Economics

Lecture #4

Nonlinear Regression Functions II

Outline

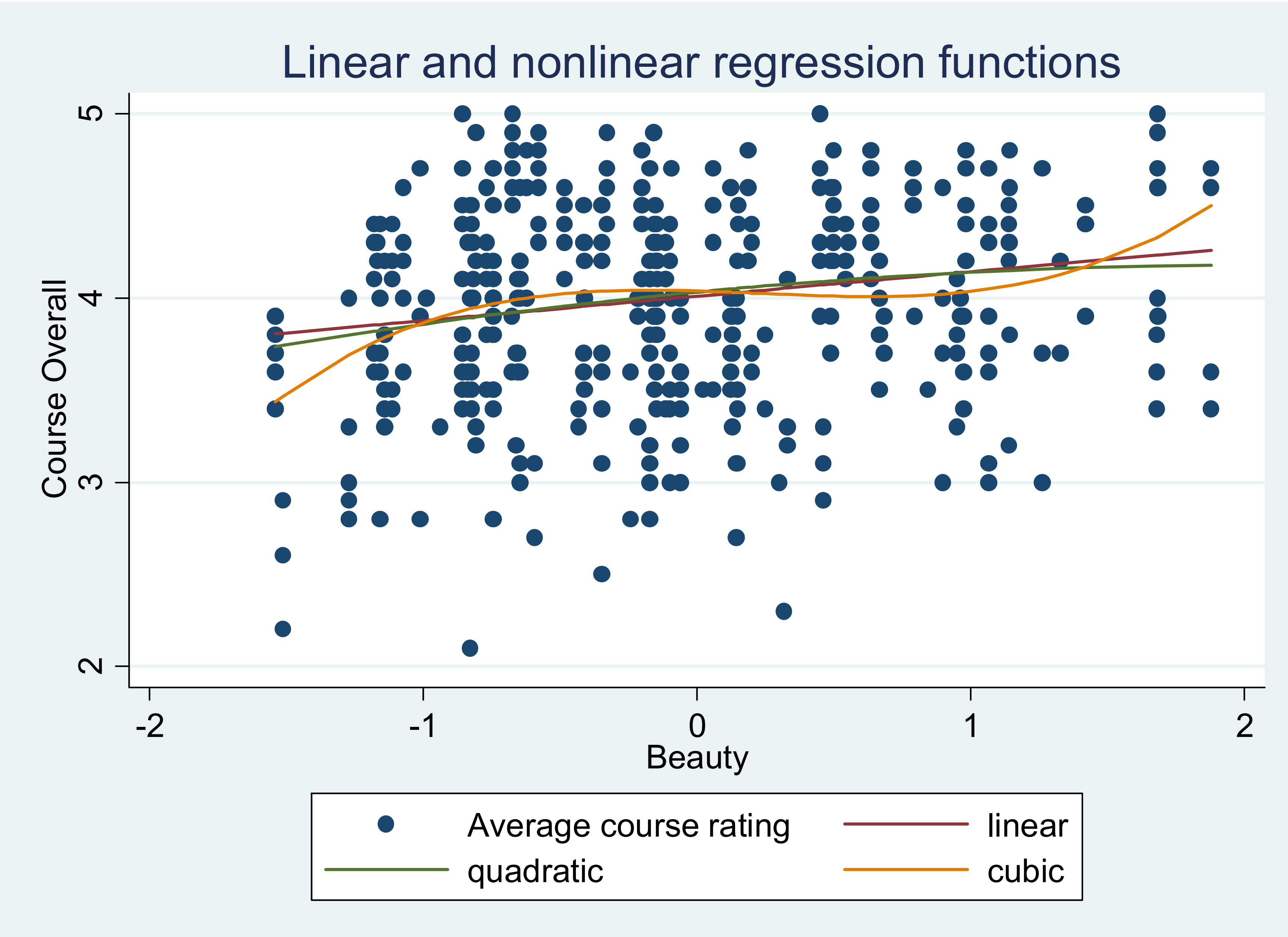
1. Standard errors for predicted effects, nonlinear specifications
2. Logarithms: gasoline demand elasticity (cross-section data)
3. Class size – test score example:
    a. Tabulation of regression results
    b. Logarithmic specifications

Computing standard errors of predicted effect in nonlinear regression functions: cubic example

$$\widehat{CourseEval} = \underset{(.035)}{4.037} - \underset{(.065)}{.043}Beauty - \underset{(.039)}{.0858}Beauty^2 + \underset{(.041)}{.1276}Beauty$$

Predicted change in *CourseEval* for a change in *Beauty* from 1 to 1.5:

$$\Delta\widehat{CourseEval} = 4.037 - .043\times1.5 - .0858\times1.5^2 + .1276\times1.5^3$$
$$- (4.037 - .043\times1 - .0858\times1^2 + .1276\times1^3)$$
$$= 0.17$$


Linear and nonlinear regression functions
Course Overall
5
4
3
2
-2
-1
0
1
2
Beauty
Average course rating
linear
quadratic
cubic

Predicted "effects" for different values of *X*:

Change in *Beauty*	$\Delta\widehat{CourseEval}$	*Std. Error*
from -1.5 to -1.0	0.39	0.10
from 0 to 0.5	-0.03	0.03
from 1.0 to 1.5	0.17	0.07

What is the effect of a change from 2.0 to 2.5? (*caution!*)

STATA: computing the SE of this predicted change

The easiest approach is to use the `lincom` command:

```
. sca a1 = (1.5) - (1);          Note: sca means "create this scalar"
. sca b1 = (1.5)*(1.5) - (1)*(1);
. sca c1 = (1.5)*(1.5)*(1.5) - (1)*(1)*(1);
. lincom a1*btystdave+b1*bty2+c1*bty3;

 ( 1)  .5 btystdave + 1.25 bty2 + 2.375 bty3 = 0

------------------------------------------------------------------------------
courseeval~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .1741214   .0669055     2.60   0.010     .0426424    .3056004
------------------------------------------------------------------------------
```

This standard error can also be computed by printing out the estimated variance matrix of the parameters. Right after running the regression execute the STATA command:

```
. matrix list e(V);    This command prints out the variance matrix

symmetric e(V)[4,4]
            btystdave        bty2        bty3       _cons
btystdave   .00424225
     bty2    .0005358   .00152125
     bty3  -.00230495  -.00059672   .00166406
    _cons    .0001839  -.00087789   .00005669   .00119767
```

Now use the "variance of sums" formula:

$$\widehat{\text{var}}(0.5\hat{\beta}_1 + 1.25\hat{\beta}_2 + 2.375\hat{\beta}_3)$$

$$= 0.5^2 \times .00424 + 1.25^2 \times .00152 + 2.375^2 \times .00166$$

$$+ 2 \times .5 \times 1.25 \times .00054 + 2 \times .5 \times 2.375 \times (-.00231)$$

$$+ 2 \times 1.25 \times 2.375 \times (-.00059) \quad = .004476$$

so $\quad SE(0.5\hat{\beta}_1 + 1.25\hat{\beta}_2 + 2.375\hat{\beta}_3) = \sqrt{.004476} = .0669$

Gasoline demand elasticity using logarithms (PS1 data set)

```
. gen lpumpprice = ln(pumpprice)
. gen lgaspc = ln(gaspc)

. reg lgaspc lpumpprice popdensity unemployment if (statename~="DC"), r

Linear regression                                      Number of obs =      47
                                                       F(  3,    43) =   16.04
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4121
                                                       Root MSE      =  .07985

------------------------------------------------------------------------------
             |               Robust
      lgaspc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  lpumpprice |  -1.770943   .3375229    -5.25   0.000    -2.451622   -1.090263
  popdensity |  -.0001178   .0000376    -3.13   0.003    -.0001937   -.0000419
unemployment |  -.0090777   .0127185    -0.71   0.479    -.0347271    .0165716
       _cons |   15.83669   1.811951     8.74   0.000     12.18254    19.49083
------------------------------------------------------------------------------
```

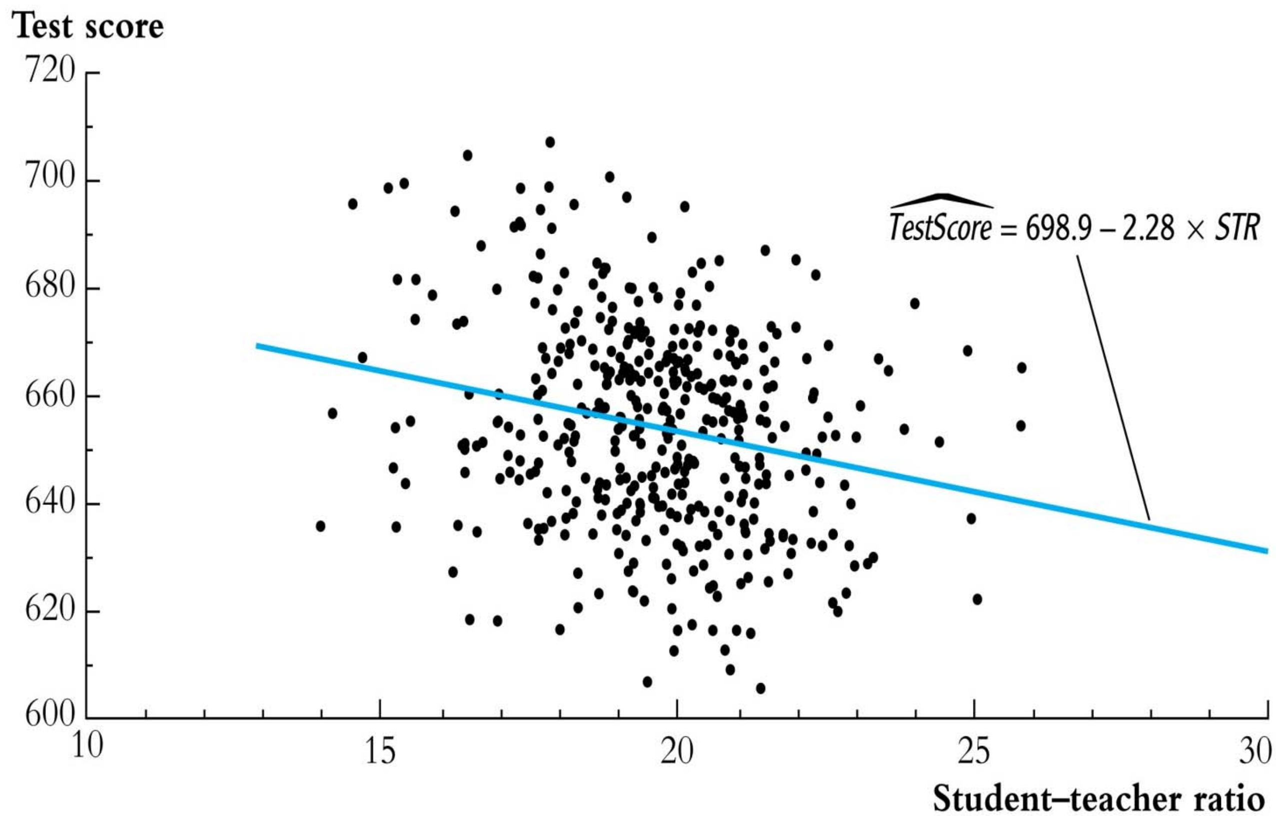
Recall PS1 estimate was -1.71 using linear demand function with elasticity evaluated at the mean – so, in this application, results are similar if we use log-log or linear demand function

Tabular Presentation of Regression Results (& new example)

Issue: effect of class size on elementary student achievement.

Data: $n = 420$ California school districts, 1998-99

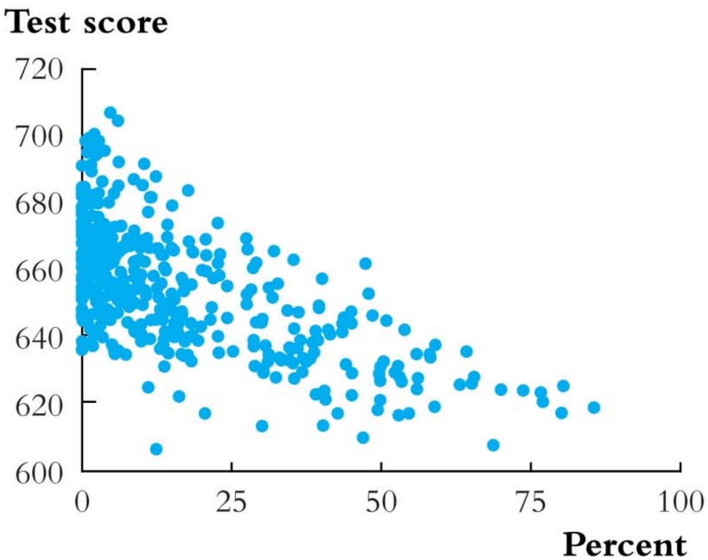
Test score: district average 5th grade Stanford Achievement Test



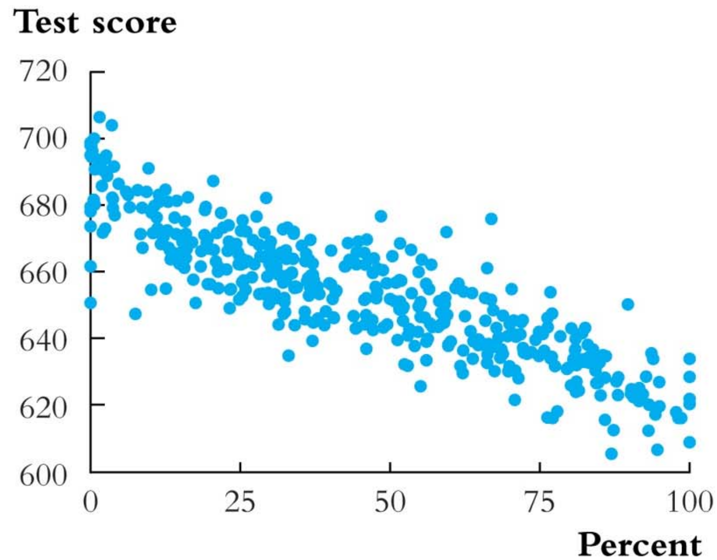
$$\widehat{TestScore} = 698.9 - 2.28STR,\ R^2 = 0.049,\ SER = 18.58$$

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

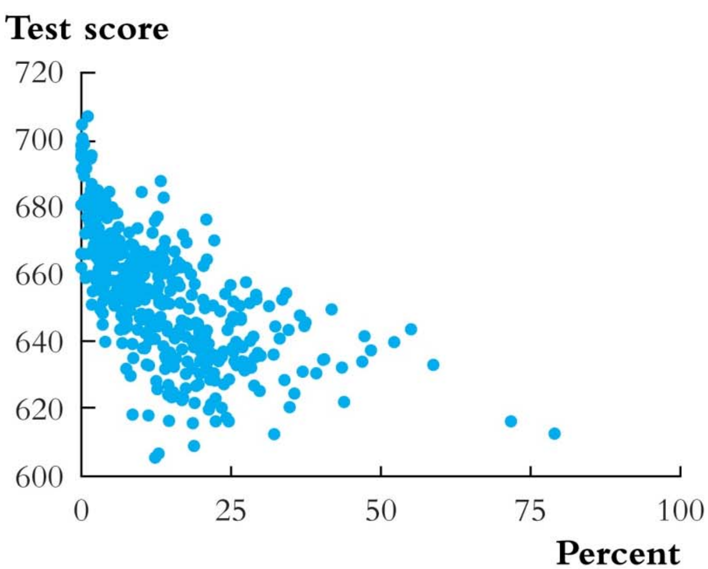
Some California data scatterplots



(a) Percentage of English language learners



(b) Percentage qualifying for reduced price lunch



(c) Percentage qualifying for income assistance

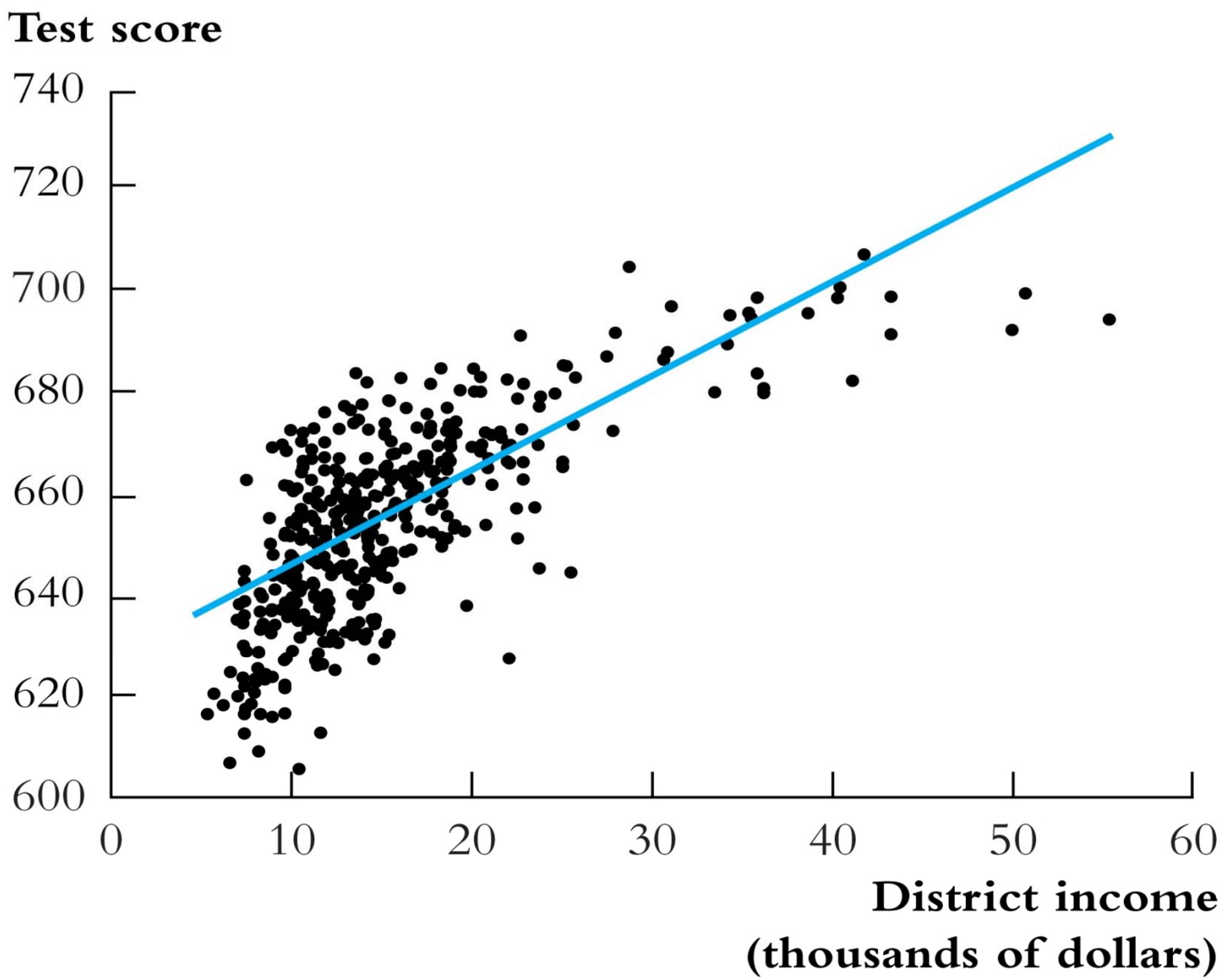
TABLE 7.1 **Results of Regressions of Test Scores on the Student–Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts**

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
$\bar{R}^2$	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. 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.

Logarithms: California data - TestScore vs. ln(Income)



Implementation:

First define the new regressor, ln(*Income*)

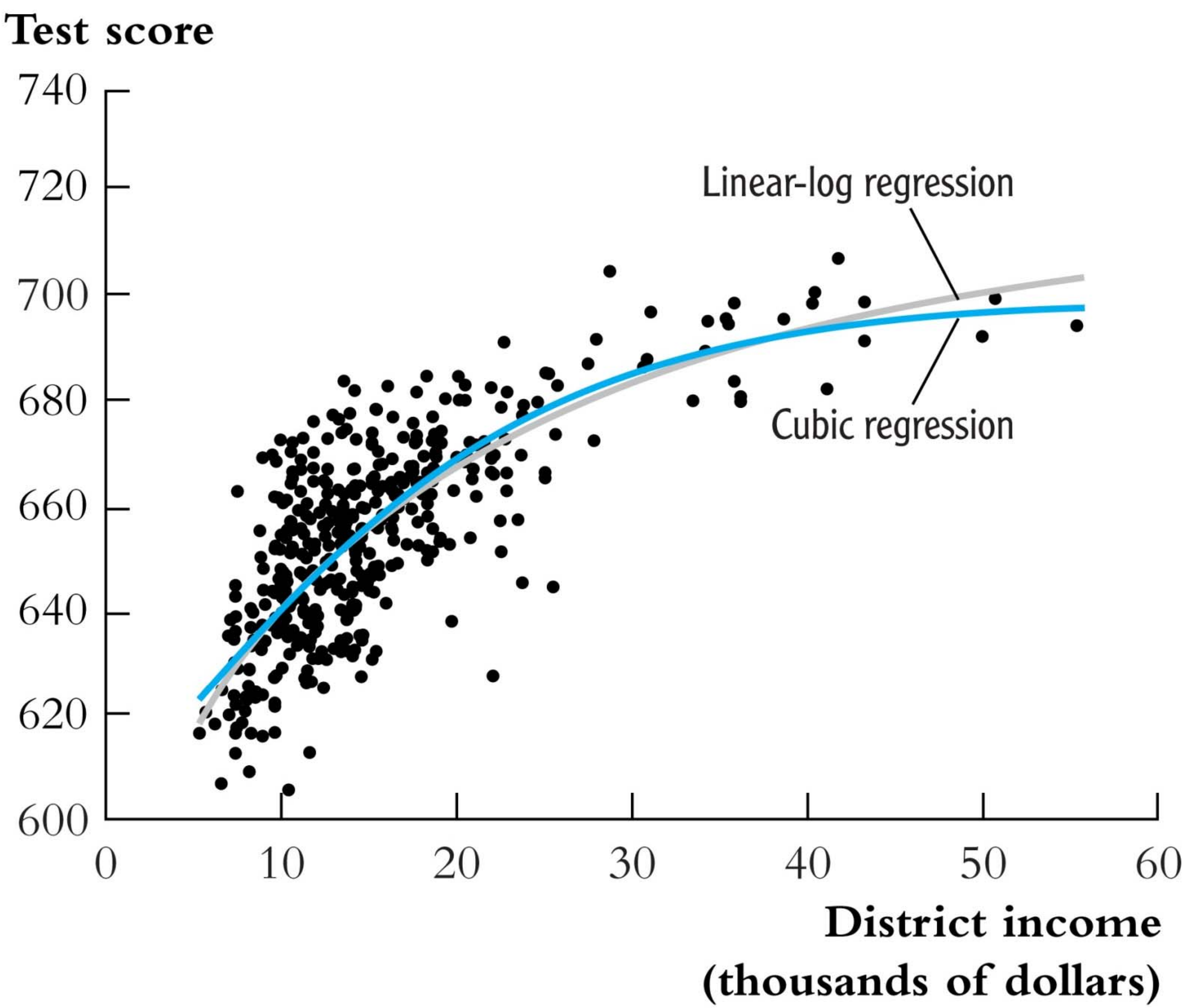
- The model is now linear in ln(*Income*), so the linear-log model can be estimated by OLS:

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

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?

Test scores: linear-log and cubic regression functions

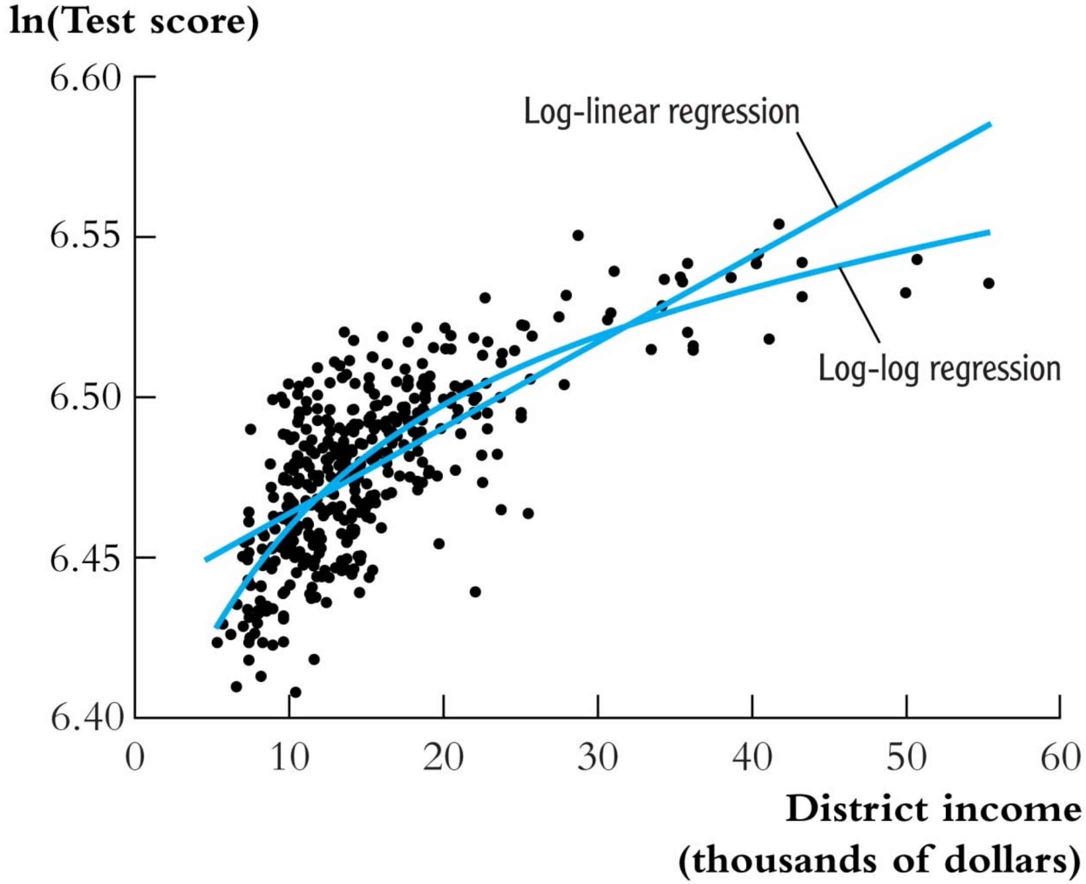


Example: ln(TestScore) vs. ln(Income)

$$\widehat{\ln(TestScore)} = 6.336 + 0.0554 \times \ln(Income_i)$$
$$(0.006) \quad (0.0021)$$

A 1% increase in *Income* is associated with an increase of .0554% in *TestScore* (*Income* up by a factor of 1.01, *TestScore* up by a factor of 1.000554)

The log-linear and log-log specifications:



- *Note vertical axis*
- *Neither log-linear nor log-log seems to fit as well as the cubic or linear-log – although you can't tell this formally here because the dependent variables are different*