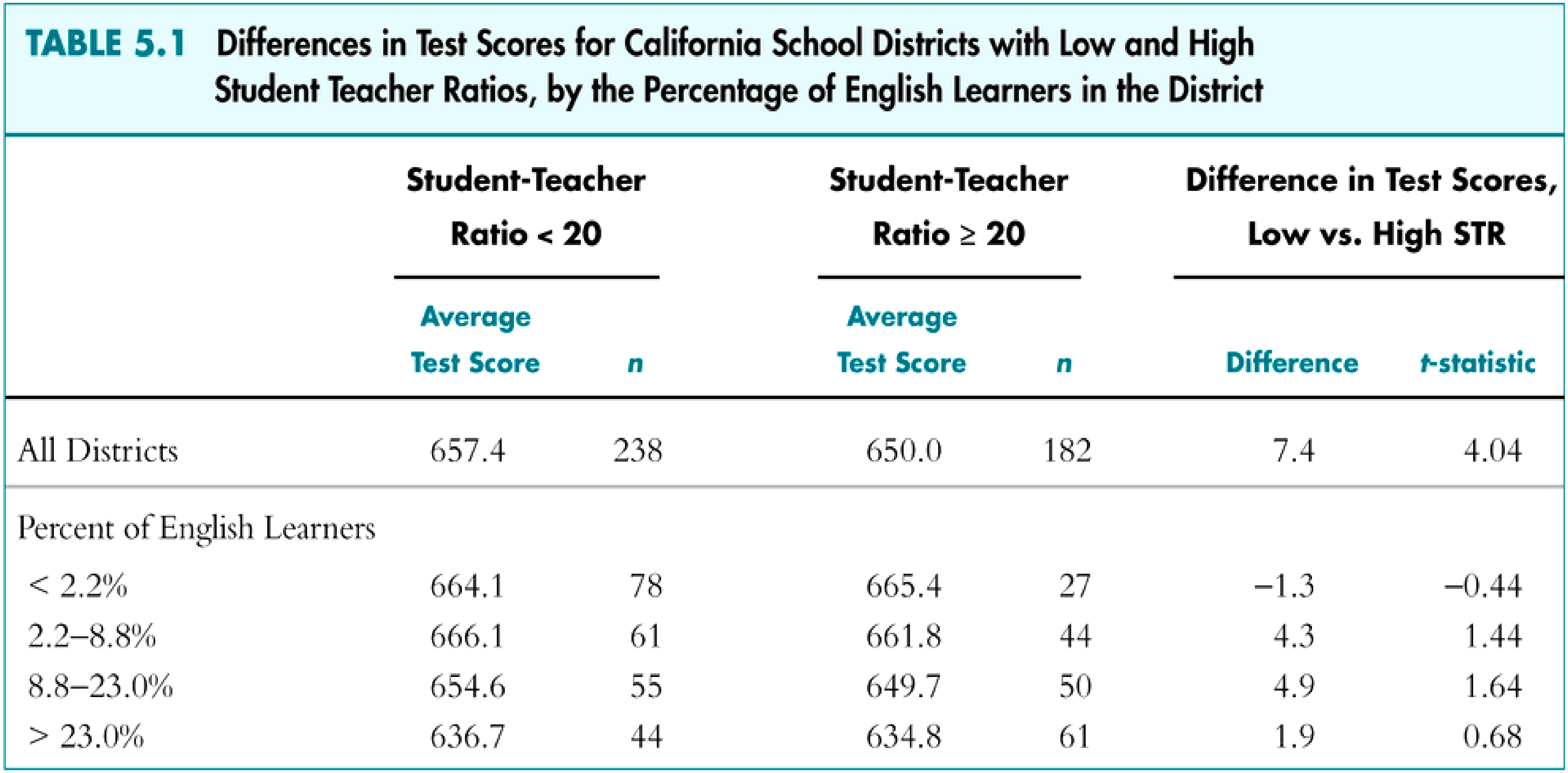


TABLE 5.1 Differences in Test Scores for California School Districts with Low and High Student Teacher Ratios, by the Percentage of English Learners in the District

	Student-Teacher Ratio < 20		Student-Teacher Ratio ≥ 20		Difference in Test Scores, Low vs. High STR	
	Average Test Score	*n*	Average Test Score	*n*	Difference	*t*-statistic
All Districts	657.4	238	650.0	182	7.4	4.04
Percent of English Learners						
< 2.2%	664.1	78	665.4	27	−1.3	−0.44
2.2–8.8%	666.1	61	661.8	44	4.3	1.44
8.8–23.0%	654.6	55	649.7	50	4.9	1.64
> 23.0%	636.7	44	634.8	61	1.9	0.68

- Districts with fewer English Learners have higher test scores
- Districts with lower percent *EL* (*PctEL*) have smaller classes

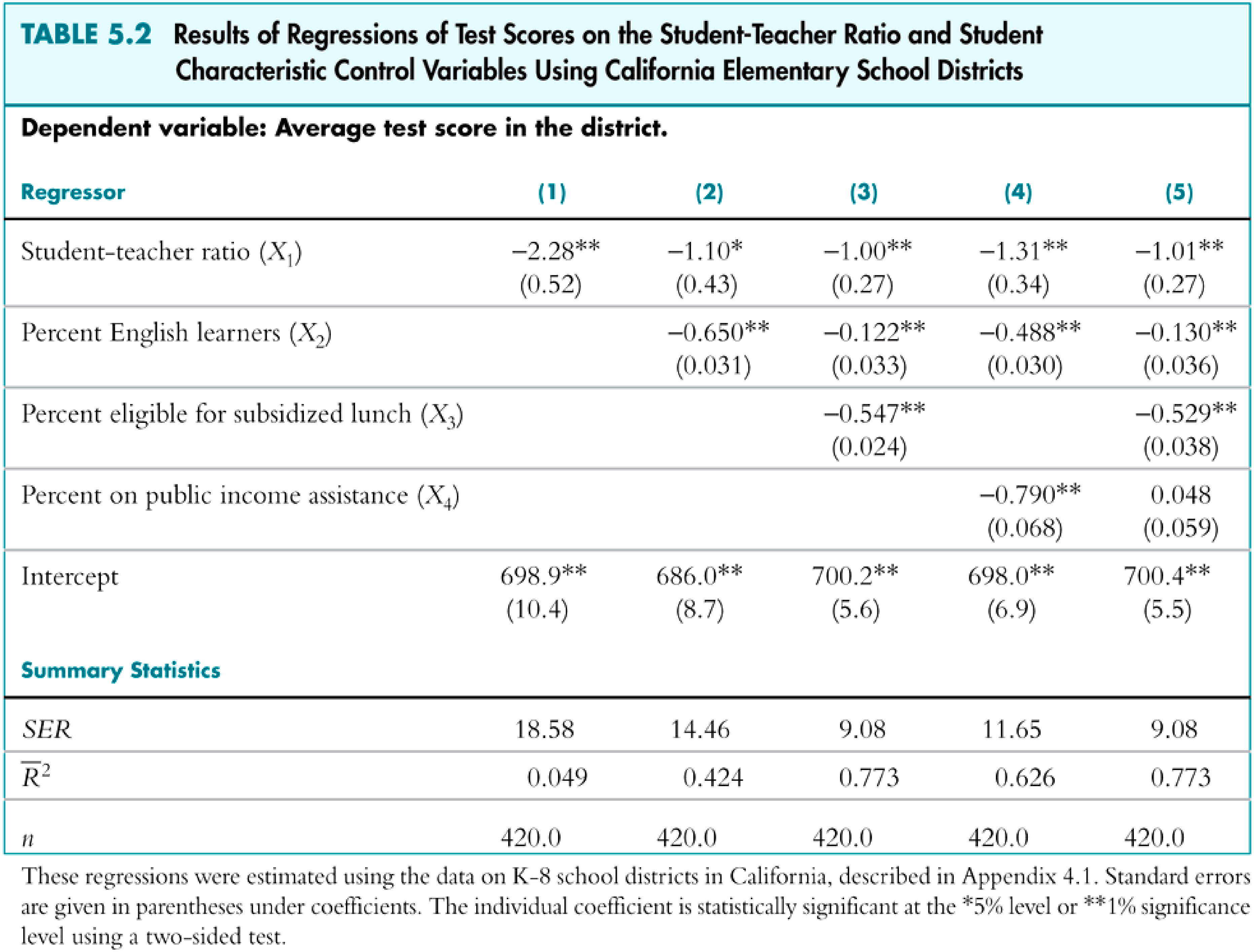


TABLE 5.2 Results of Regressions of Test Scores on the Student-Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: Average test score in the district.

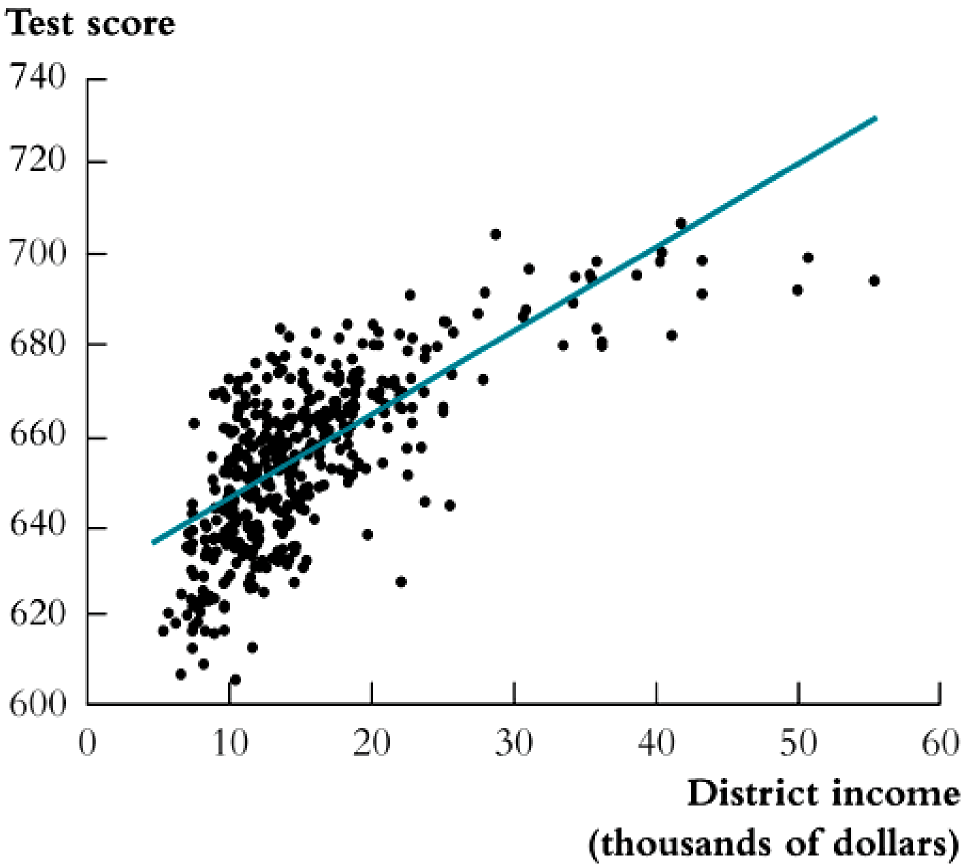
Regressor	(1)	(2)	(3)	(4)	(5)
Student-teacher ratio (X_1)	−2.28** (0.52)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (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	420.0	420.0	420.0	420.0	420.0

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.

Another omitted variable- income. But the *TestScore* – average district income relation looks like it is nonlinear.

FIGURE 6.2 Scatterplot of Test Score vs. District Income with a Linear OLS Regression Function

There is a positive correlation between test scores and district income (correlation = 0.71), but the linear OLS regression line does not adequately describe the relationship between these variables.



Estimation of the cubic specification in STATA

```
gen avginc3 = avginc*avginc2;              Create the cubic regressor
reg testscr avginc avginc2 avginc3, r;

Regression with robust standard errors                 Number of obs =     420
                                                       F(  3,   416) =  270.18
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5584
                                                       Root MSE      =  12.707

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   5.018677   .7073505     7.10   0.000     3.628251    6.409104
     avginc2 |  -.0958052   .0289537    -3.31   0.001    -.1527191   -.0388913
     avginc3 |   .0006855   .0003471     1.98   0.049     3.27e-06    .0013677
       _cons |    600.079   5.102062   117.61   0.000     590.0499     610.108
------------------------------------------------------------------------------
```

The cubic term is statistically significant at the 5%, but not 1%, level

Testing the null hypothesis of linearity, against the alternative that the population regression is quadratic and/or cubic, that is, it is a polynomial of degree up to 3:

H_0: pop'n coefficients on $Income^2$ and $Income^3 = 0$

H_1: at least one of these coefficients is nonzero.

```
test avginc2 avginc3;  Execute the test command after running the regression

 ( 1)  avginc2 = 0.0
 ( 2)  avginc3 = 0.0

       F(  2,   416) =   37.69
            Prob > F =    0.0000
```

The hypothesis that the population regression is linear is rejected at the 1% significance level against the alternative that it is a polynomial of degree up to 3.

Another idea: TestScore vs. ln(Income)

- The model is now linear in ln(*Income*), so the linear-log model can be estimated by OLS:

$$\widehat{TestScore} = \underset{(3.8)}{557.8} + \underset{(1.40)}{36.42} \ln(Income_i)$$

so a 1% increase in *Income* is associated with an increase in *TestScore* of 0.36 points on the test.

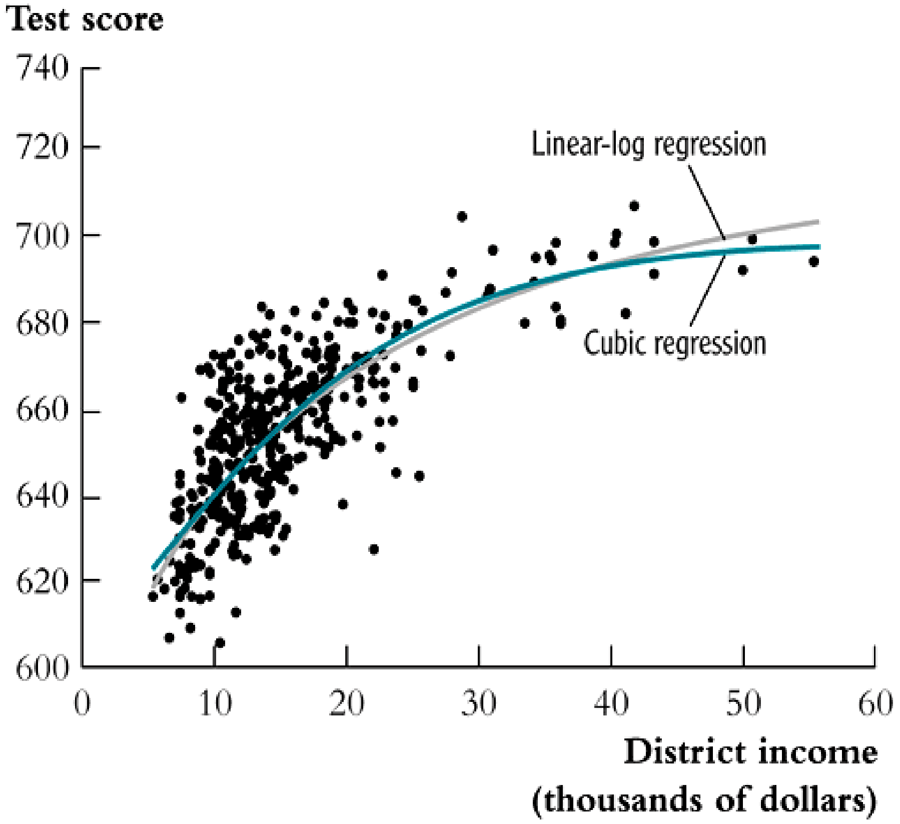
- Standard errors, confidence intervals, R^2 – all the usual tools of regression apply here.

- How does this compare to the cubic model?

$$\widehat{TestScore} = 557.8 + 36.42\ln(Income_i)$$

FIGURE 6.7 **The Linear-Log and Cubic Regression Functions**

The estimated cubic regression function (Equation (6.11)) and the estimated linear-log regression function (Equation (6.18)) are nearly identical in this sample.



Neither specification seems to fit as well as the cubic or linear-log

Application: Nonlinear Effects on Test Scores of the Student-Teacher Ratio

Focus on two questions:

1. Are there nonlinear effects of class size reduction on test scores? (Does a reduction from 35 to 30 have same effect as a reduction from 20 to 15?)

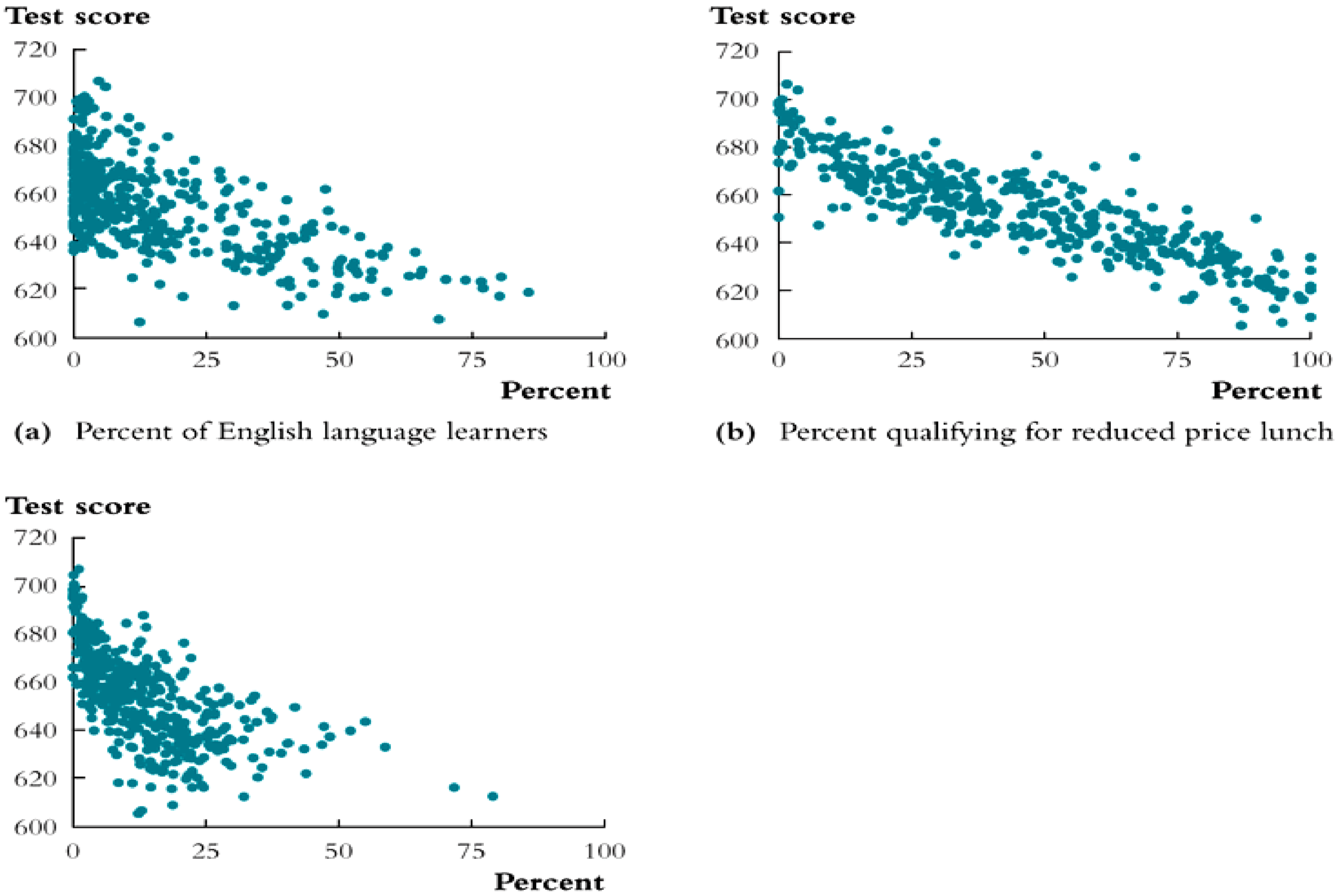
2. Are there nonlinear interactions between *PctEL* and *STR*? (Are small classes more effective when there are many English learners?)

Strategy for Question #1 (different effects for different *STR*?)

- Estimate linear and nonlinear functions of *STR*, holding constant relevant demographic variables
  - *PctEL*
  - *Income* (remember the nonlinear *TestScore*-*Income* relation!)
  - *LunchPCT* (fraction on free/subsidized lunch)
- See whether adding the nonlinear terms makes an "economically important" quantitative difference ("economic" or "real-world" importance is different than statistically significant)
- Test for whether the nonlinear terms are significant

What is a good “base” specification?

FIGURE 5.2 **Scatterplots of Test Scores vs. Three Student Characteristics**



(a) Percent of English language learners

(b) Percent qualifying for reduced price lunch

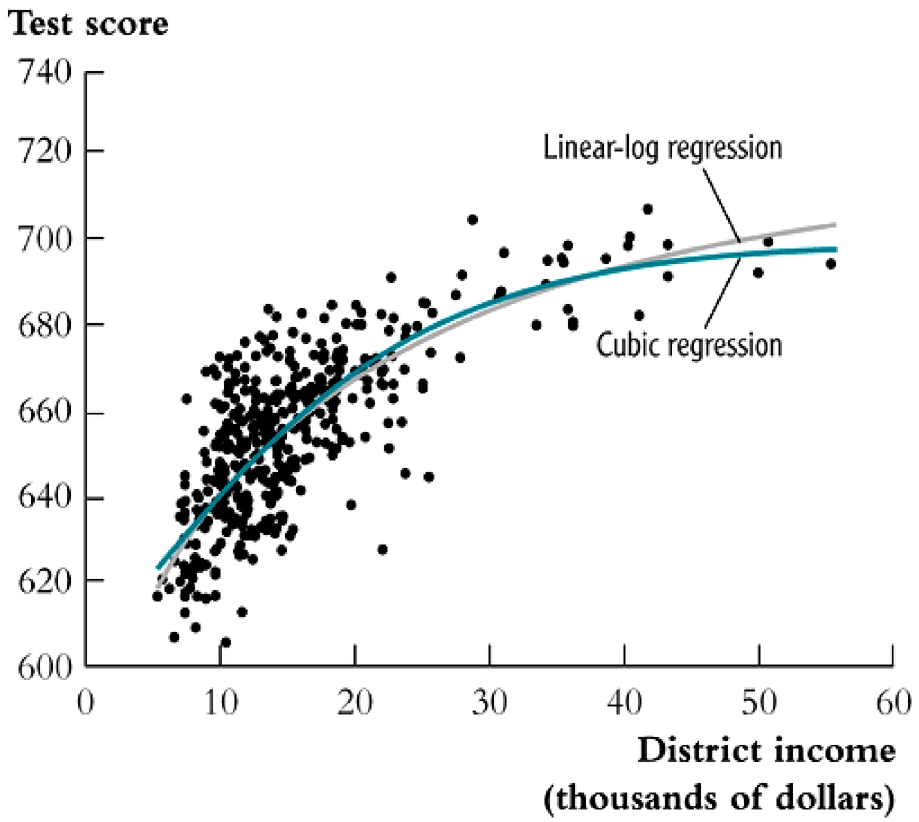
(c) Percent qualifying for income assistance

The scatterplots show a negative relationship between test scores and (a) the percentage of English learners (correlation = –0.64), (b) the percentage of students qualifying for a subsidized lunch (correlation = –0.87); and (c) the percentage qualifying for income assistance (correlation = –0.63).

The *TestScore – Income* relation

FIGURE 6.7 **The Linear-Log and Cubic Regression Functions**

The estimated cubic regression function (Equation (6.11)) and the estimated linear-log regression function (Equation (6.18)) are nearly identical in this sample.



An advantage of the logarithmic specification is that it is better behaved near the ends of the sample, especially large values of income.

Base specification

From the scatterplots and preceding analysis, here are plausible starting points for the demographic control variables:

Dependent variable: *TestScore*

Independent variable	Functional form
$PctEL$	$HiEL = \begin{cases} 1 \text{ if } PctEL \geq \mathrm{l0} \\ 0 \text{ if } PctEL < 10 \end{cases}$
$LunchPCT$	linear
$Income$	ln(*Income*) (or could use cubic)

Question #1:
Investigate by considering a polynomial in *STR*

$$\widehat{TestScore} = \underset{(163.6)}{252.0} + \underset{(24.86)}{64.33}STR – \underset{(1.25)}{3.42}STR^2 + \underset{(.021)}{.059}STR^3$$

$$– \underset{(1.03)}{5.47}HiEL – \underset{(.029)}{.420}LunchPCT + \underset{(1.78)}{11.75}\ln(Income)$$

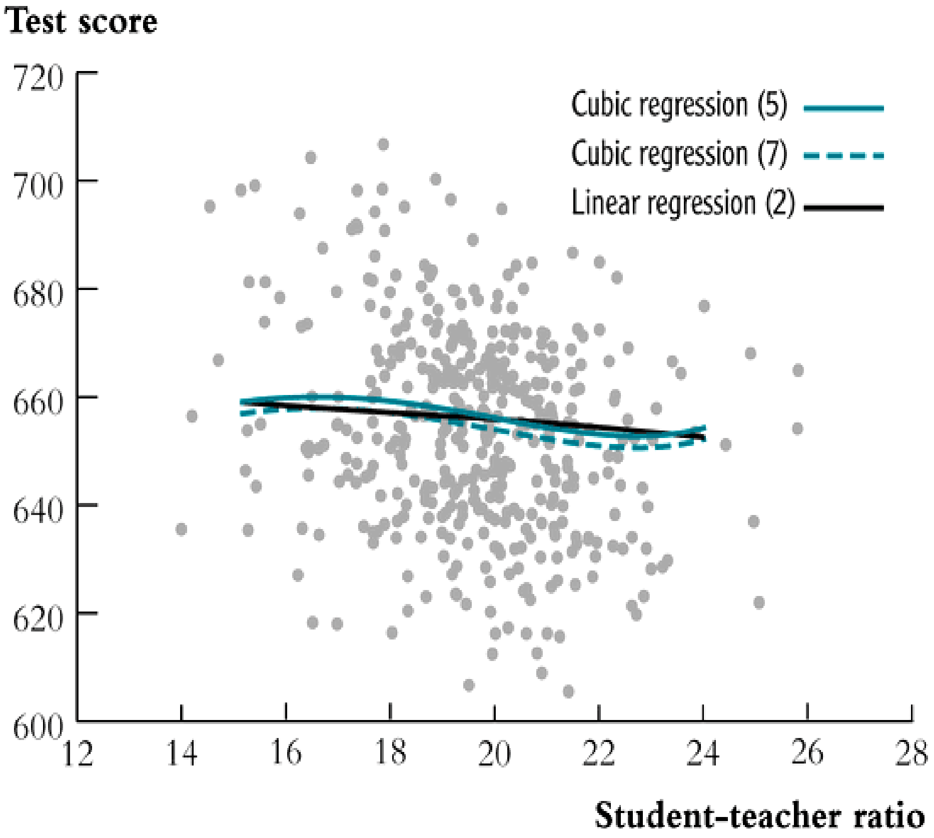
Interpretation of coefficients on:

- *HiEL*?
- *LunchPCT*?
- ln(*Income*)?
- STR, STR^2, STR^3?

- **Interpreting the regression function via plots**
  (preceding regression is labeled (5) in this figure)

FIGURE 6.10 Three Regression Functions Relating Test Scores and Student-Teacher Ratio

The cubic regressions from columns (5) and (7) of Table 6.2 are nearly identical. They indicate a small amount of nonlinearity in the relation between test scores and student-teacher ratio.



Are the higher order terms in *STR* statistically significant?

$$\begin{array}{llllll} \widehat{TestScore} = & 252.0 & +\ 64.33STR & -\ 3.42STR^2 & +\ .059STR^3 & \\ & (163.6) & (24.86) & (1.25) & (.021) & \end{array}$$

$$\begin{array}{llll} & -\ 5.47HiEL & -\ .420LunchPCT & +\ 11.75\ln(Income) \\ & (1.03) & (.029) & (1.78) \end{array}$$

(a) H_0: quadratic in *STR* v. H_1: cubic in *STR*?

t = .059/.021 = 2.86 (p = .005)

(b) H_0: linear in *STR* v. H_1: nonlinear/up to cubic in *STR*?

F = 6.17 (p = .002)

Question #2: STR-PctEL interactions

(to simplify things, ignore STR^2, STR^3 terms for now)

$$\widehat{TestScore} = \underset{(9.9)}{653.6} - \underset{(.34)}{.53}STR + \underset{(9.80)}{5.50}HiEL - \underset{(.50)}{.58}HiEL*STR$$

$$- \underset{(.029)}{.411}LunchPCT + \underset{(1.80)}{12.12}\ln(Income)$$

Interpretation of coefficients on:

- *STR*?
- *HiEL*? (wrong sign?)
- *HiEL*STR*?
- *LunchPCT*?
- ln(*Income*)?

Interpreting the regression functions via plots:

$$\widehat{TestScore} = \underset{(9.9)}{653.6} - \underset{(.34)}{.53}STR + \underset{(9.80)}{5.50}HiEL - \underset{(.50)}{.58}HiEL*STR$$

$$- \underset{(.029)}{.411}LunchPCT + \underset{(1.80)}{12.12}\ln(Income)$$

"Real-world" ("policy" or "economic") importance of the interaction term:

$$\frac{\Delta\widehat{TestScore}}{\Delta STR} = -.53 - .58HiEL = \begin{cases} -1.12 \text{ if } HiEL = 1 \\ -.53 \text{ if } HiEL = 0 \end{cases}$$

The difference in the estimated effect of reducing the *STR* is substantial; class size reduction is more effective in districts with more English learners

- **Is the interaction effect statistically significant?**

$$\widehat{TestScore} = \underset{(9.9)}{653.6} - \underset{(.34)}{.53}STR + \underset{(9.80)}{5.50}HiEL - \underset{(.50)}{.58}HiEL*STR$$

$$- \underset{(.029)}{.411}LunchPCT + \underset{(1.80)}{12.12}\ln(Income)$$

(a) H_0: coeff. on interaction=0 v. H_1: nonzero interaction

t = –1.17 not significant at the 10% level

(b) H_0: both coeffs involving *STR* = 0 vs.

H_1: at least one coefficient is nonzero (*STR* enters)

F = 5.92 (p = .003)

Next: specifications with polynomials + interactions!

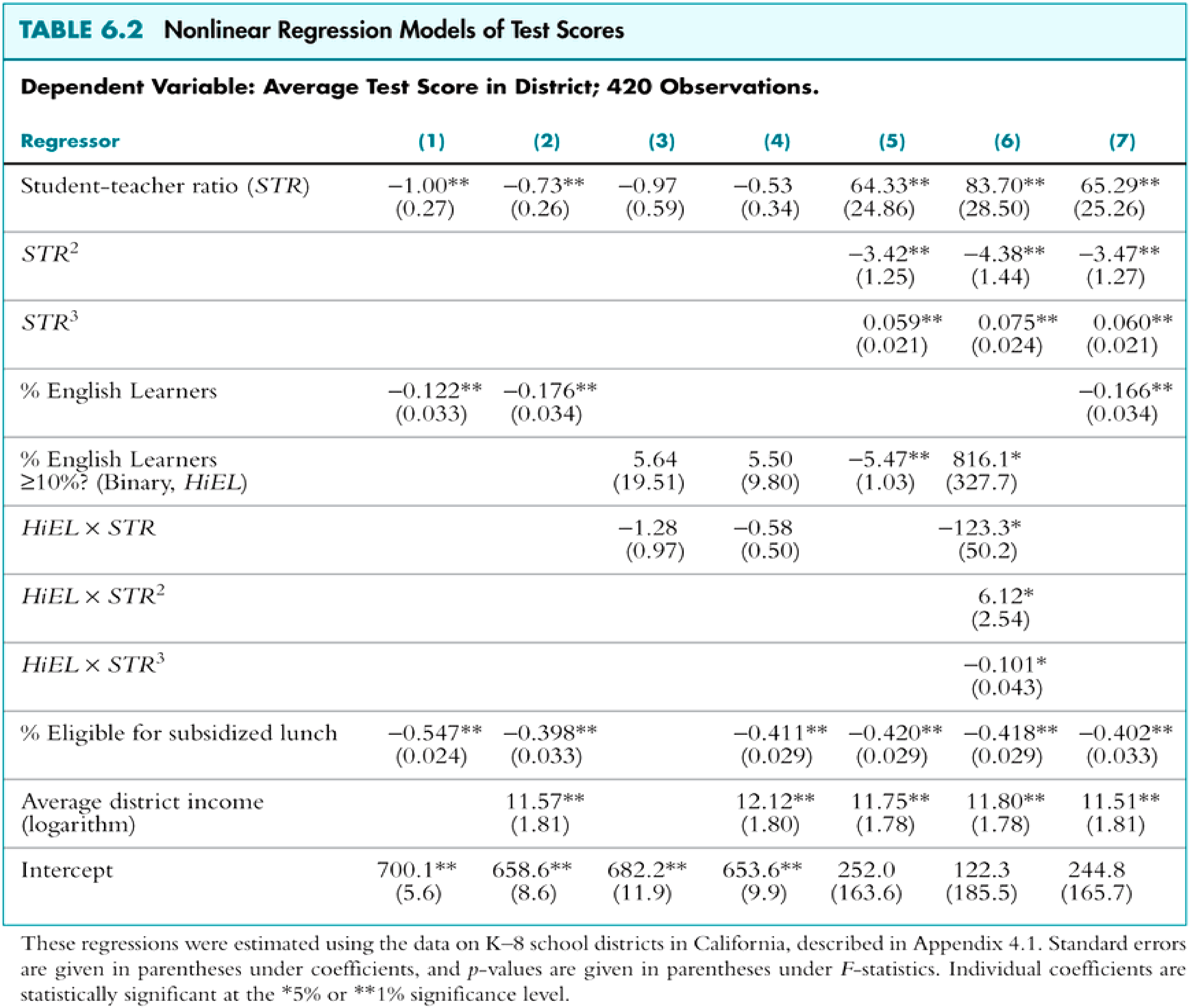


TABLE 6.2 Nonlinear Regression Models of Test Scores

Dependent Variable: Average Test Score in District; 420 Observations.

Regressor	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Student–teacher ratio (*STR*)	−1.00** (0.27)	−0.73** (0.26)	−0.97 (0.59)	−0.53 (0.34)	64.33** (24.86)	83.70** (28.50)	65.29** (25.26)
STR^2					−3.42** (1.25)	−4.38** (1.44)	−3.47** (1.27)
STR^3					0.059** (0.021)	0.075** (0.024)	0.060** (0.021)
% English Learners	−0.122** (0.033)	−0.176** (0.034)					−0.166** (0.034)
% English Learners ≥10%? (Binary, *HiEL*)			5.64 (19.51)	5.50 (9.80)	−5.47** (1.03)	816.1* (327.7)	
$HiEL \times STR$			−1.28 (0.97)	−0.58 (0.50)		−123.3* (50.2)	
$HiEL \times STR^2$						6.12* (2.54)	
$HiEL \times STR^3$						−0.101* (0.043)	
% Eligible for subsidized lunch	−0.547** (0.024)	−0.398** (0.033)		−0.411** (0.029)	−0.420** (0.029)	−0.418** (0.029)	−0.402** (0.033)
Average district income (logarithm)		11.57** (1.81)		12.12** (1.80)	11.75** (1.78)	11.80** (1.78)	11.51** (1.81)
Intercept	700.1** (5.6)	658.6** (8.6)	682.2** (11.9)	653.6** (9.9)	252.0 (163.6)	122.3 (185.5)	244.8 (165.7)

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, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

Interpreting the regression functions via plots:

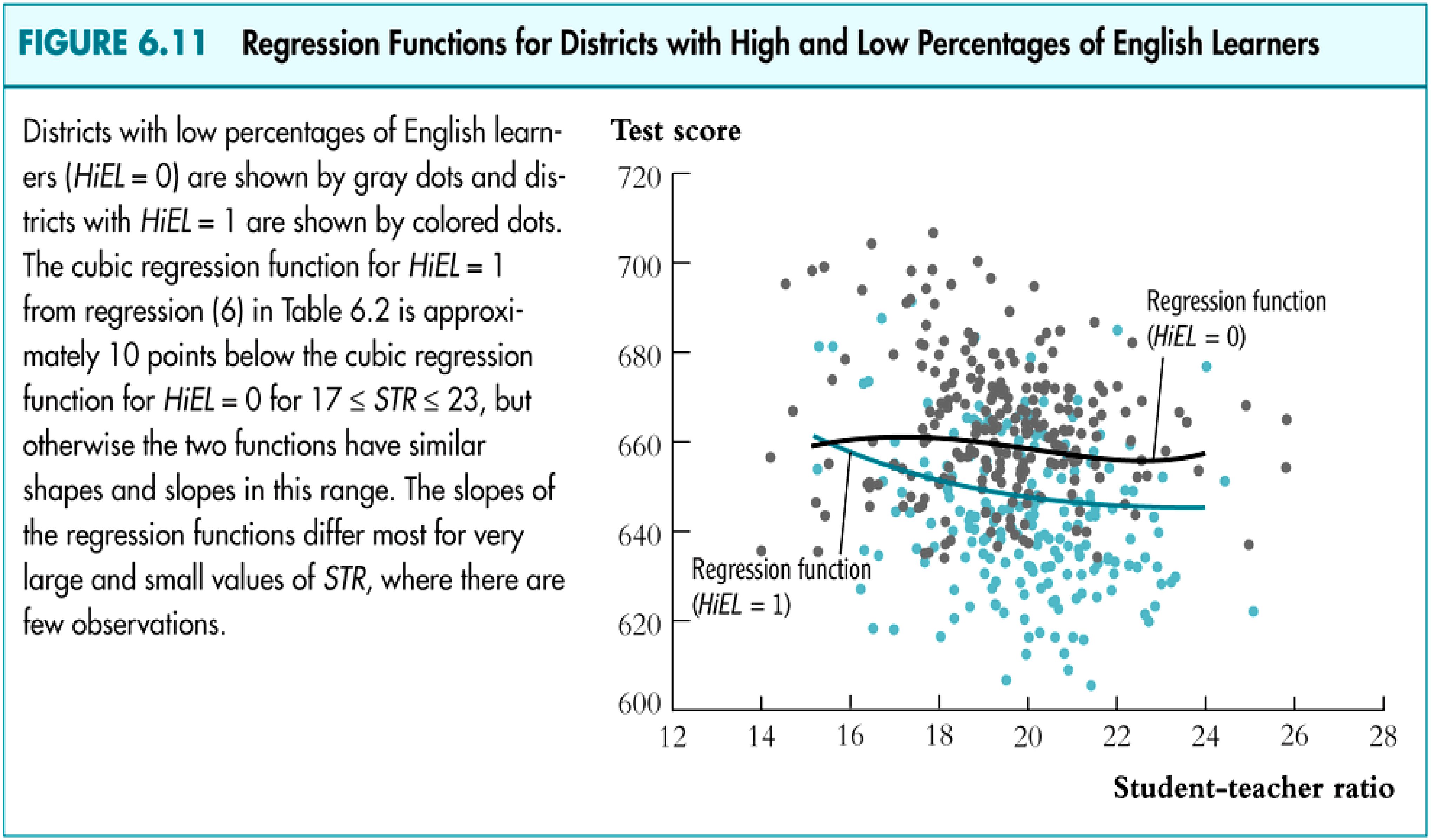


FIGURE 6.11 Regression Functions for Districts with High and Low Percentages of English Learners

Districts with low percentages of English learners (*HiEL* = 0) are shown by gray dots and districts with *HiEL* = 1 are shown by colored dots. The cubic regression function for *HiEL* = 1 from regression (6) in Table 6.2 is approximately 10 points below the cubic regression function for *HiEL* = 0 for 17 ≤ *STR* ≤ 23, but otherwise the two functions have similar shapes and slopes in this range. The slopes of the regression functions differ most for very large and small values of *STR*, where there are few observations.

Tests of joint hypotheses:

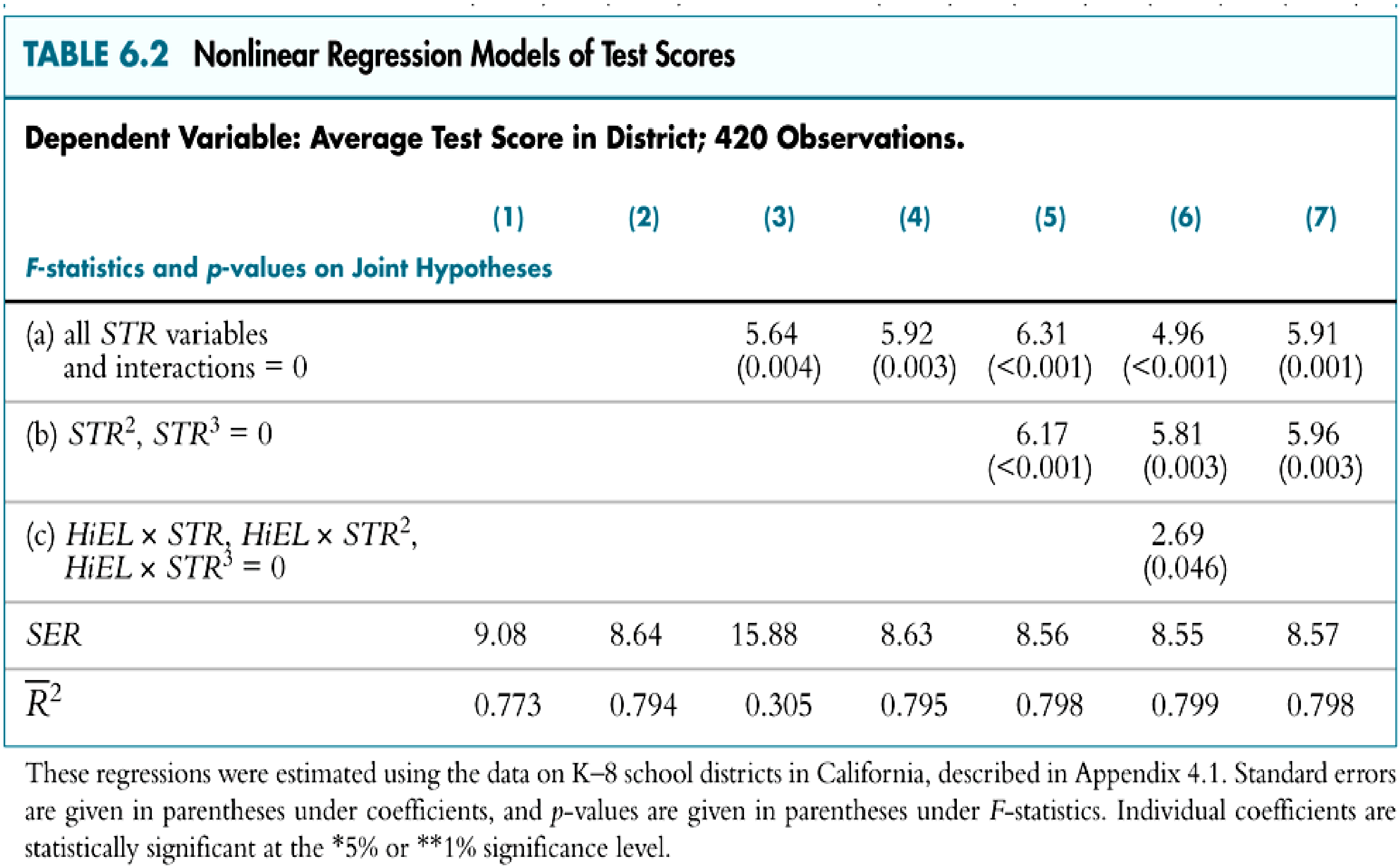


TABLE 6.2 Nonlinear Regression Models of Test Scores

Dependent Variable: Average Test Score in District; 420 Observations.

F-statistics and *p*-values on Joint Hypotheses	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(a) all *STR* variables and interactions = 0			5.64 (0.004)	5.92 (0.003)	6.31 (<0.001)	4.96 (<0.001)	5.91 (0.001)
(b) STR^2, $STR^3 = 0$					6.17 (<0.001)	5.81 (0.003)	5.96 (0.003)
(c) $HiEL \times STR$, $HiEL \times STR^2$, $HiEL \times STR^3 = 0$						2.69 (0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\bar{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

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, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

Summary: Nonlinear Regression Functions

- Using functions of the independent variables such as ln(X) or $X_{1*}X_2$, allows recasting a large family of nonlinear regression functions as multiple regression.
- Estimation and inference proceeds in the same way as in the linear multiple regression model.
- Interpretation of the coefficients is model-specific, but the general rule is to compute effects by comparing different cases (different value of the original X's)
- Many nonlinear specifications are possible, so you must use judgment: What nonlinear effect you want to analyze? What makes sense in your application?

External and internal validity

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

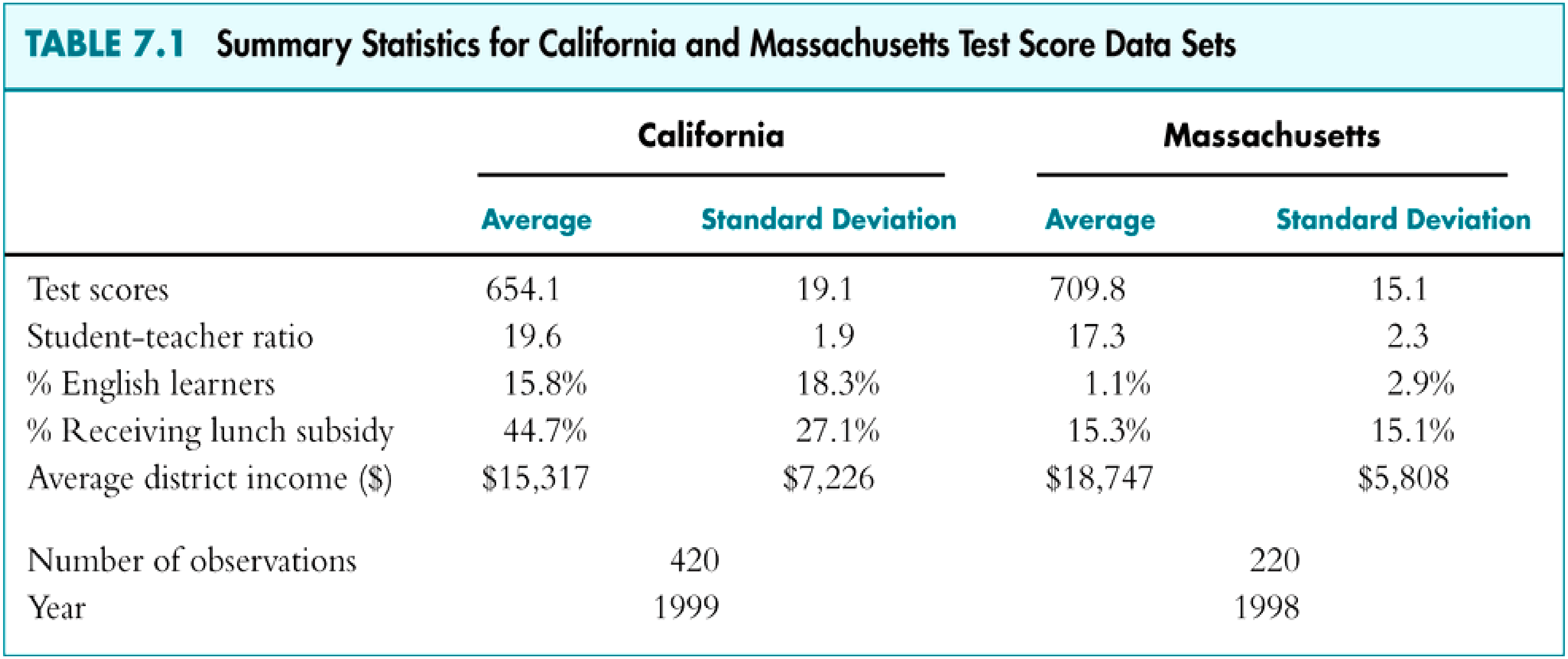
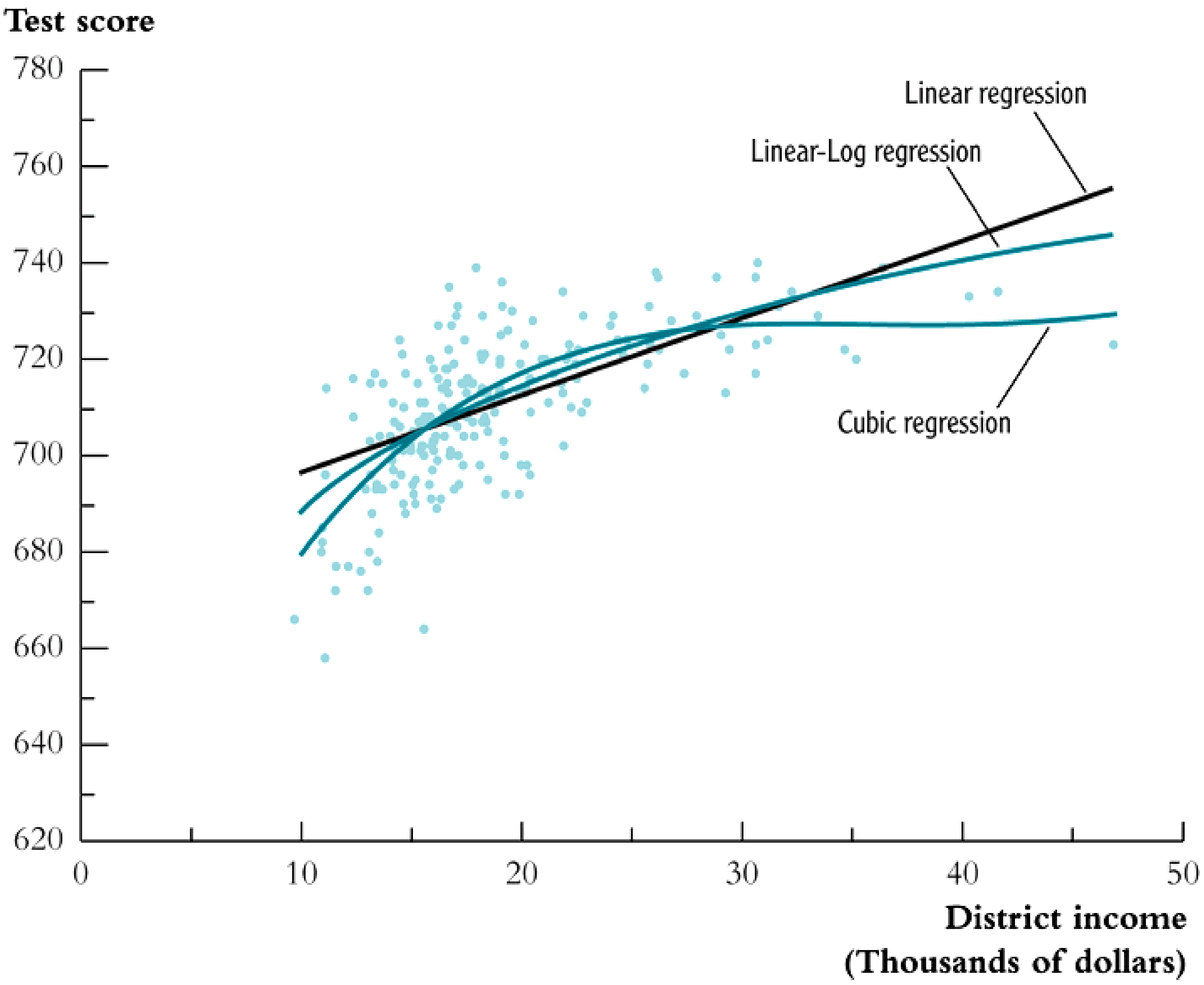


TABLE 7.1 Summary Statistics for California and Massachusetts Test Score Data Sets

	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	

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.



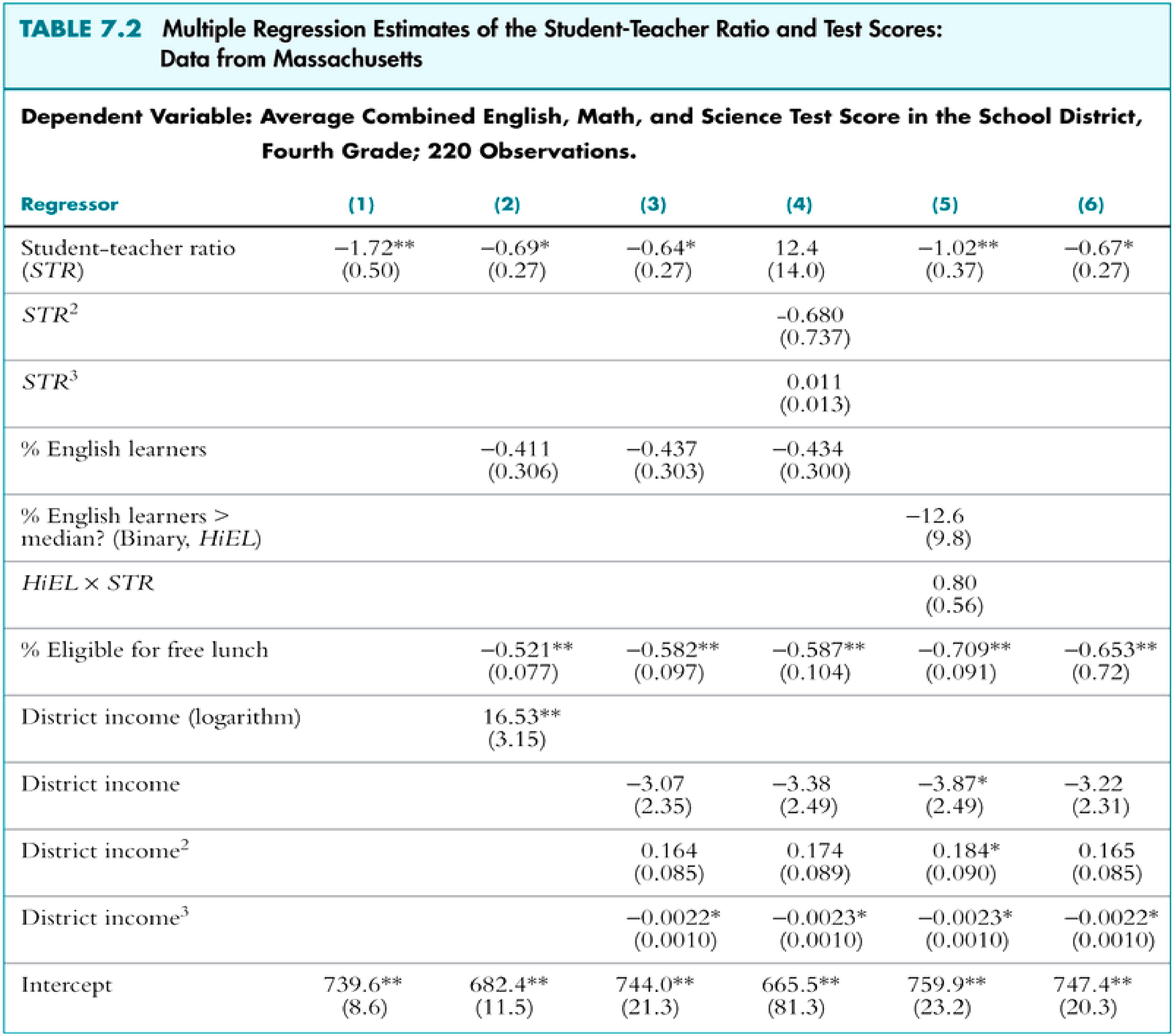
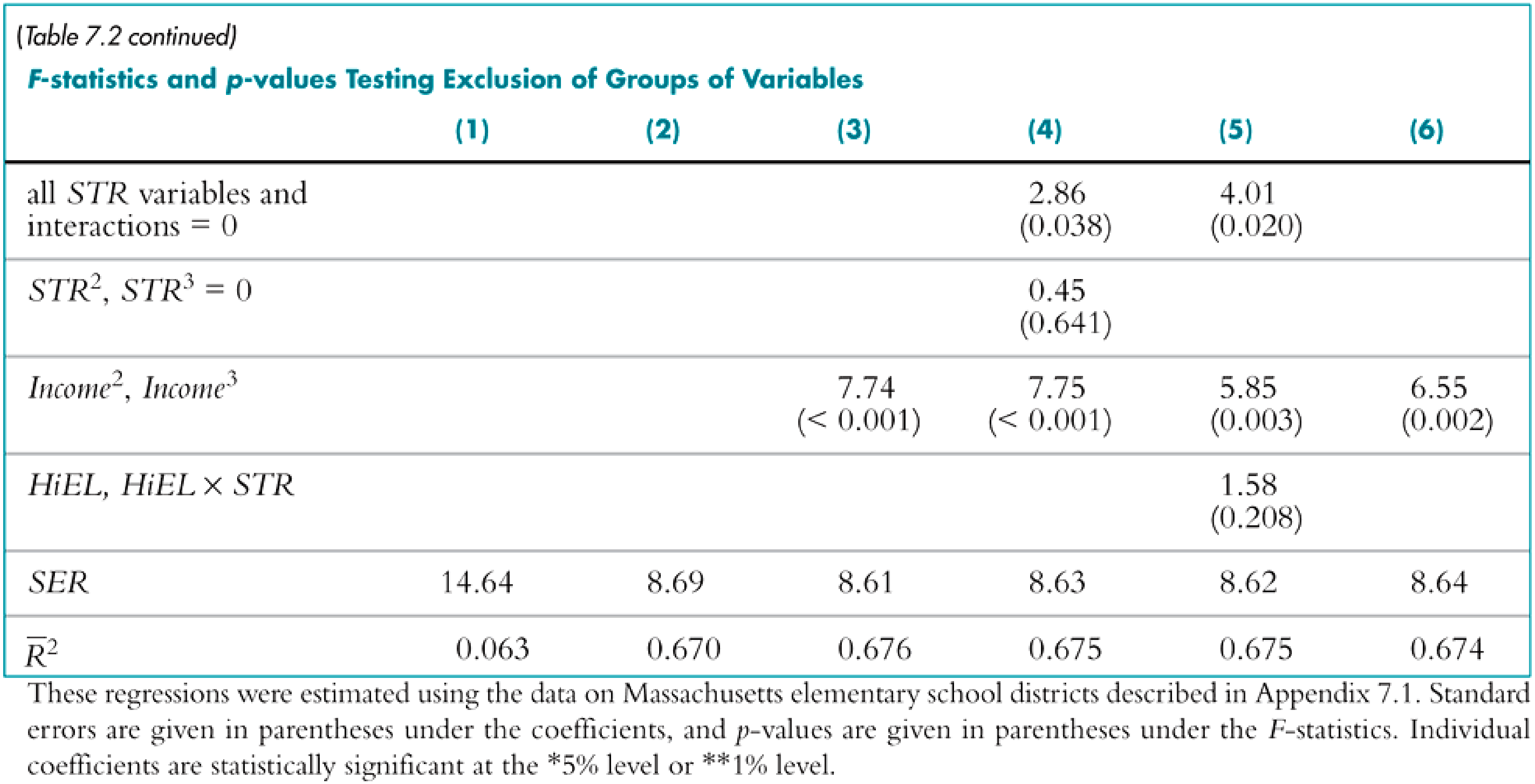


TABLE 7.2 Multiple Regression Estimates of the Student-Teacher Ratio and Test Scores: Data from Massachusetts

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)**	**(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* × *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 7.1. 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.

- 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

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

Comparison of estimated class size effects: CA vs. MA

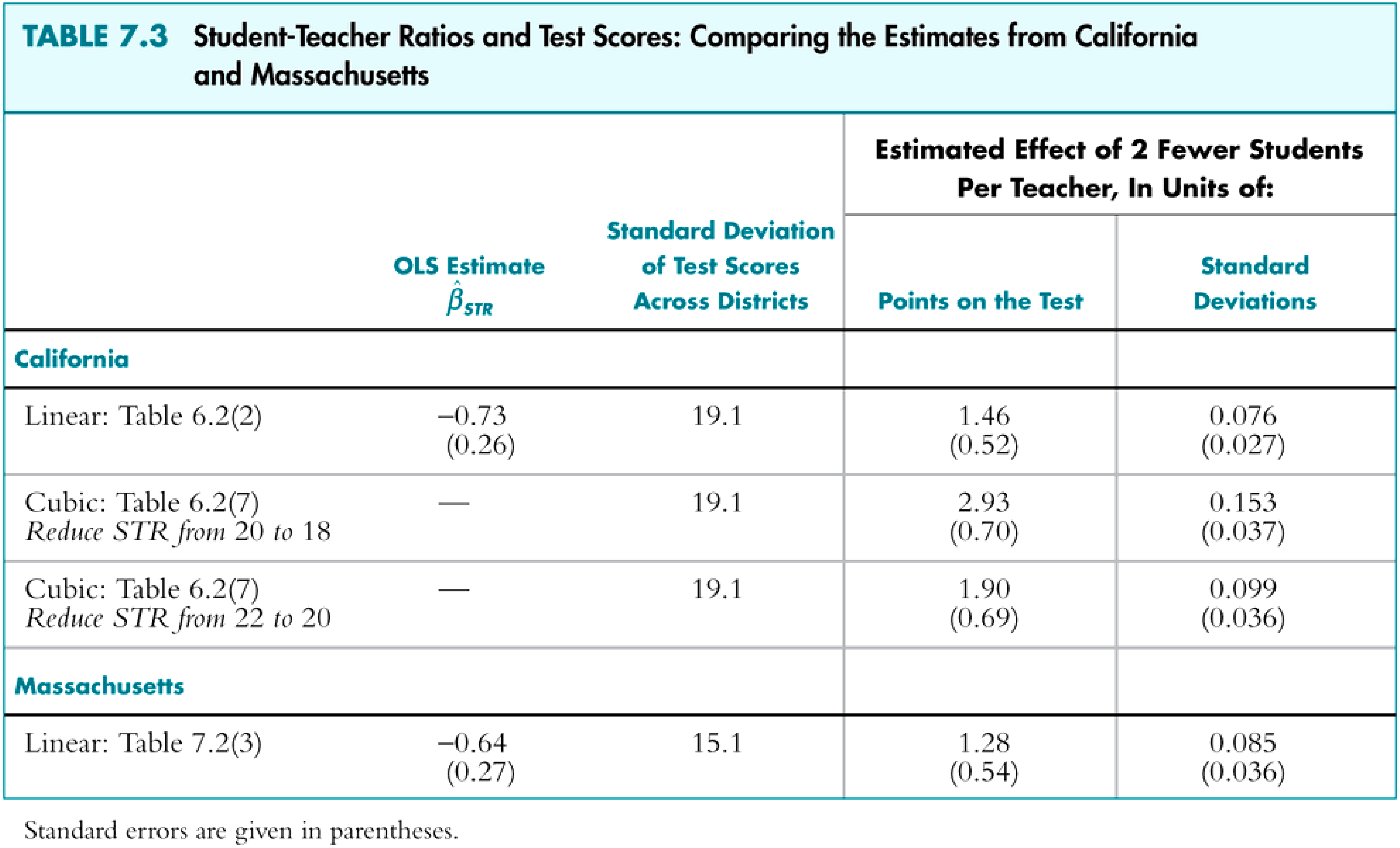


TABLE 7.3 Student-Teacher Ratios and Test Scores: Comparing the Estimates from California and Massachusetts

			Estimated Effect of 2 Fewer Students Per Teacher, In Units of:	
	OLS Estimate $\hat{\beta}_{STR}$	Standard Deviation of Test Scores Across Districts	Points on the Test	Standard Deviations
California				
Linear: Table 6.2(2)	–0.73 (0.26)	19.1	1.46 (0.52)	0.076 (0.027)
Cubic: Table 6.2(7) *Reduce STR from* 20 *to* 18	—	19.1	2.93 (0.70)	0.153 (0.037)
Cubic: Table 6.2(7) *Reduce STR from* 22 *to* 20	—	19.1	1.90 (0.69)	0.099 (0.036)
Massachusetts				
Linear: Table 7.2(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.

Remaining threats to internal validity

What the CA v. MA comparison does and doesn’t show

1. 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.