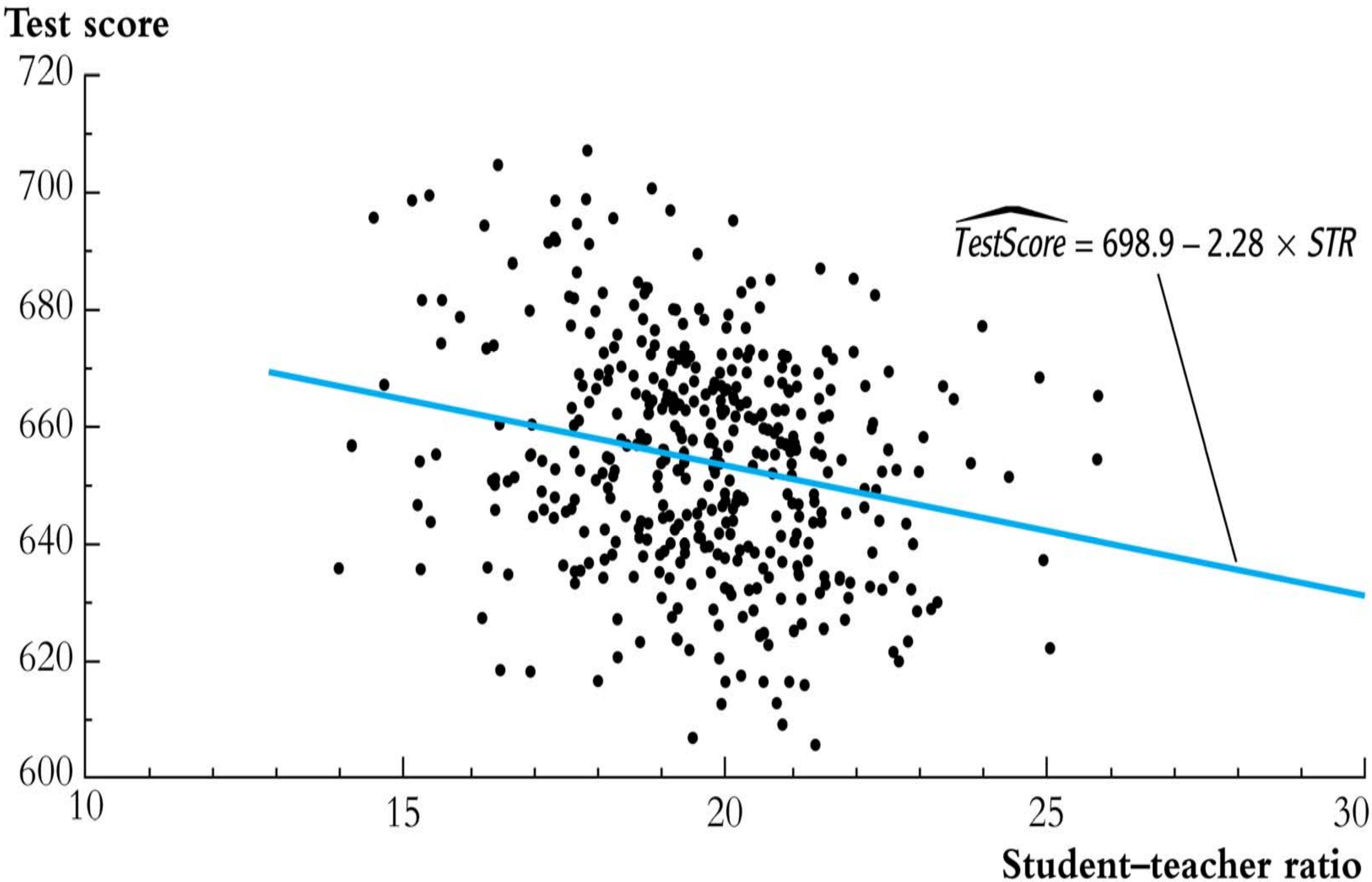
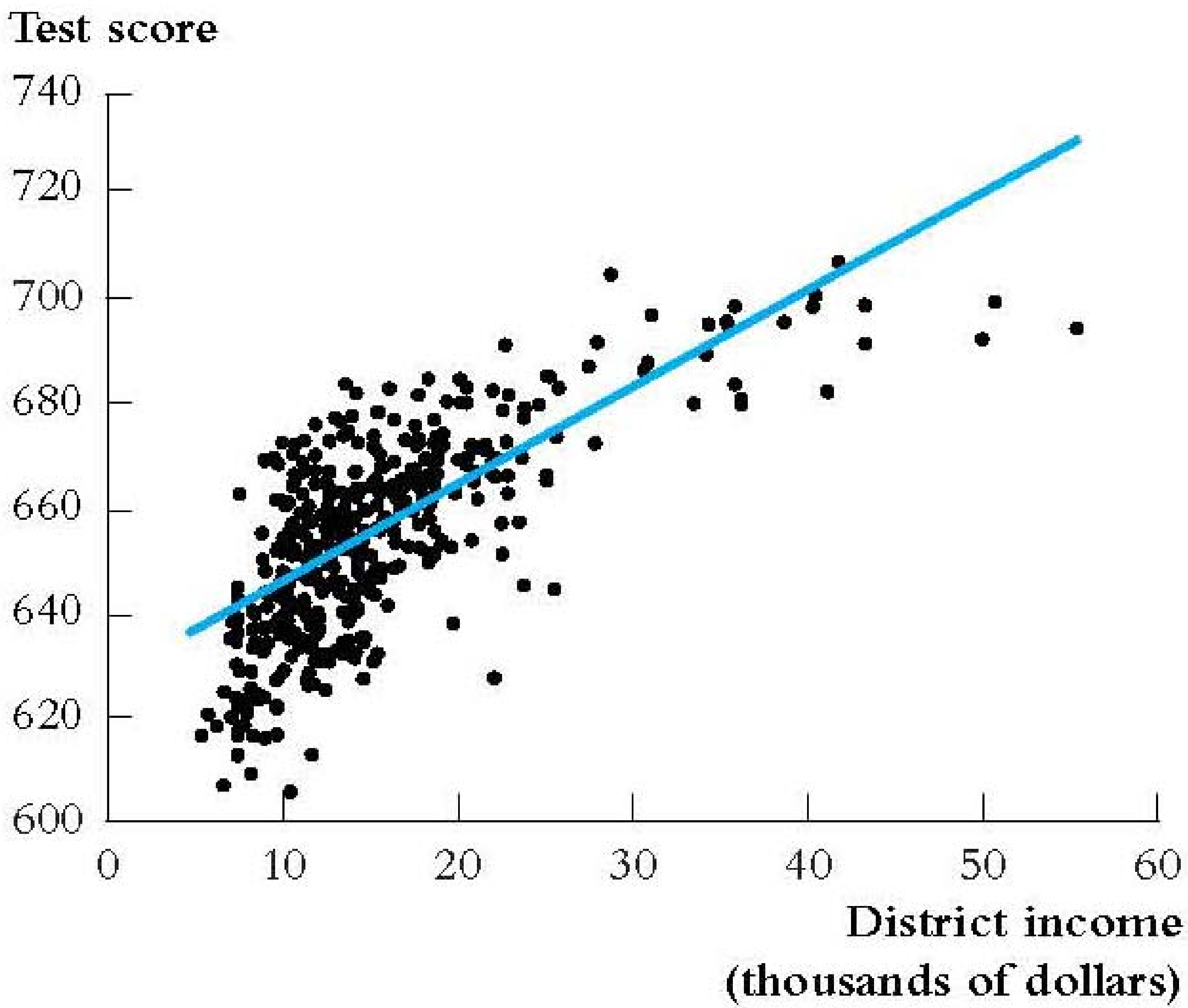
Nonlinear Regression Functions

The *TestScore* – *STR* relation looks linear (maybe)…



But the *TestScore* – *Income* relation looks nonlinear...



Nonlinear Regression – General Ideas

If a relation between Y and X is **nonlinear**:

- The effect on Y of a change in X depends on the value of X – that is, the marginal effect of X is not constant
- A linear regression is mis-specified: the functional form is wrong
- The estimator of the effect on Y of X is biased: in general it isn't even right on average.
- The solution is to estimate a regression function that is nonlinear in X

The general nonlinear population regression function

$$Y_i = f(X_{1i}, X_{2i},\ldots, X_{ki}) + u_i,\ i = 1,\ldots, n$$

Assumptions

1. $E(u_i | X_{1i}, X_{2i},\ldots, X_{ki}) = 0$ (same)
2. $(X_{1i},\ldots, X_{ki}, Y_i)$ are i.i.d. (same)
3. Big outliers are rare (same idea; the precise mathematical condition depends on the specific f)
4. No perfect multicollinearity (same idea; the precise statement depends on the specific f)

Outline

1. Nonlinear (polynomial) functions of one variable

2. Polynomial functions of multiple variables: Interactions

3. Application to the California Test Score data set

4. Addendum: Fun with logarithms

Nonlinear (Polynomial) Functions of a One RHS Variable

Approximate the population regression function by a polynomial:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- This is just the linear multiple regression model – except that the regressors are powers of *X*!
- Estimation, hypothesis testing, etc. proceeds as in the multiple regression model using OLS
- The coefficients are difficult to interpret, but the regression function itself is interpretable

Example: the *TestScore* – *Income* relation

$Income_i$ = average district income in the i^{th} district (thousands of dollars per capita)

Quadratic specification:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + u_i$$

Cubic specification:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 + \beta_3 (Income_i)^3 + u_i$$

Estimation of the quadratic specification in STATA

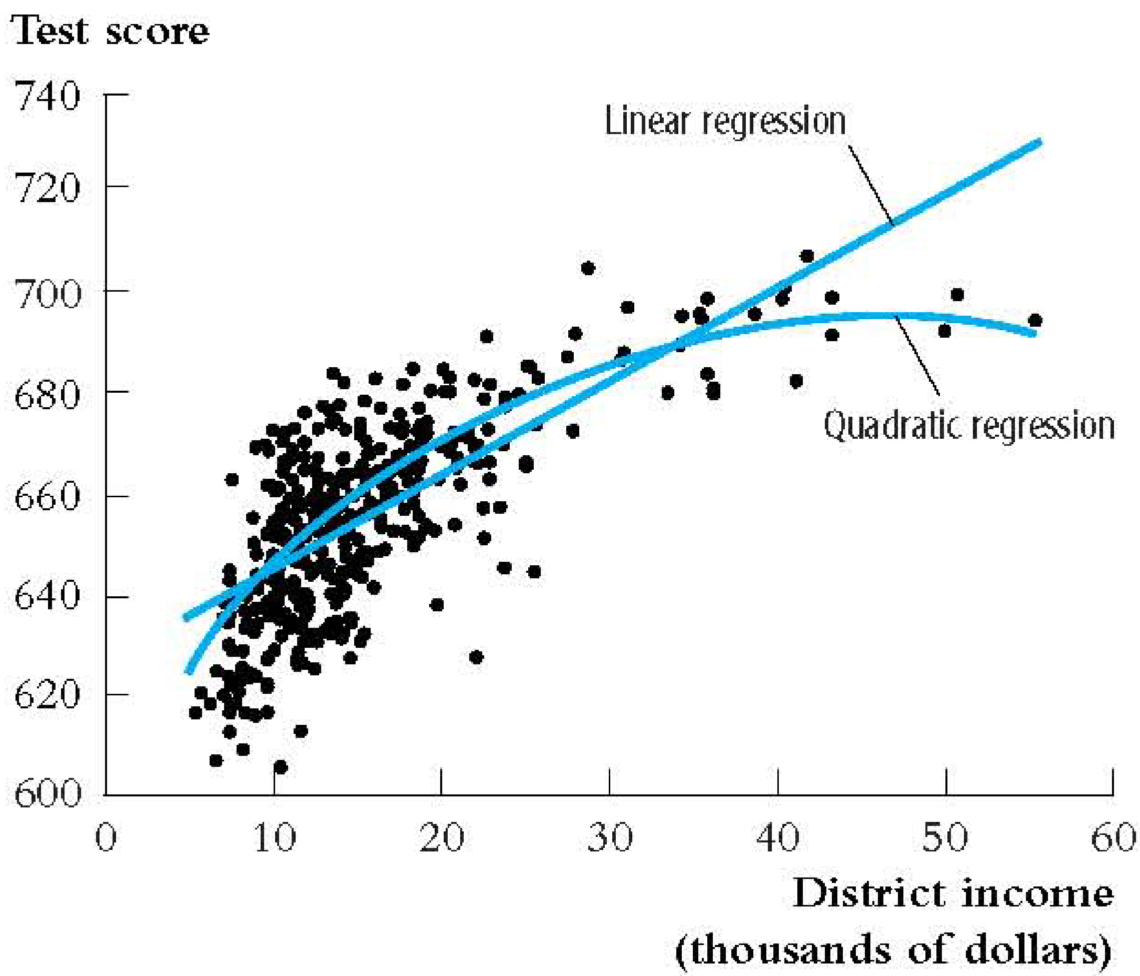
```
generate avginc2 = avginc*avginc;        Create a new regressor
reg testscr avginc avginc2, r;

Regression with robust standard errors                 Number of obs =     420
                                                       F(  2,   417) =  428.52
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5562
                                                       Root MSE      =  12.724

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      avginc |   3.850995   .2680941    14.36   0.000      3.32401    4.377979
     avginc2 |  -.0423085   .0047803    -8.85   0.000     -.051705   -.0329119
       _cons |   607.3017   2.901754   209.29   0.000     601.5978    613.0056
------------------------------------------------------------------------------
```

Test the null hypothesis of linearity against the alternative that the regression function is a quadratic….

Interpreting the estimated regression function:

(a) Plot the predicted values

$$\widehat{TestScore} = 607.3 + 3.85 Income_i - 0.0423(Income_i)^2$$
$$(2.9) \quad (0.27) \qquad\qquad (0.0048)$$



***Interpreting the estimated regression function, ctd*:**

(b) Compute "effects" for different values of X

$$\widehat{TestScore} = 607.3 + 3.85 Income_i - 0.0423(Income_i)^2$$
$$\qquad\qquad (2.9) \quad (0.27) \qquad\qquad\qquad (0.0048)$$

Predicted change in *TestScore* for a change in income from $5,000 per capita to $6,000 per capita:

$$\Delta\widehat{TestScore} = 607.3 + 3.85\times 6 - 0.0423\times 6^2$$
$$- (607.3 + 3.85\times 5 - 0.0423\times 5^2)$$
$$= 3.4$$

$$\widehat{TestScore} = 607.3 + 3.85Income_i - 0.0423(Income_i)^2$$

Predicted "effects" for different values of X:

Change in *Income* ($1000 per capita)	$\Delta\widehat{TestScore}$
from 5 to 6	3.4
from 25 to 26	1.7
from 45 to 46	0.0

The "effect" of a change in income is greater at low than high income levels (perhaps, a declining marginal benefit of an increase in school budgets?)

Caution! What is the effect of a change from 65 to 66?

Don't extrapolate outside the range of the data!

Estimation of a 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
------------------------------------------------------------------------------
```

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: population 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.

Summary: polynomial regression functions

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_r X_i^r + u_i$$

- Estimation: by OLS after defining new regressors
- Coefficients have complicated interpretations
- To interpret the estimated regression function:
  - plot predicted values as a function of x
  - compute predicted $\Delta Y/\Delta X$ at different values of x
- Hypotheses concerning degree r can be tested by t- and F-tests on the appropriate (blocks of) variable(s).
- Choice of degree r
  - plot the data; t- and F-tests, check sensitivity of estimated effects; judgment.
  - *Or use model selection criteria* (*later*)

Polynomials in Multiple Variables: Interactions

- Perhaps a class size reduction is more effective in some circumstances than in others…
- Perhaps smaller classes help more if there are many English learners, who need individual attention
- That is, $\frac{\Delta TestScore}{\Delta STR}$ might depend on *PctEL*
- More generally, $\frac{\Delta Y}{\Delta X_1}$ might depend on X_2
- How to model such "interactions" between X_1 and X_2?
- We first consider binary *X*'s, then continuous *X*'s

(a) Interactions between two binary variables

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

- D_{1i}, D_{2i} are binary
- β_1 is the effect of changing $D_1=0$ to $D_1=1$. In this specification, *this effect doesn't depend on the value of* D_2.
- To allow the effect of changing D_1 to depend on D_2, include the "interaction term" $D_{1i} \times D_{2i}$ as a regressor:

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$$

Interpreting the coefficients

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3(D_{1i} \times D_{2i}) + u_i$$

- The effect of D_1 depends on d_2 (what we wanted)
- β_3 = increment to the effect of D_1, when $D_2 = 1$

Example: *TestScore*, *STR*, *English learners*

Let

$$HiSTR = \begin{cases} 1 \text{ if } STR \geq 20 \\ 0 \text{ if } STR < 20 \end{cases} \quad \text{and} \quad HiEL = \begin{cases} 1 \text{ if } PctEL \geq 10 \\ 0 \text{ if } PctEL < 10 \end{cases}$$

$$\widehat{TestScore} = \underset{(1.4)}{664.1} - \underset{(2.3)}{18.2}HiEL - \underset{(1.9)}{1.9}HiSTR - \underset{(3.1)}{3.5}(HiSTR \times HiEL)$$

- "Effect" of *HiSTR* when *HiEL* = 0 is –1.9
- "Effect" of *HiSTR* when *HiEL* = 1 is –1.9 – 3.5 = –5.4
- Class size reduction is estimated to have a bigger effect when the percent of English learners is large
- This interaction isn't statistically significant: $t = 3.5/3.1$

(b) Interactions between continuous and binary variables

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$$

- D_i is binary, X is continuous
- As specified above, the effect on Y of X (holding constant D) = β_2, which does not depend on D
- To allow the effect of X to depend on D, include the "interaction term" $D_i \times X_i$ as a regressor:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3(D_i \times X_i) + u_i$$

Binary-continuous interactions: the two regression lines

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3(D_i \times X_i) + u_i$$

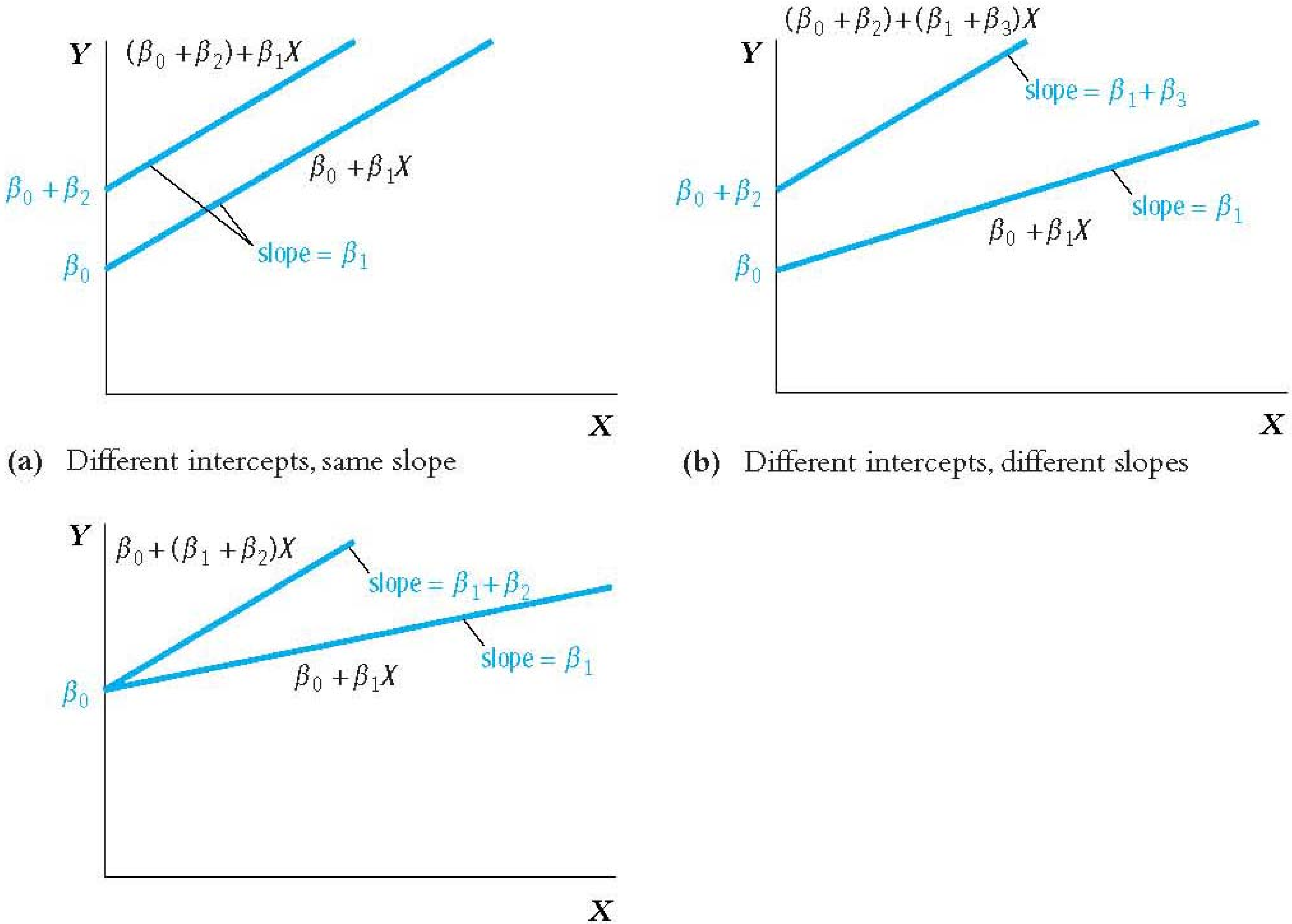
Observations with D_i= 0 (the "D = 0" group):

$$Y_i = \beta_0 + \beta_2 X_i + u_i \quad \textbf{\textit{The D=0 regression line}}$$

Observations with D_i= 1 (the "D = 1" group):

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 + \beta_2 X_i + \beta_3 X_i + u_i \\ &= (\beta_0+\beta_1) + (\beta_2+\beta_3)X_i + u_i \quad \textbf{\textit{The D=1 regression line}} \end{aligned}$$

Binary-continuous interactions, ctd.



(a) Different intercepts, same slope

(b) Different intercepts, different slopes

(c) Same intercept, different slopes

Interpreting the coefficients

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3(D_i \times X_i) + u_i$$

- β_1 = increment to intercept when D=1
- β_3 = increment to slope when D = 1

Example: *TestScore*, *STR*, *HiEL* (=1 if *PctEL* ≥ 10)

$$\widehat{TestScore} = \underset{(11.9)}{682.2} - \underset{(0.59)}{0.97}STR + \underset{(19.5)}{5.6}HiEL - \underset{(0.97)}{1.28}(STR \times HiEL)$$

- When $HiEL = 0$:

$$\widehat{TestScore} = 682.2 - 0.97STR$$

- When $HiEL = 1$,

$$\widehat{TestScore} = 682.2 - 0.97STR + 5.6 - 1.28STR$$
$$= 687.8 - 2.25STR$$

- Two regression lines: one for each *HiSTR* group.
- Class size reduction is estimated to have a larger effect when the percent of English learners is large.

***Example, ctd**: Testing hypotheses*

$$\widehat{TestScore} = \underset{(11.9)}{682.2} - \underset{(0.59)}{0.97}STR + \underset{(19.5)}{5.6}HiEL - \underset{(0.97)}{1.28}(STR \times HiEL)$$

- The two regression lines have the same **slope** ⇔ the coefficient on $STR \times HiEL$ is zero: $t = -1.28/0.97 = -1.32$
- The two regression lines have the same **intercept** ⇔ the coefficient on $HiEL$ is zero: $t = -5.6/19.5 = 0.29$
- The two regression **lines** are the same ⇔ population coefficient on $HiEL = 0$ ***and*** population coefficient on $STR \times HiEL = 0$: $F = 89.94$ (p-value < .001) **!!**
- We reject the joint hypothesis but neither individual hypothesis (*how can this be?*)

(c) Interactions between two continuous variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$$

- X_1, X_2 are continuous
- As specified, the effect of X_1 doesn't depend on X_2
- As specified, the effect of X_2 doesn't depend on X_1
- To allow the effect of X_1 to depend on X_2, include the "interaction term" $X_{1i}\times X_{2i}$ as a regressor:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3(X_{1i}\times X_{2i}) + u_i$$

Interpreting the coefficients:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3(X_{1i}\times X_{2i}) + u_i$$

- The effect of X_1 depends on X_2 (what we wanted)
- β_3 = increment to the effect of X_1 from a unit change in X_2

Example: *TestScore*, *STR*, *PctEL*

$$\widehat{TestScore} = \underset{(11.8)}{686.3} - \underset{(0.59)}{1.12}STR - \underset{(0.37)}{0.67}PctEL + \underset{(0.019)}{.0012}(STR \times PctEL),$$

The estimated effect of class size reduction is nonlinear because the size of the effect itself depends on *PctEL*:

$$\frac{\Delta TestScore}{\Delta STR} = -1.12 + .0012PctEL$$

$PctEL$	$\frac{\Delta TestScore}{\Delta STR}$
0	–1.12
20%	–1.12+.0012×20 = –1.10

***Example, ctd**: hypothesis tests*

$$\widehat{TestScore} = \underset{(11.8)}{686.3} - \underset{(0.59)}{1.12}STR - \underset{(0.37)}{0.67}PctEL + \underset{(0.019)}{.0012}(STR\times PctEL),$$

- Does population coefficient on $STR\times PctEL = 0$?

  $t = .0012/.019 = .06 \Rightarrow$ can't reject null at 5% level
- Does population coefficient on $STR = 0$?

  $t = -1.12/0.59 = -1.90 \Rightarrow$ can't reject null at 5% level
- Do the coefficients on ***both*** *STR* ***and*** $STR\times PctEL = 0$?

  $F = 3.89$ (p-value = .021) $\Rightarrow$ reject null at 5% level(!!)

  (Why?)

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

Nonlinear specifications let us examine more nuanced questions about the Test score – *STR* relation, such as:

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*?)

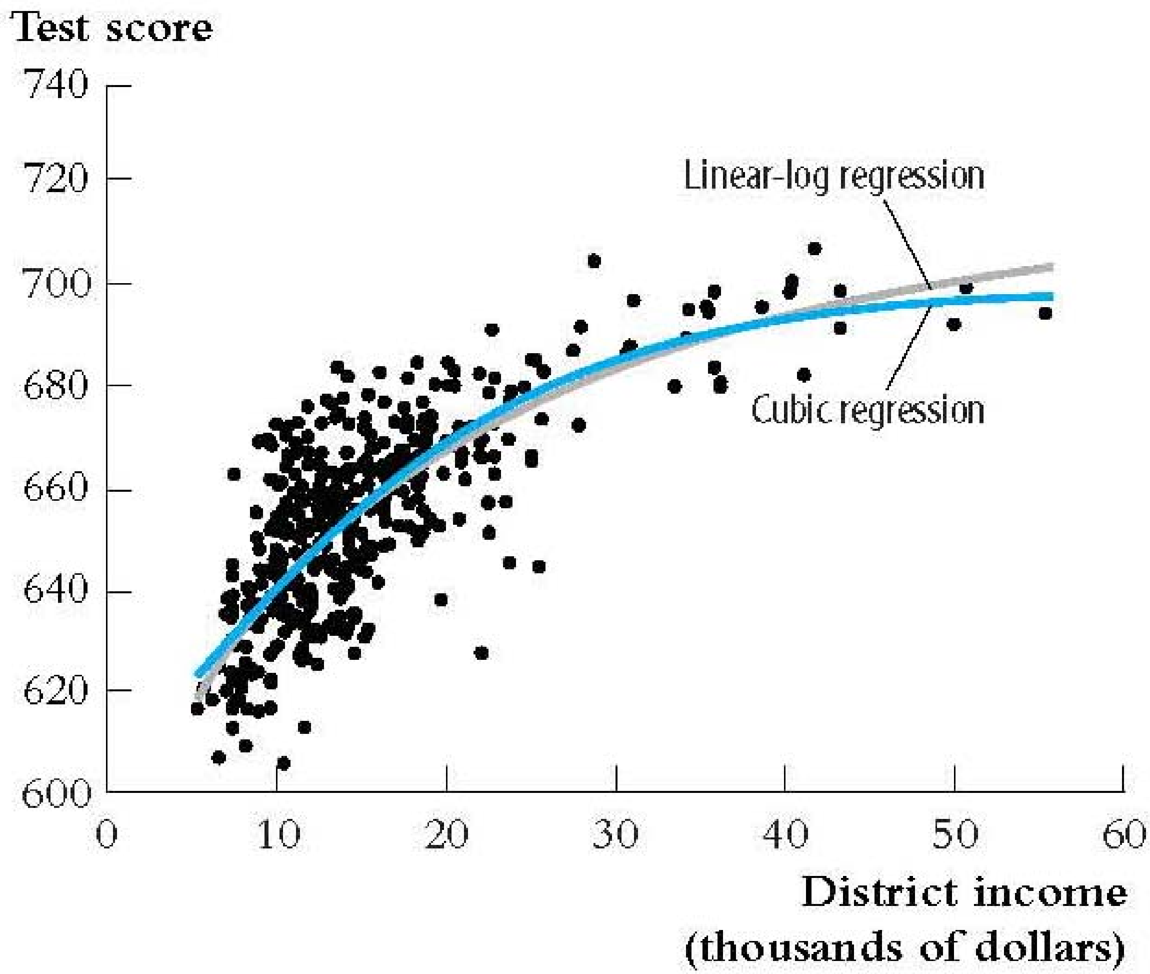
- Estimate linear and nonlinear functions of *STR*, holding constant relevant demographic variables
    - *PctEL*
    - *Income* (remember the nonlinear *TestScore-Income* relation!)
    - *LunchPCT* (fraction on free/subsidized lunch)
- See whether adding the nonlinear terms makes an "economically important"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

Strategy for Question #2 (interactions between *PctEL* and *STR***?)**

- Estimate linear and nonlinear functions of *STR*, interacted with *PctEL*.
- If the specification is nonlinear (with STR, STR^2, STR^3), then you need to add interactions with all the terms so that the entire functional form can be different, depending on the level of *PctEL*.
- We will use a binary-continuous interaction specification by adding $HiEL \times STR$, $HiEL \times STR^2$, and $HiEL \times STR^3$.

What is a good “base” specification?

The *TestScore* – *Income* relation:



The logarithmic specification is better behaved near the extremes of the sample, especially for large values of income.

TABLE 8.3 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	700.2** (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)

***Tests of joint hypotheses*:**

F-Statistics and p-Values on Joint Hypotheses

(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
$\overline{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

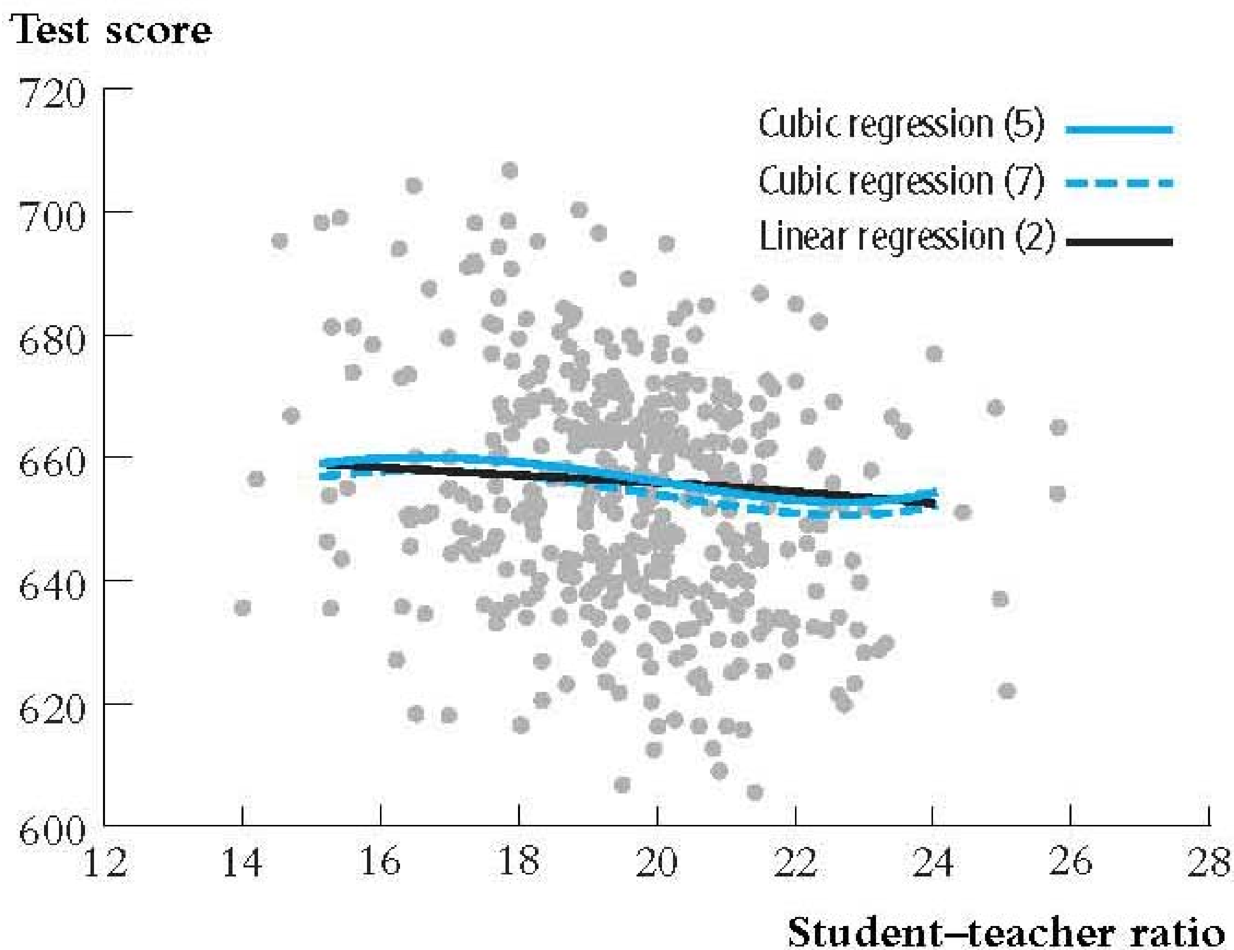
These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and p-values are given in parentheses under F-statistics. Individual coefficients are statistically significant at the *5% or **1% significance level.

What can you conclude about question #1?

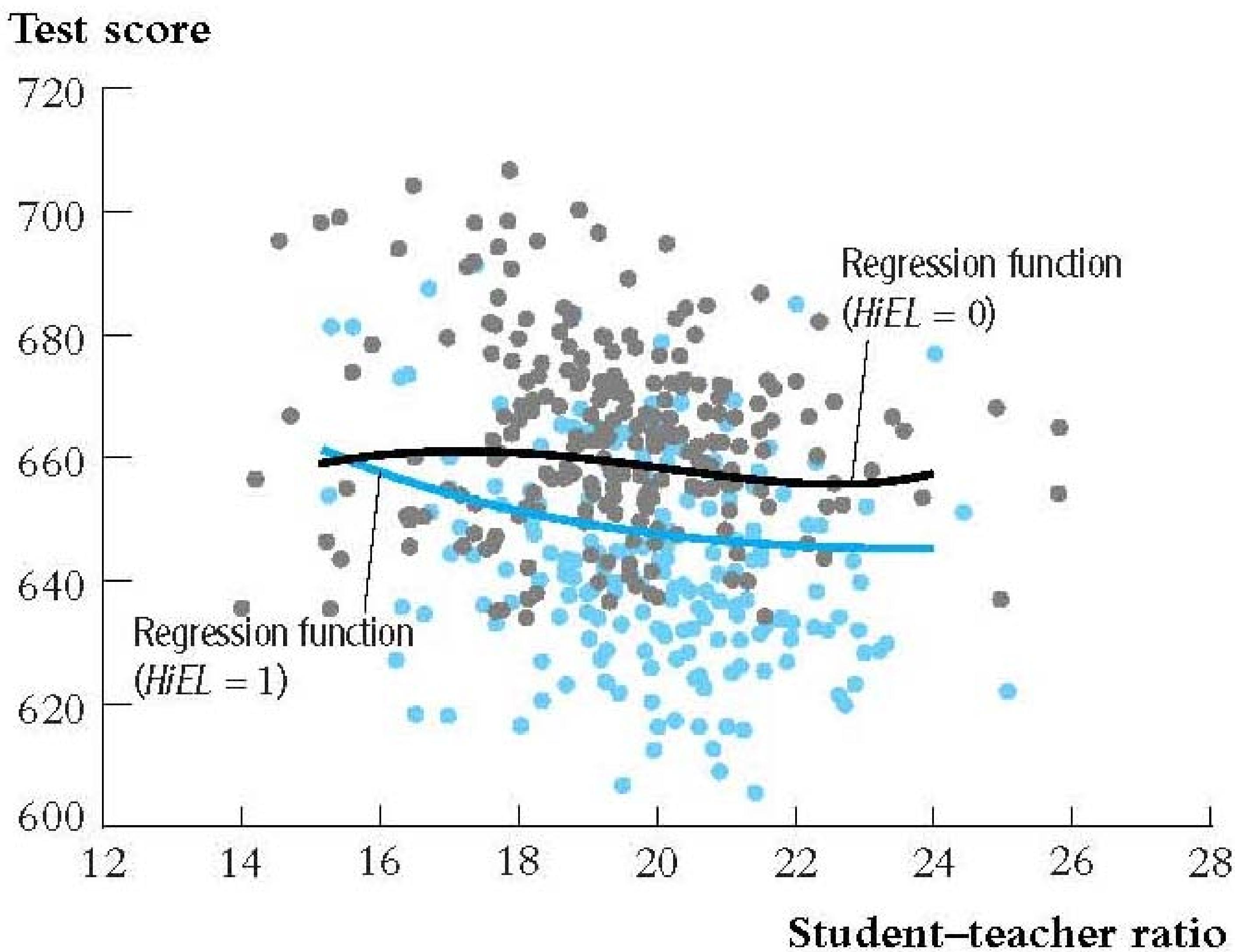
About question #2?

***Interpreting the regression functions via plots*:**

First, compare the linear and nonlinear specifications:



Next, compare the regressions with interactions:



Addendum

Fun with logarithms

- Y and/or X is transformed by taking its logarithm
- this gives a "percentages" interpretation that makes sense in many applications

2. Logarithmic functions of *Y* and/or *X*

- $\ln(X)$ = the natural logarithm of X
- Logarithmic transforms permit modeling relations in "percentage" terms (like elasticities), rather than linearly.

Key result (recall from calculus):

For small changes, the change in the log is approximately the percent change (expressed as a decimal).

The three log regression specifications:

Case	Population regression function
I. linear-log	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$
II. log-linear	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

- The interpretation of the slope coefficient differs in each case.
- The interpretation is found by applying the general "before and after" rule: "figure out the change in Y for a given change in X."
- Each case has a natural interpretation (for small changes in X)

I.Linear-log population regression function

a 1% increase in X (multiplying X by* 1.01*)
***is associated with a* .01β_1 *change in Y*.**

$$\begin{aligned}(1\% \text{ increase in } X &\Rightarrow .01 \text{ increase in } \ln(X)\\ &\Rightarrow .01\beta_1 \text{ increase in } Y)\end{aligned}$$

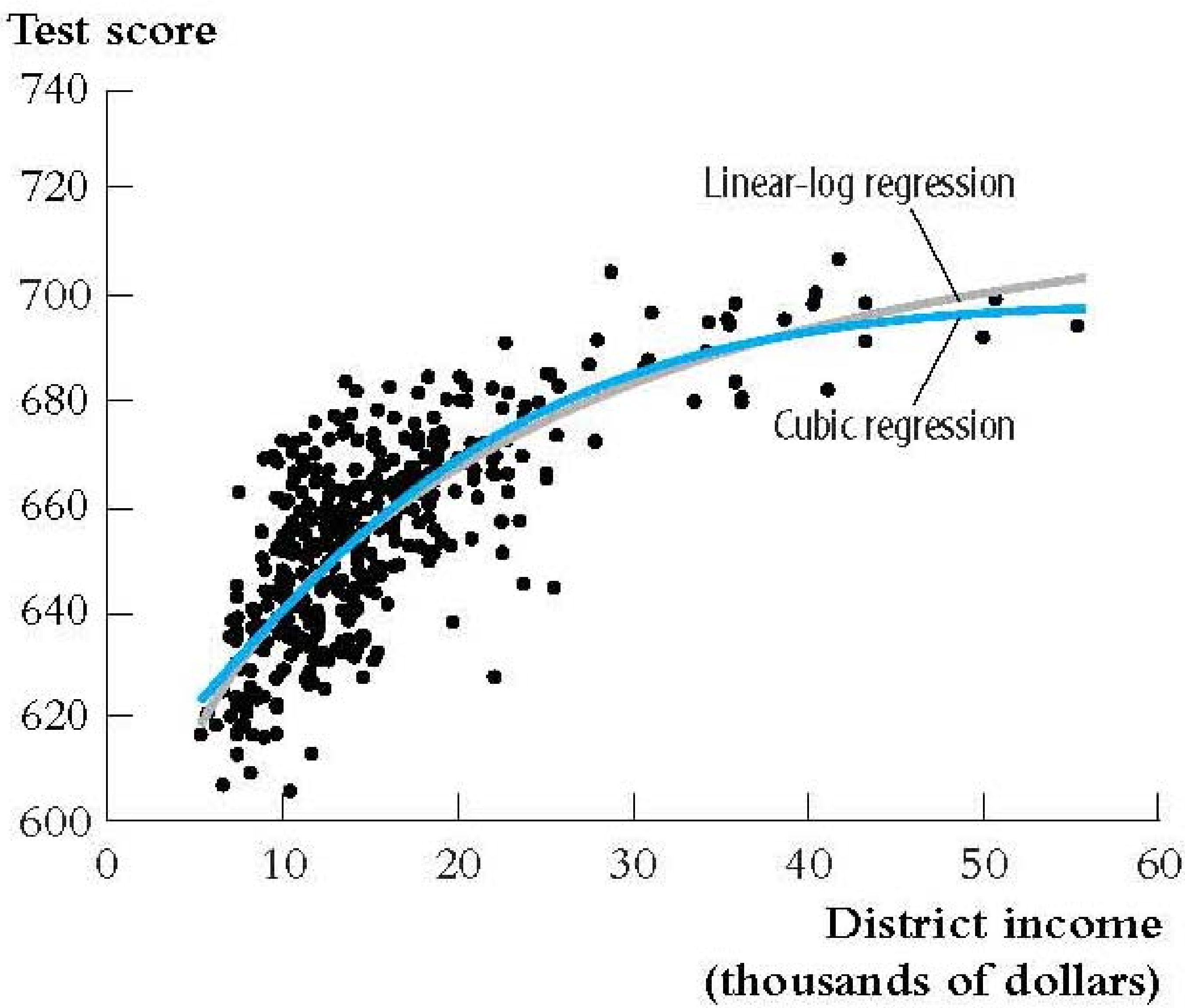
Example: TestScore vs. ln(Income)

- 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:

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

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

The linear-log and cubic regression functions



II. Log-linear population regression function

***a change in X by one unit* (ΔX = 1)**
is associated with a* $100\beta_1$*% change in Y

- 1 unit increase in $X \Rightarrow \beta_1$ increase in $\ln(Y)$

$$\Rightarrow 100\beta_1\% \text{ increase in } Y$$

III. Log-log population regression function

a 1% change in X is associated with a β_1% change in Y.

In the log-log specification, β_1 has the interpretation of an elasticity.

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

- First define a new dependent variable, ln(*TestScore*), ***and*** the new regressor, ln(*Income*)
- The model is now a linear regression of ln(*TestScore*) against ln(*Income*), which can be estimated by OLS:

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

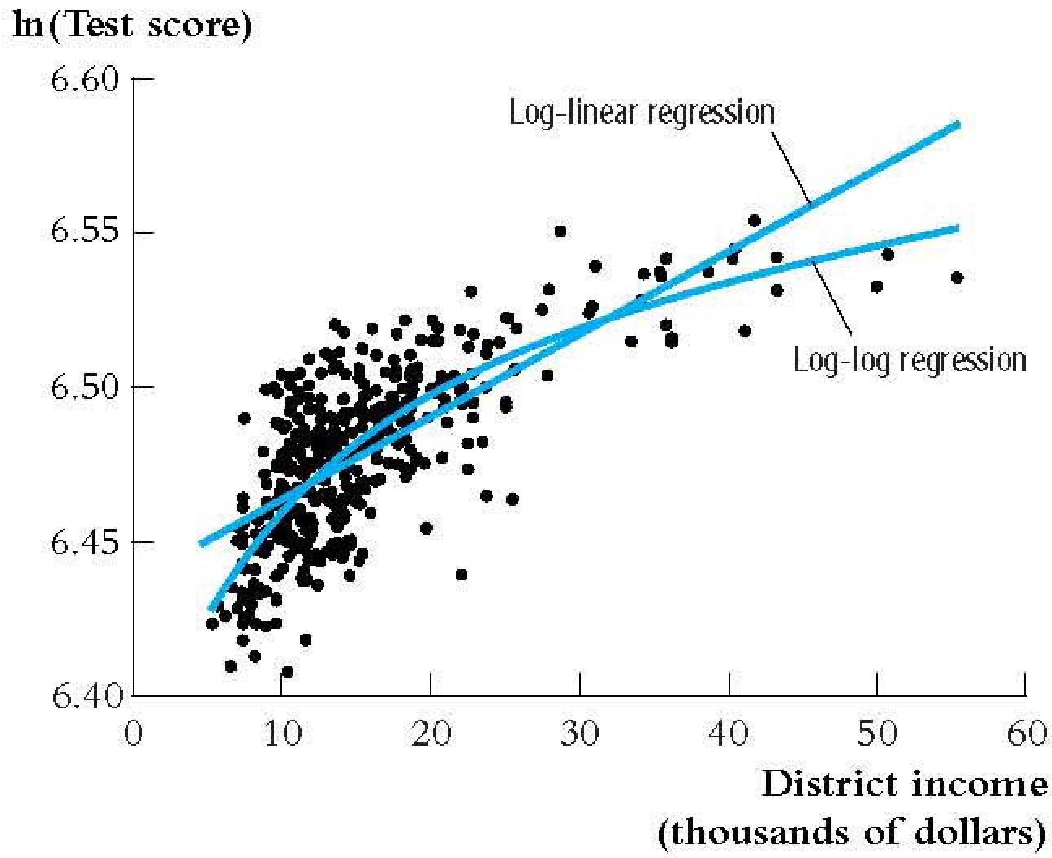
An 1% increase in *Income* is associated with an increase of .0554% in *TestScore* (*Income* up by a factor of 1.01, *TestScore* up by a factor of 1.000554)

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

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

- For example, suppose income increases from \$10,000 to \$11,000, or by 10%. Then *TestScore* increases by approximately .0554×10% = .554%. If *TestScore* = 650, this corresponds to an increase of .00554×650 = 3.6 points.
- How does this compare to the log-linear model?

The log-linear and log-log specifications:



- *Note vertical axis*
- *Neither seems to fit as well as the cubic or linear-log, at least based on visual inspection (formal comparison is difficult because the dependent variables differ)*

Summary: Logarithmic transformations

- Three cases, differing in whether Y and/or X is transformed by taking logarithms.
- The regression is linear in the new variable(s) $\ln(Y)$ and/or $\ln(X)$, and the coefficients can be estimated by OLS.
- Hypothesis tests and confidence intervals are now implemented and interpreted "as usual."
- The interpretation of β_1 differs from case to case.

The choice of specification (functional form) should be guided by judgment (which interpretation makes the most sense in your application?), tests, and plotting predicted values

Other nonlinear functions (and nonlinear least squares)

The foregoing regression functions have limitations…

- Polynomial: test score can decrease with income
- Linear-log: test score increases with income, but without bound
- Here is a nonlinear function in which Y always increases with X *and* there is a maximum (asymptote) value of Y:

$$Y = \beta_0 - \alpha e^{-\beta_1 X}$$

β_0, β_1, and α are unknown parameters. This is called a negative exponential growth curve. The asymptote as $X \to \infty$ is β_0.

Negative exponential growth

We want to estimate the parameters of

$$Y_i = \beta_0 - \alpha e^{-\beta_1 X_i} + u_i$$

Compare to linear-log or cubic models:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_2 X_i^3 + u_i$$

Linear-log and polynomial models are *linear in the parameters* β_0 and β_1, but the negative exponential model is not.

Nonlinear Least Squares

- Models that are linear in the parameters can be estimated by OLS.
- Models that are nonlinear in one or more parameters can be estimated by nonlinear least squares (NLS) (but not by OLS)
- What is the NLS problem for the proposed specification?

- This is a nonlinear minimization problem (a "hill-climbing" problem). How could you solve this?
  - Guess and check
  - There are better ways…
  - Implementation ...

Negative exponential growth; *RMSE* = 12.675

Linear-log; *RMSE* = 12.618 (oh well…)