General Linear Models

Nonlinear Regression Functions

– O.Örsan Özener

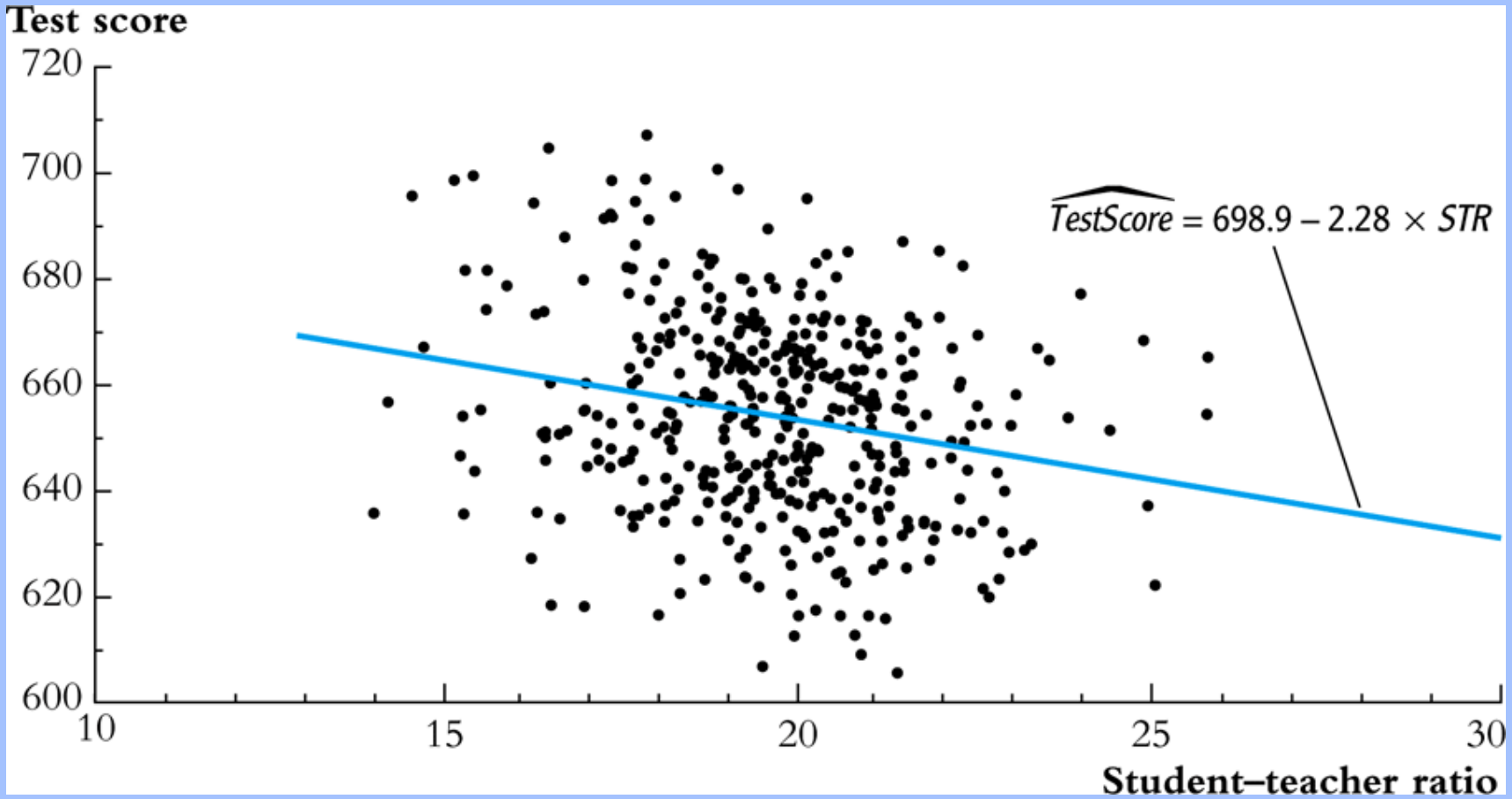
Outline

1. Nonlinear regression functions – general comments
2. Nonlinear functions of one variable
3. Nonlinear functions of two variables: interactions
4. Application to the California Test Score data set

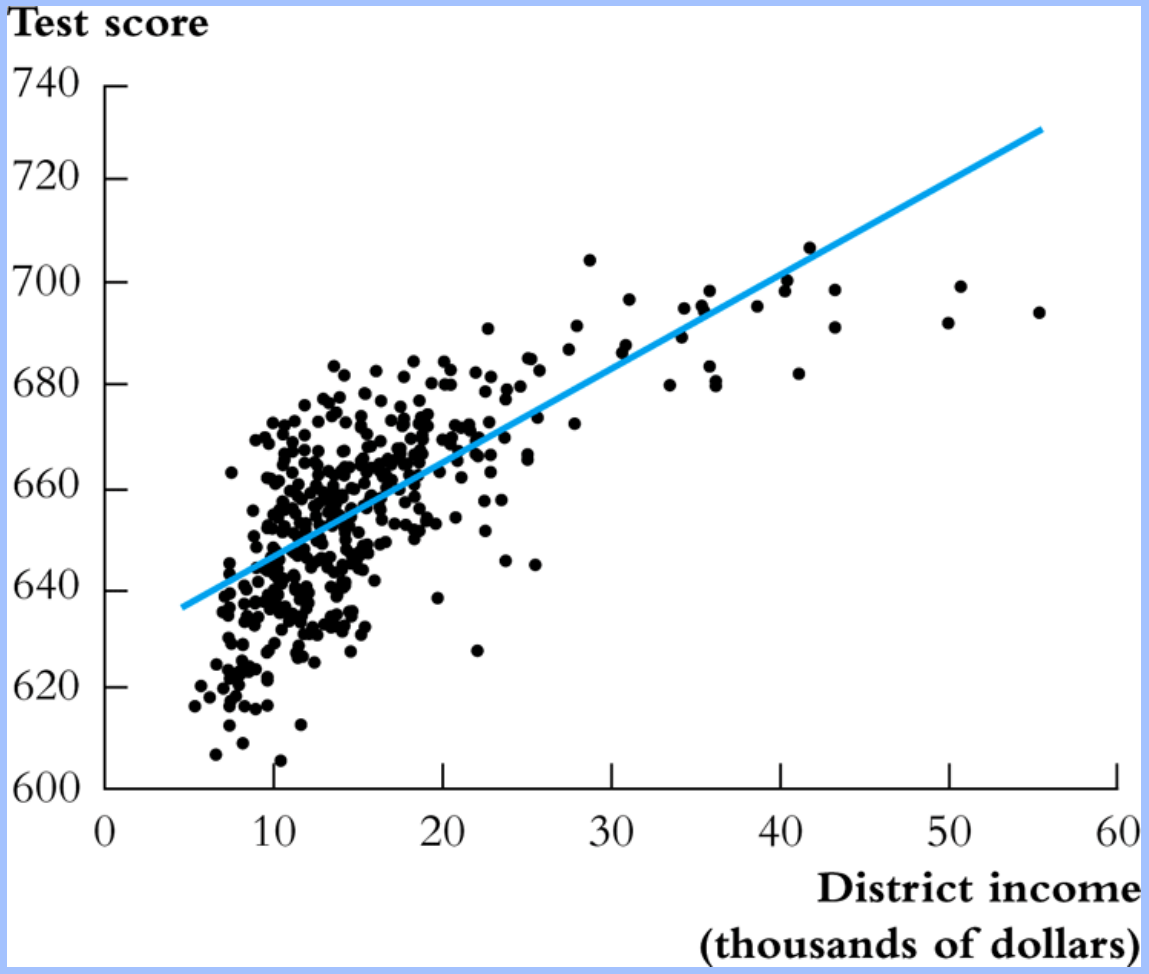
Nonlinear regression functions

- The regression functions so far have been linear in the *X*'s
- But the linear approximation is not always a good one
- The multiple regression model can handle regression functions that are nonlinear in one or more *X*.

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



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



Nonlinear Regression Population Regression Functions – 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); implies that f is the conditional expectation of Y given the X's.
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).

The change in Y associated with a change in X_1, holding $X_2,\ldots, X_k$ constant is:

$$\Delta Y = f(X_1 + \Delta X_1, X_2,\ldots, X_k) - f(X_1, X_2,\ldots, X_k)$$

The Expected Effect on Y of a Change in X_1 in the Nonlinear Regression Model (8.3)

KEY CONCEPT 8.1

The expected change in Y, ΔY, associated with the change in X_1, ΔX_1, holding $X_2, \ldots, X_k$ constant, is the difference between the value of the population regression function before and after changing X_1, holding $X_2, \ldots, X_k$ constant. That is, the expected change in Y is the difference:

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k) - f(X_1, X_2, \ldots, X_k). \quad (8.4)$$

The estimator of this unknown population difference is the difference between the predicted values for these two cases. Let $\hat{f}(X_1, X_2, \ldots, X_k)$ be the predicted value of Y based on the estimator $\hat{f}$ of the population regression function. Then the predicted change in Y is

$$\Delta \hat{Y} = \hat{f}(X_1 + \Delta X_1, X_2, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k). \quad (8.5)$$

Nonlinear Functions of a Single Independent Variable

We'll look at two complementary approaches:

1. Polynomials in *X*

   The population regression function is approximated by a quadratic, cubic, or higher-degree polynomial

2. Logarithmic transformations

   Y and/or *X* is transformed by taking its logarithm

   this gives a "percentages" interpretation that makes sense in many applications

1. Polynomials in *X*

Approximate the population regression function by a polynomial:

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

- 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

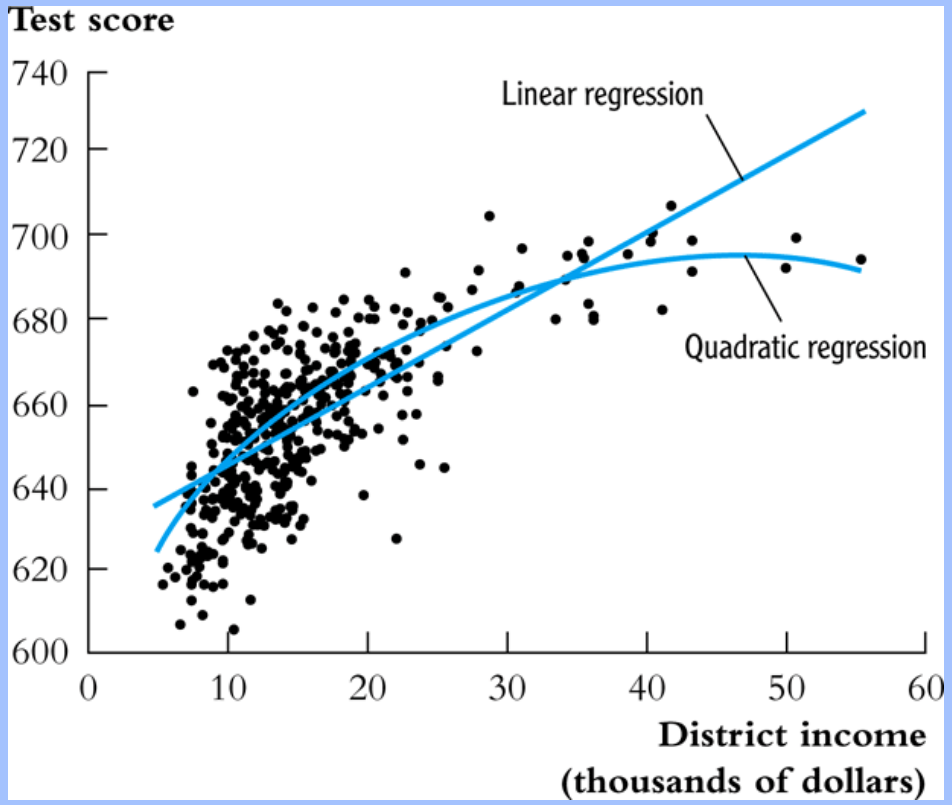
```
generate avginc2 = avginc*avginc;       Create a new regressor
reg testscr avginc avginc2, r;

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

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



Interpreting the estimated regression function, ctd:

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

$$\underset{}{\textit{Test Score}} = 607.3 + 3.85\underset{(2.9)}{\textit{Income}_i} - 0.0423\underset{(0.0048)}{(\textit{Income}_i)^2}$$
$$\qquad\qquad\qquad\qquad\qquad\quad (0.27)$$

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

$$\begin{aligned}\Delta \textit{Test Score} &= 607.3 + 3.85\times 6 - 0.0423\times 6^2 \\ &\quad - (607.3 + 3.85\times 5 - 0.0423\times 5^2) \\ &= 3.4\end{aligned}$$

$$\textit{Test Score} = 607.3 + 3.85\textit{Income}_i - 0.0423(\textit{Income}_i)^2$$

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

Change in *Income* ($1000 per capita)	Δ *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;
reg testscr avginc avginc2 avginc3, r;
```

Create the cubic regressor

```
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.

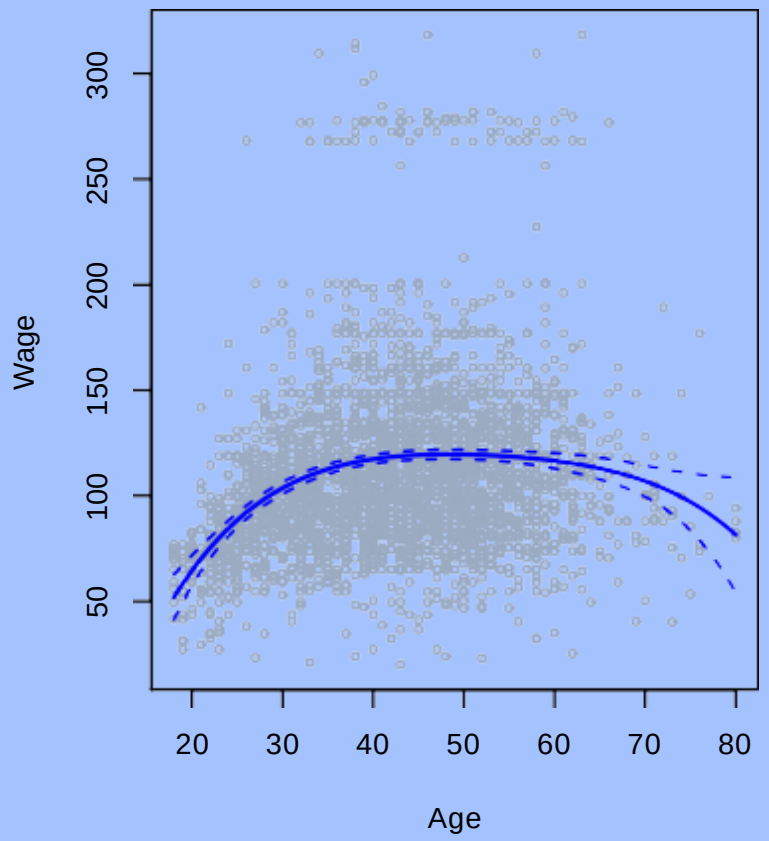
Summary: polynomial regression functions

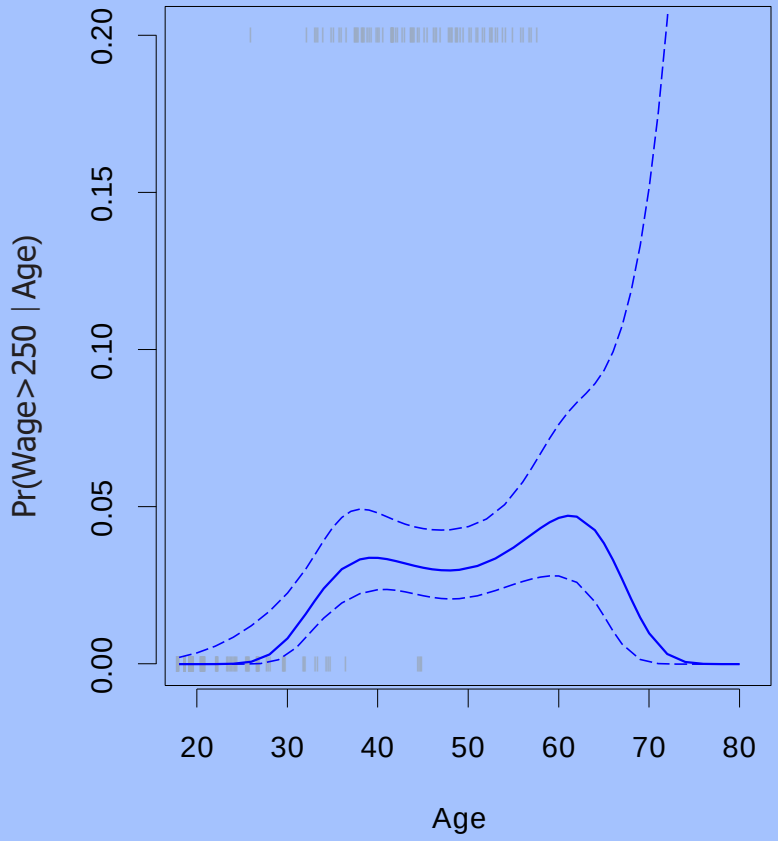
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 \ +\ldots+ \beta_r \ + u_i$$

- 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.

Wage Data

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i$$

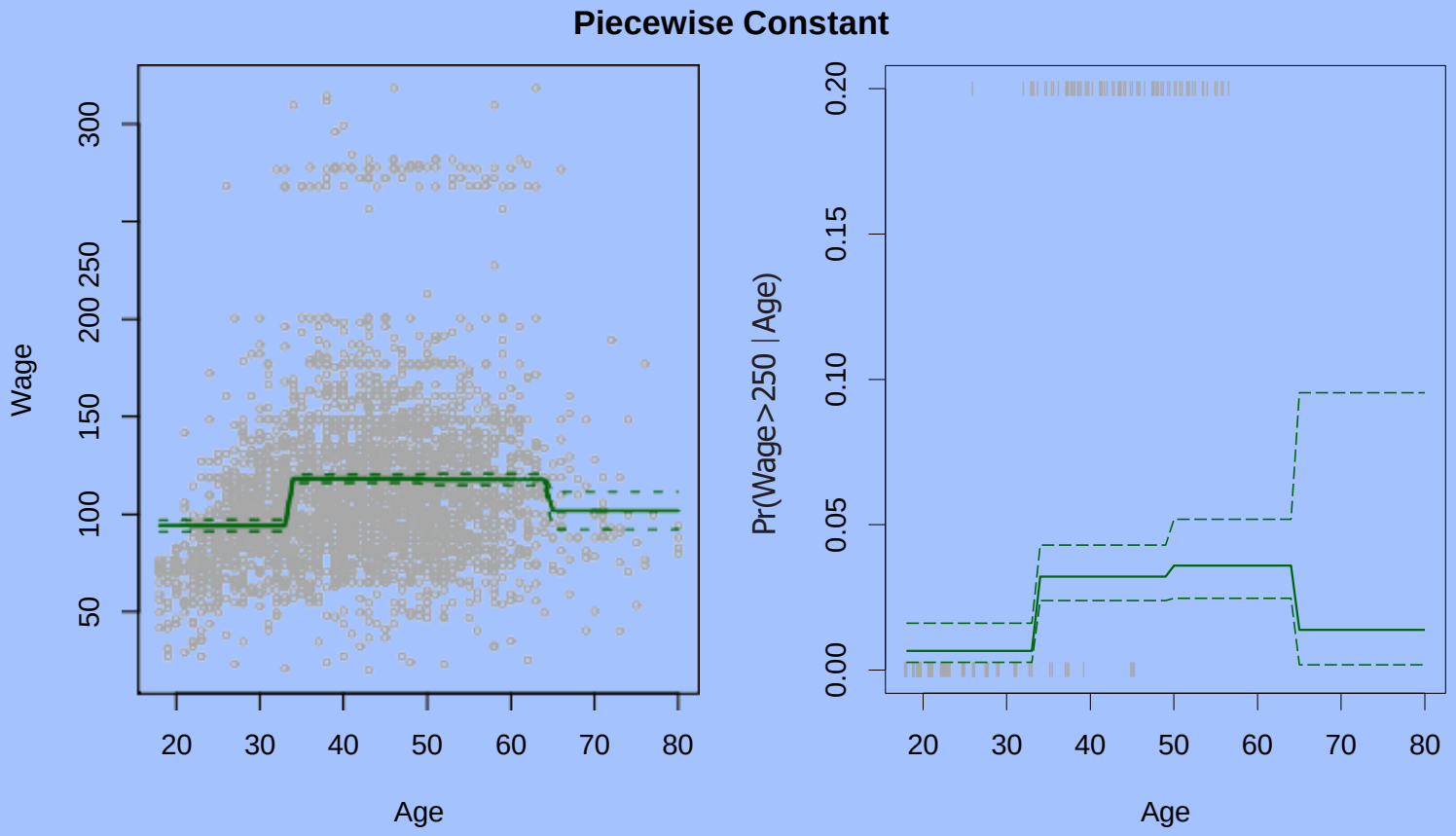




Step Functions

Another way of creating transformations of a variable — cut the variable into distinct regions.

$$C_1(X) = I(X < 35), \quad C_2(X) = I(35 \leq X < 50), \ldots, C_3(X) = I(X \geq 65)$$



Step Functions

- Easy to work with. Creates a series of dummy variables representing each group.
- Useful way of creating interactions that are easy to interpret. For example, interaction effect of Year and Age:

$$I(\text{Year} < 2005) \cdot \text{Age}, \quad I(\text{Year} \geq 2005) \cdot \text{Age}$$

  would allow for different linear functions in each age category.
- In R: I(year < 2005) or cut(age, c(18, 25, 40, 65, 90)).
- Choice of cutpoints or *knots* can be problematic. For creating nonlinearities, smoother alternatives such as *splines* are available.

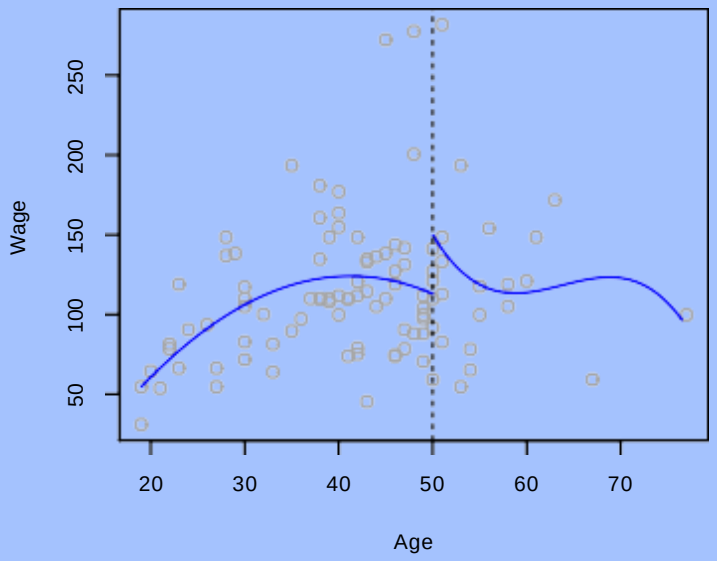
Piecewise Polynomials

- Instead of a single polynomial in X over its whole domain, we can rather use different polynomials in regions defined by knots.

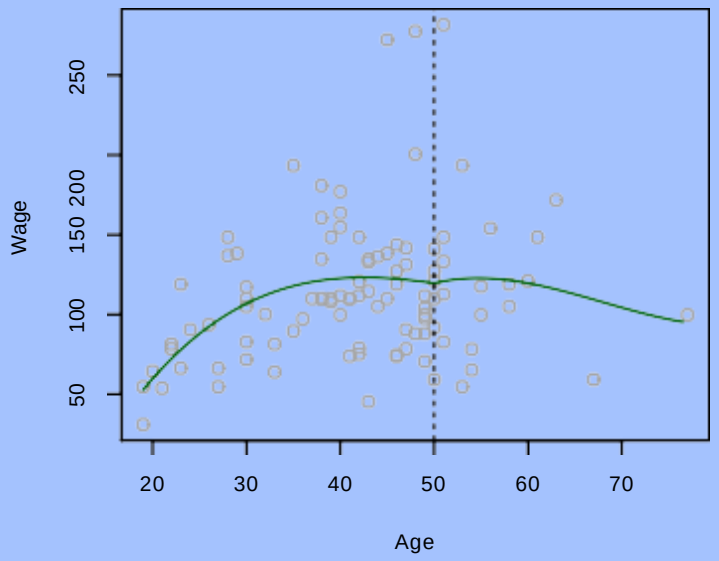
$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & \text{if } x_i < c; \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & \text{if } x_i \geq c. \end{cases}$$

- Better to add constraints to the polynomials, e.g. continuity.
- *Splines* have the "maximum" amount of continuity.

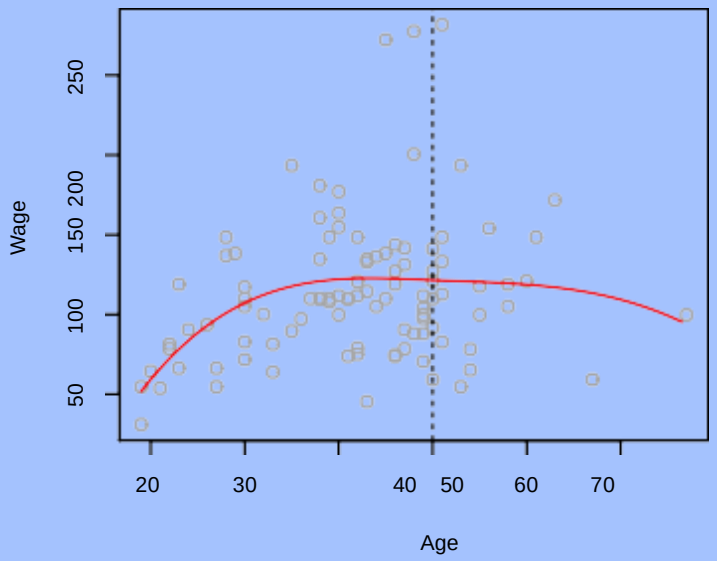
Piecewise Cubic



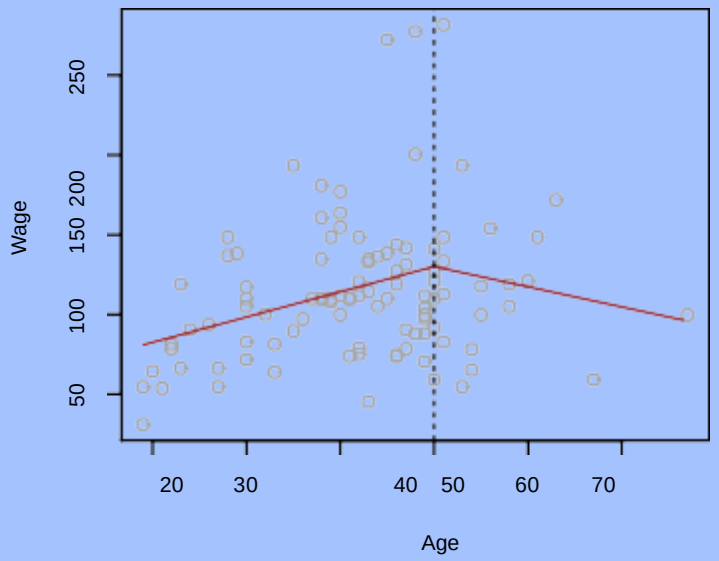
Continuous Piecewise Cubic



Cubic Spline



Linear Spline



Linear Splines

A linear spline with knots at ξ_k, $k = 1, \ldots, K$ is a piecewise linear polynomial continuous at each knot.

We can represent this model as

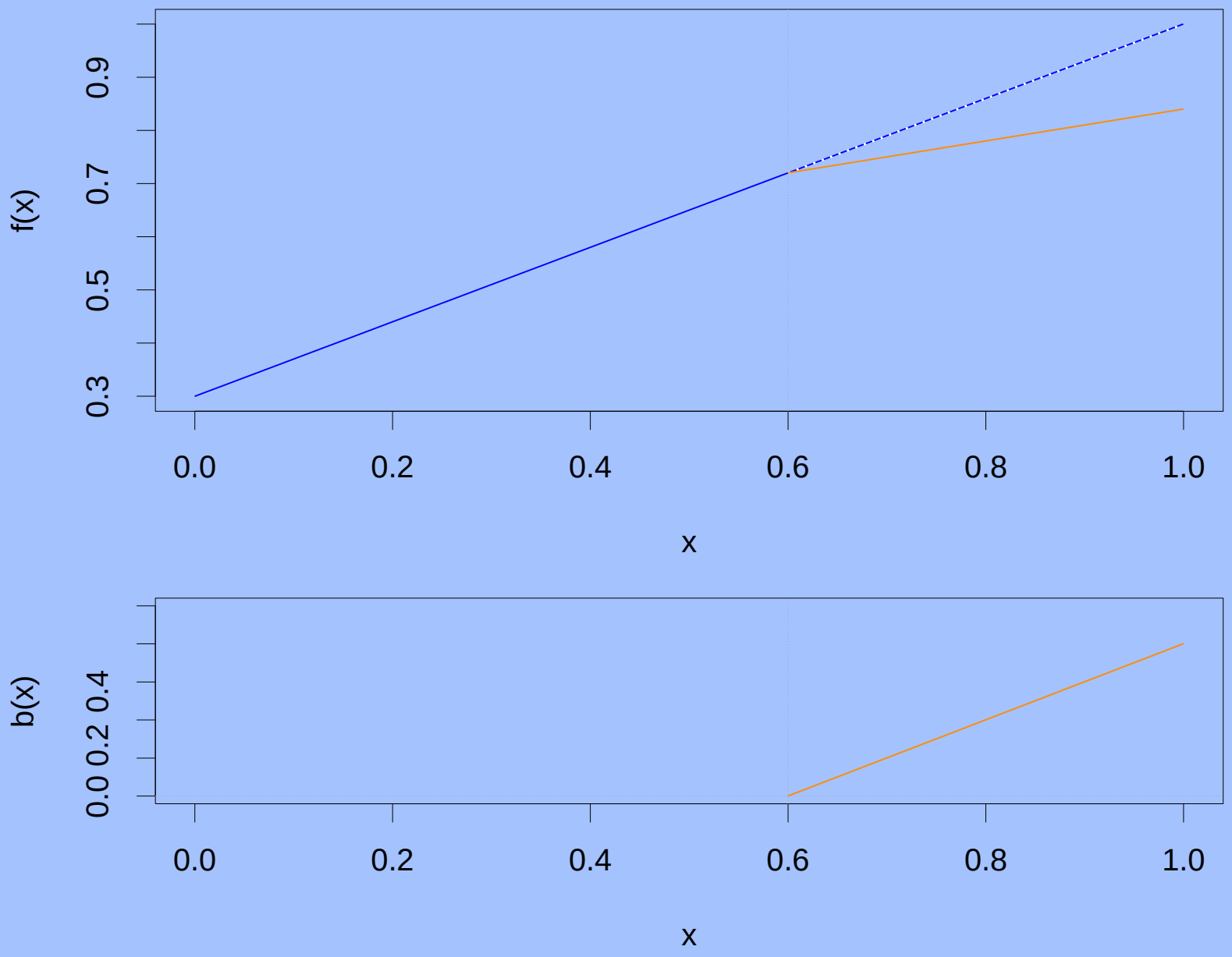
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

where the b_k are *basis functions*.

$$\begin{aligned} b_1(x_i) &= x_i \\ b_{k+1}(x_i) &= (x_i - \xi_k)_+, \quad k = 1, \ldots, K \end{aligned}$$

Here the $()_+$ means *positive part*; i.e.

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$


f(x)
0.3
0.5
0.7
0.9
0.0
0.2
0.4
0.6
0.8
1.0
x
b(x)
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
x

Cubic Splines

A cubic spline with knots at ξ_k, $k = 1, \ldots, K$ is a piecewise cubic polynomial with continuous derivatives up to order 2 at each knot.

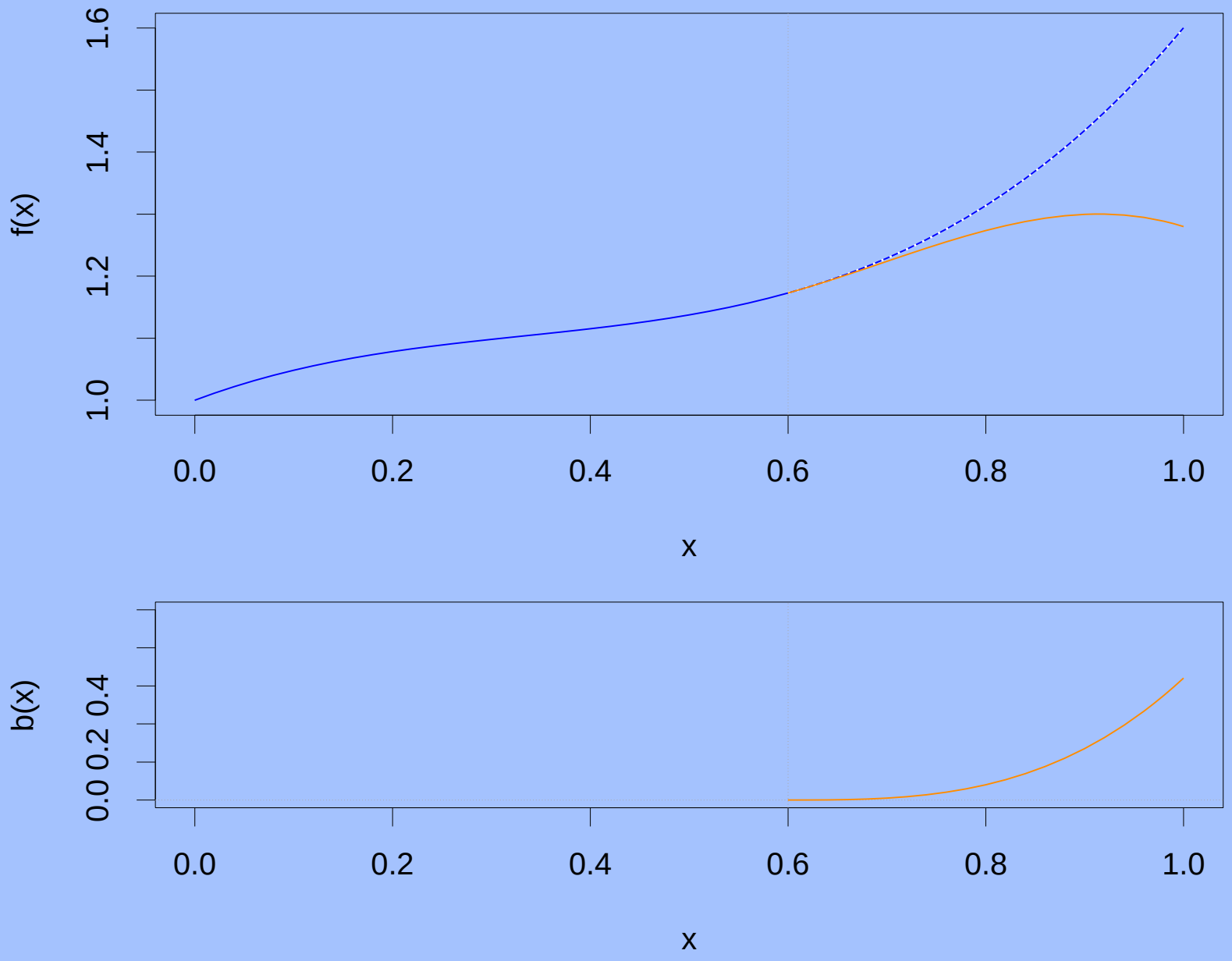
Again we can represent this model with truncated power basis functions

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i,$$

$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3, \quad k = 1, \ldots, K \end{aligned}$$

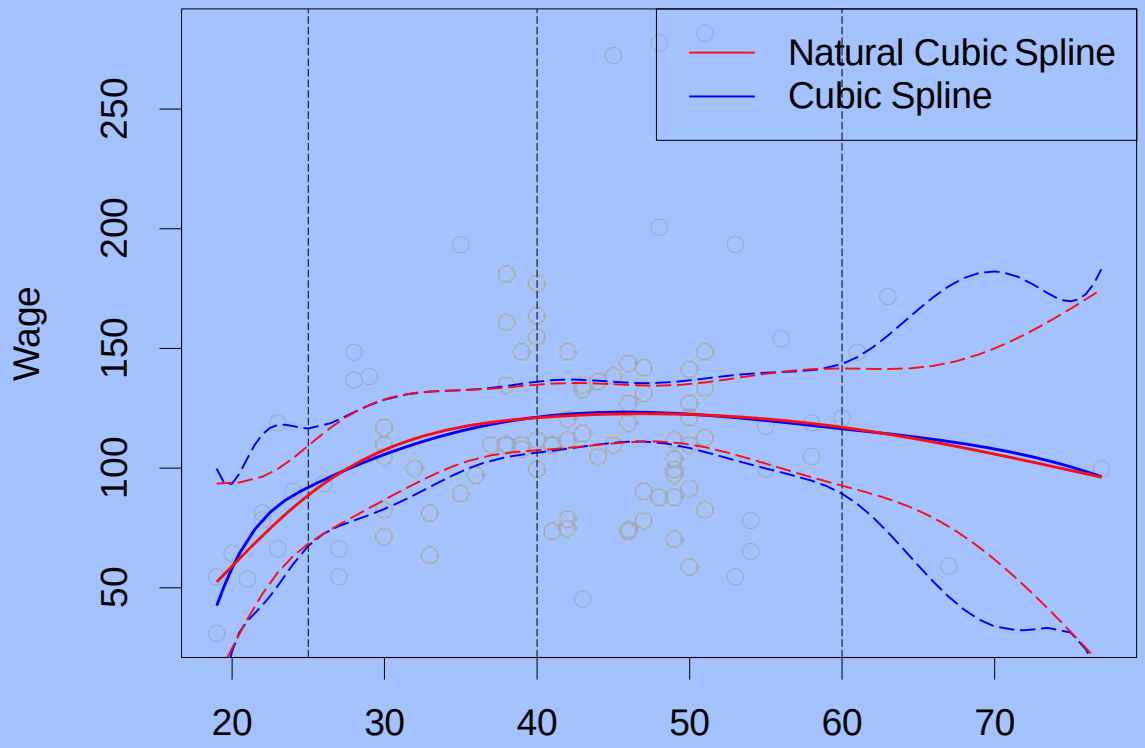
where

$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & \text{if } x_i > \xi_k \\ 0 & \text{otherwise} \end{cases}$$

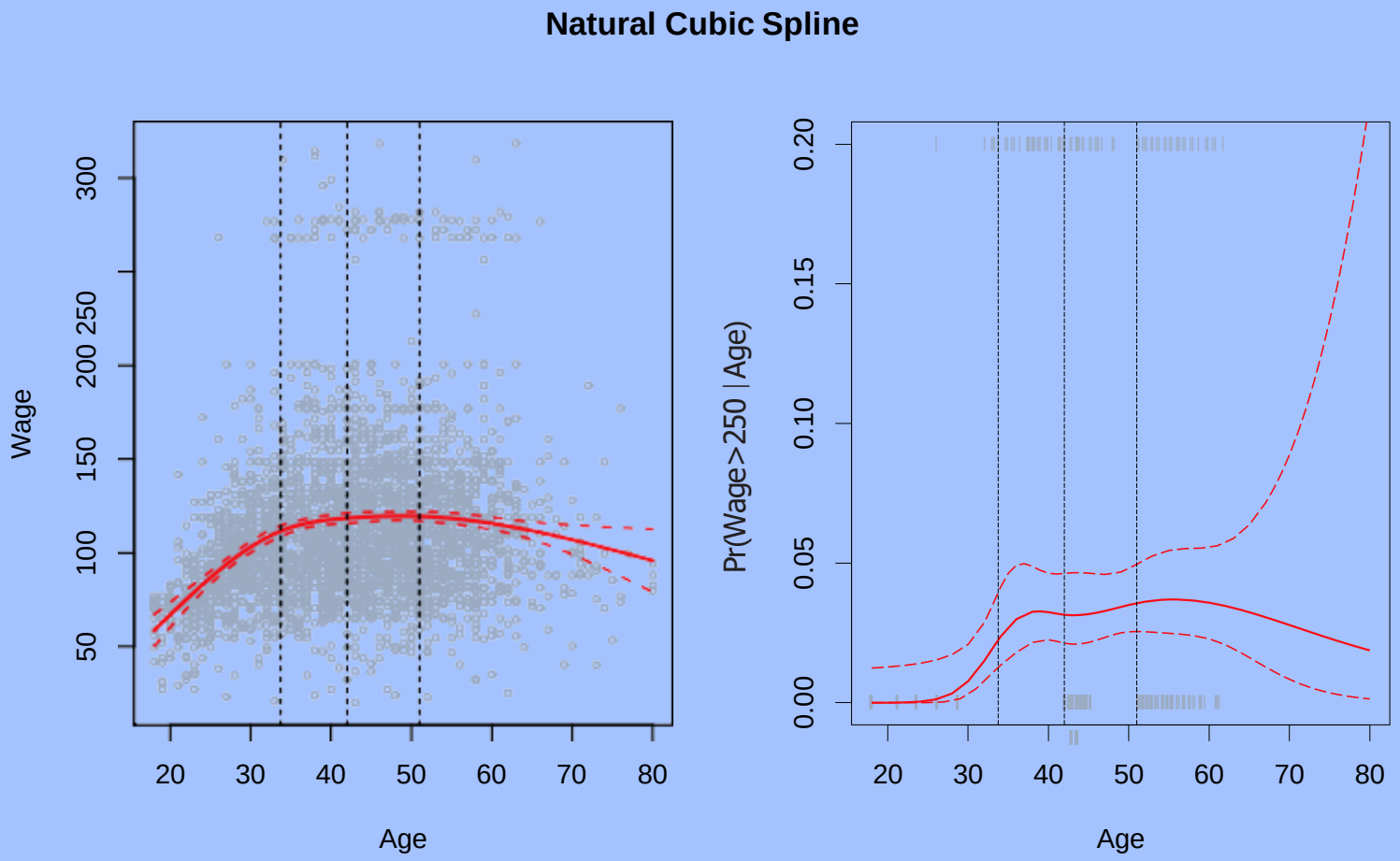

f(x)
1.0
1.2
1.4
1.6
0.0
0.2
0.4
0.6
0.8
1.0
x
b(x)
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
x

Natural Cubic Splines

A natural cubic spline extrapolates linearly beyond the boundary knots. This adds $4 = 2 \times 2$ extra constraints, and allows us to put more internal knots for the same degrees of freedom as a regular cubic spline.

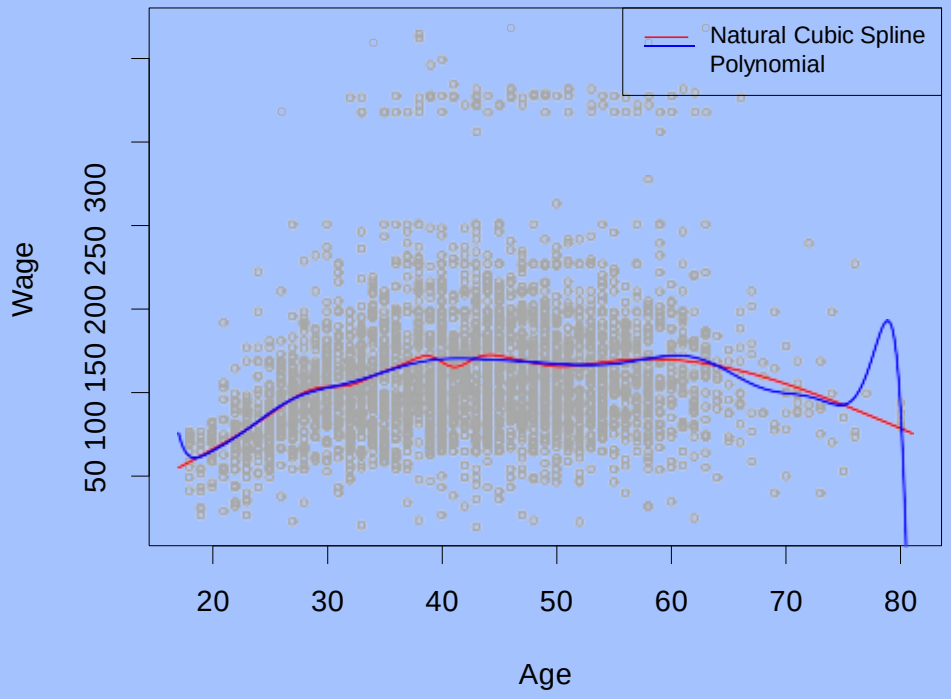


Natural Cubic Splines



Knot Placement

- One strategy is to decide K, the number of knots, and then place them at appropriate quantiles of the observed X.
- A cubic spline with K knots has $K + 4$ parameters or degrees of freedom.
- A natural spline with K knots has K degrees of freedom.



Comparison of a degree-14 polynomial and a natural cubic spline, each with 15df.

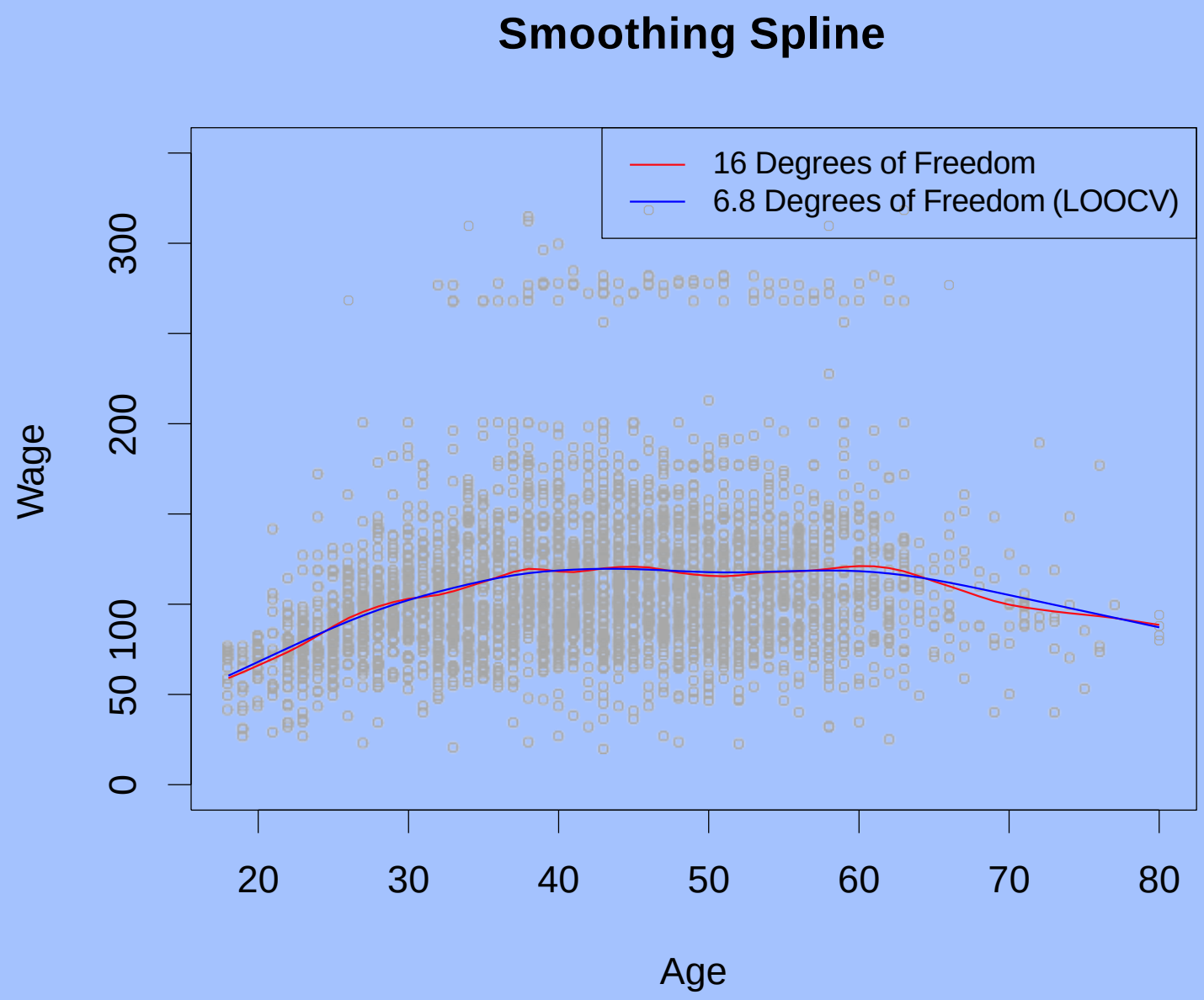
```
ns(age,df=14)
poly(age,deg=14)
```

Smoothing Splines

Consider this criterion for fitting a smooth function $g(x)$ to some data:

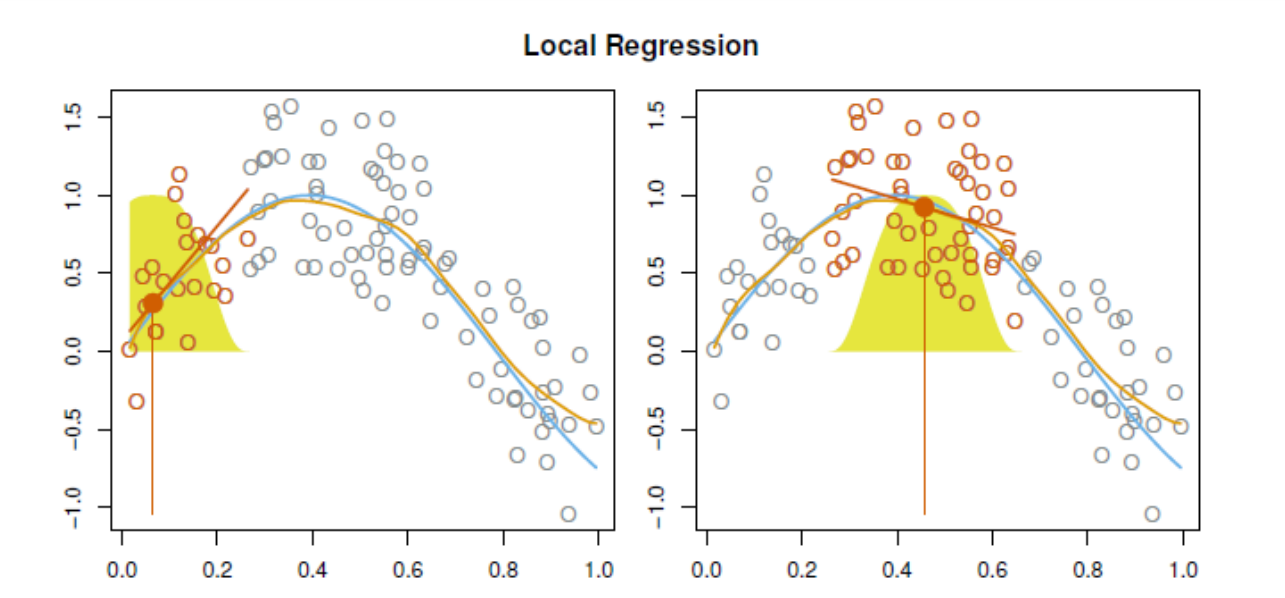
$$\underset{g \in \mathcal{S}}{\text{minimize}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- The first term is RSS, and tries to make $g(x)$ match the data at each x_i.
- The second term is a *roughness penalty* and controls how wiggly $g(x)$ is. It is modulated by the *tuning parameter* $\lambda \geq 0$.
  - The smaller λ, the more wiggly the function, eventually interpolating y_i when $\lambda = 0$.
  - As $\lambda \to \infty$, the function $g(x)$ becomes linear.
  - Smoothing splines avoid the knot-selection issue, leaving a single λ to be chosen.


Smoothing Spline
16 Degrees of Freedom
6.8 Degrees of Freedom (LOOCV)
Wage
0
50
100
200
300
20
30
40
50
60
70
80
Age

Local Regression

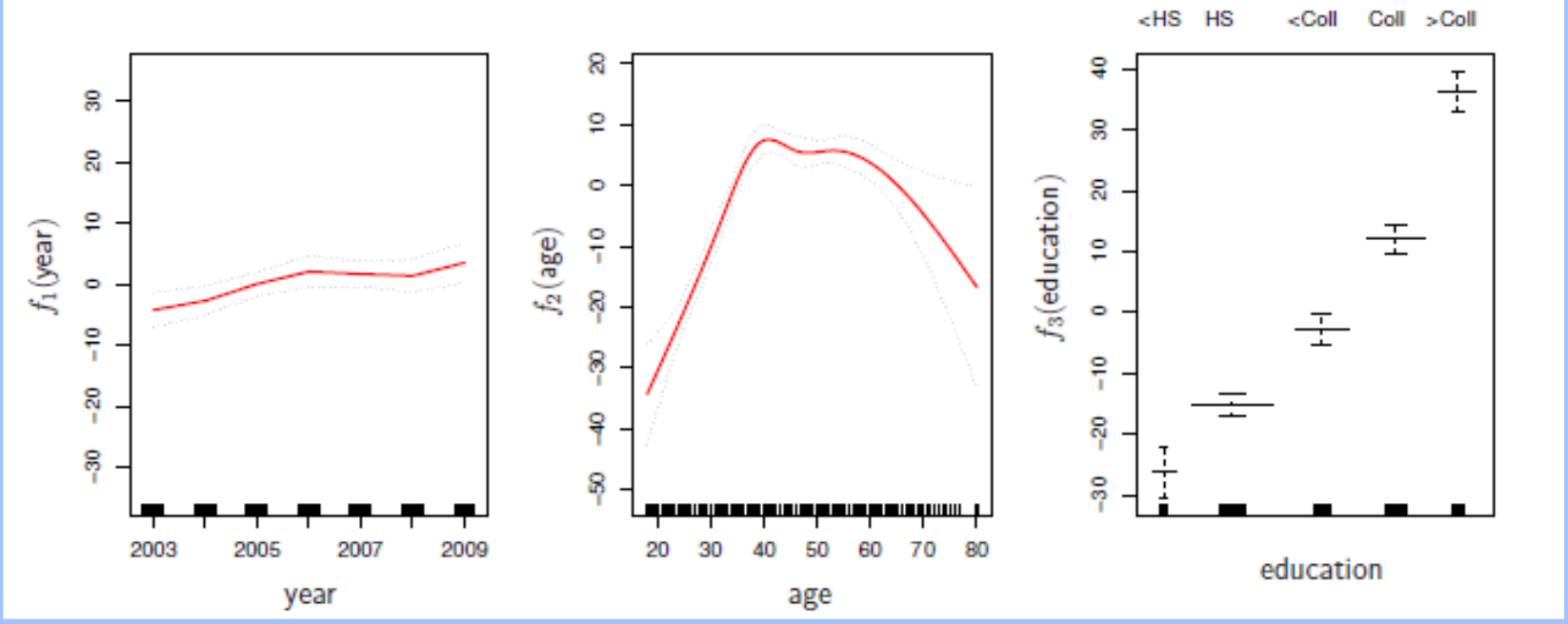
With a sliding weight function we fit separate linear fits over the range of X by weighted least squares



Generalized Additive Models

Allows for flexible nonlinearities in several variables, but retains the additive structure of linear models

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i.$$



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.

Here's why: $\ln(x+\Delta x) - \ln(x) = \ln\left(1+\frac{\Delta x}{x}\right) \cong \frac{\Delta x}{x}$

(calculus: $\frac{d\ln(x)}{dx} = \frac{1}{x}$)

Numerically:

ln(1.01) = .00995 ≅ .01;

ln(1.10) = .0953 ≅ .10 (sort of)

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

Compute Y "before" and "after" changing X:

$Y = \beta_0 + \beta_1 \ln(X)$ ("before")

Now change X: $Y + \Delta Y = \beta_0 + \beta_1 \ln(X + \Delta X)$ ("after")

Subtract ("after") – ("before"): $\Delta Y = \beta_1[\ln(X + \Delta X) - \ln(X)]$

now $\ln(X + \Delta X) - \ln(X) \cong \frac{\Delta X}{X}$,

so $\Delta Y \cong \beta_1 \frac{\Delta X}{X}$

or $\beta_1 \cong \frac{\Delta Y}{\Delta X / X}$ (small ΔX)

Linear-log case, continued

$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$

for small ΔX,

$$\beta_1 \cong \frac{\Delta Y}{\Delta X / X}$$

Now $100 \times \frac{\Delta X}{X}$ = percentage change in X, so ***a 1% increase in X (multiplying X by* 1.01*) is associated with a .01β_1 change in Y*.**

(1% increase in X --> .01 increase in $\ln(X)$
--> .01β_1 increase in Y)

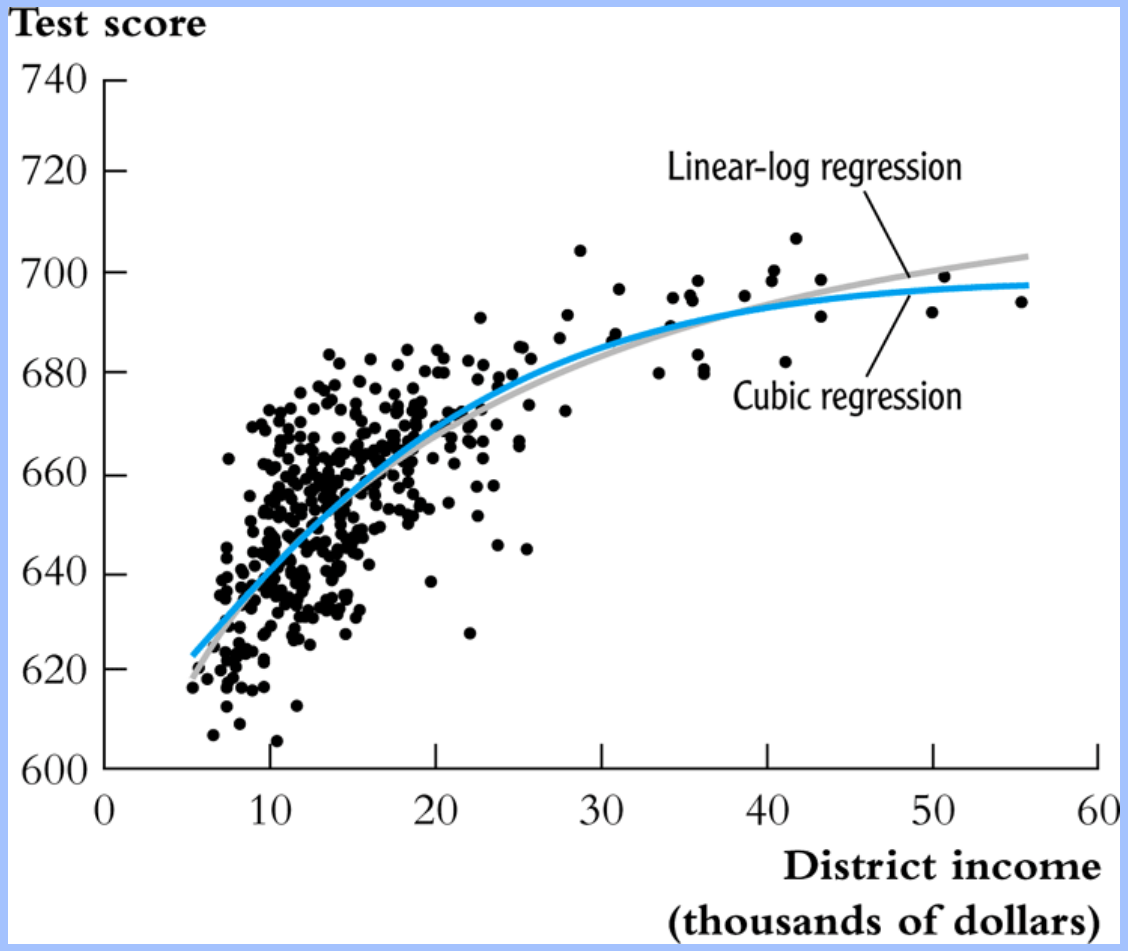
Example: TestScore vs. ln(Income)

- First defining 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?

The linear-log and cubic regression functions



II. Log-linear population regression function

$$\ln(Y) = \beta_0 + \beta_1 X \qquad \text{(b)}$$

Now change X: $\ln(Y + \Delta Y) = \beta_0 + \beta_1(X + \Delta X)$ (a)

Subtract (a) – (b): $\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$

so $$\frac{\Delta Y}{Y} \cong \beta_1 \Delta X$$

or $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X} \quad \text{(small } \Delta X\text{)}$$

Log-linear case, continued

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$$

for small ΔX, $\beta_1 \cong \dfrac{\Delta Y / Y}{\Delta X}$

- Now $100 \times \dfrac{\Delta Y}{Y}$ = percentage change in Y, so ***a change in X by one unit* ($\Delta X = 1$) *is associated with a* $100\beta_1$*% change in Y*.**
- 1 unit increase in $X \rightarrow$ β_1 increase in $\ln(Y)$
  $\rightarrow$ $100\beta_1$% increase in Y
- *Note*: What are the units of u_i and the SER?
  - fractional (proportional) deviations
  - for example, *SER* = .2 means…

III. Log-log population regression function

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i \qquad \text{(b)}$$

Now change X: $\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X)$ (a)

Subtract: $\ln(Y + \Delta Y) - \ln(Y) = \beta_1[\ln(X + \beta X) - \ln(X)]$

so $$\frac{\Delta Y}{Y} \cong \beta_1 \frac{\Delta X}{X}$$

or $$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X} \text{ (small } \Delta X\text{)}$$

Log-log case, continued

$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$

for small ΔX,

$$\beta_1 \cong \frac{\Delta Y / Y}{\Delta X / X}$$

Now $100 \times \frac{\Delta Y}{Y}$ = percentage change in Y, and $100 \times \frac{\Delta X}{X}$ = percentage change in X, so ***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 defining 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:

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

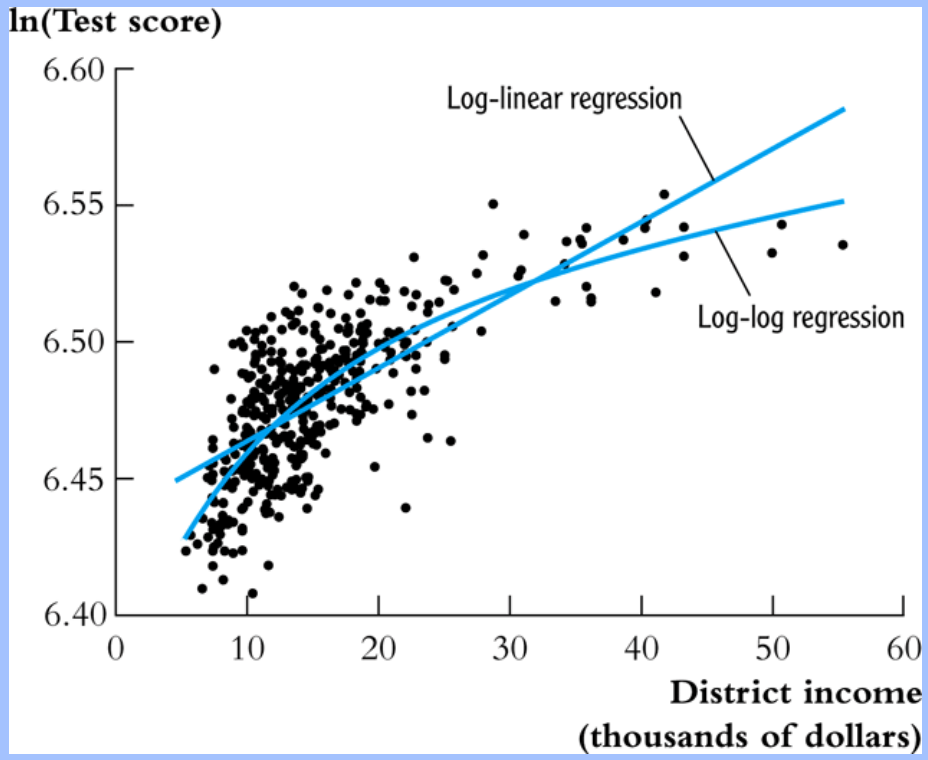
An 1% increase in *Income* is associated with an increase of .0554% in *TestScore* (*Income* up by a factor of 1.01, *TestScore* up by a factor of 1.000554)

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

$$\ln(\text{Test Score}) = 6.336 + 0.0554 \times \ln(Income_i)$$
$$(0.006) \quad (0.0021)$$

- 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)*

Poisson Regression

- Poisson regression is also a type of GLM model where the random component is specified by the Poisson distribution of the response variable which is a count.
- When all explanatory variables are discrete, log-linear model is equivalent to poisson regression model.

Poisson Regression - Applications

Give data about crabs we are interested in knowing

- How does the number of satellites, (male crabs residing near a female crab), for a female horseshoe crab depend on the width of her back?, and
- What is the rate of satellites per unit width?

Given data about credit cards we are interested in knowing

- What is the expected number of credit cards a person may have, given his/her income?, or
- What is the sample rate of possession of credit cards?

Poisson Regression

Variables:

- In Poisson regression **Response/outcome** variable Y is a count. But we can also have Y/t, the rate (or incidence) as the response variable, where t is an interval representing time, space or some other grouping.

Explanatory Variable(s):

- Explanatory variables, $X = (X_1, X_2, \ldots X_k)$, can be continuous or a combination of continuous and categorical variables. Convention is to call such a model "Poisson Regression".
- Explanatory variables, $X = (X_1, X_2, \ldots X_k)$, can be ALL categorical. Then the counts to be modeled are the counts in a contingency table, and the convention is to call such a model log-linear model.
- If Y/t is the variable of interest then even with all categorical predictors, the regression model will be known as Poisson regression, not a log-linear model.

Poisson Regression

GLM Model for Counts with its assumptions:

- $g(\mu)=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k=x^T_i\beta$
- *Random component:* Response *Y* has a Poisson distribution that is $y_i \sim \text{Poisson}(\mu_i)$ for i=1,...,N where the expected count of y_i is $E(Y)=\mu$.
- *Systematic component:* Any set of $X = (X_1, X_2, \ldots X_k)$ are explanatory variables. For now let's focus on a single variable *X*.
- For simplicity, with a single explanatory variable, we write: $\log(\mu)=\alpha+\beta x$. This is equivalent to: $\mu=\exp(\alpha+\beta x)=\exp(\alpha)\exp(\beta x)$

Poisson Regression

Interpretation of Parameter Estimates:

- exp(α) = effect on the mean of *Y*, that is μ, when $X = 0$
- exp(β) = with every unit increase in *X*, the predictor variable has multiplicative effect of exp(β) on the mean of *Y*, that is μ
  - If $\beta = 0$, then $\exp(\beta) = 1$, and the expected count, $\mu = E(y) = \exp(\alpha)$, and *Y* and *X* are not related.
  - If $\beta > 0$, then $\exp(\beta) > 1$, and the expected count $\mu = E(y)$ is exp(β) times larger than when $X = 0$
  - If $\beta < 0$, then $\exp(\beta) < 1$, and the expected count $\mu = E(y)$ is exp(β) times smaller than when $X = 0$

Poisson Regression

GLM Model for Rates:

- $g(\mu)=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_k x_k=x^T_i\beta$
- *Random component*: Response *Y* has a Poisson distribution, and *t* is index of the time or space; more specifically the expected value of rate *Y*/*t*, is $E(Y/t)=\mu/t$.
- *Systematic component:* Any set of $X = (X_1, X_2, \ldots X_k)$ can be explanatory variables. For now let's focus on a single variable *X*.

Poisson Regression

Log of rate: log(*Y*/*t*)

Poisson loglinear regression model for the expected rate of the occurrence of event is:

- $\log(\mu/t)=\alpha+\beta x$
- This can be rearranged to:
- $\log(\mu)-\log(t)=\alpha+\beta x$
  $\log(\mu)=\alpha+\beta x+\log(t)$
- The term $\log(t)$ is referred to as an **offset**. It is an adjustment term and a group of observations may have the same offset, or each individual may have a different value of *t*. *log(t)* is an *observation* and it will change the value of estimated counts:
- $\mu=\exp(\alpha+\beta x+\log(t))=(t)\exp(\alpha)\exp(\beta x)$
- This means that mean count is proportional to *t*.

Poisson Regression

Parameter Estimation

- Similar to the case of Logistic regression, the *maximum likelihood estimators* (MLEs) for (β_0, β_1 … etc.) are obtained by finding the values that maximizes log-likelihood. In general, there are no closed-form solutions, so the ML estimates are obtained by using iterative algorithms such as *Newton-Raphson* (NR)

Inference

- The usual tools from the basic statistical inference and GLM are valid.

Poisson Regression - Applications

Data: Each female crab had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites. Explanatory variables: female crab's color (*C*), spine condition (*S*), weight (*Wt*), and carapace width (*W*). The response outcome for each female crab is her number of satellites (*Sa*).

- How does the number of satellites, (male crabs residing near a female crab), for a female horseshoe crab depend on the weight of her back?
- The estimated model is: $\log(\mu i) = -3.30476 + 0.16405 Wt_i$
- **Interpretation**: Since estimate of $\beta > 0$, the heavier the female crab, the greater expected number of male satellites on the multiplicative order as *exp*(0.1640) = 1.18. More specifically, for one unit of increase in the width, the number of *Sa* will increase and it will be multiplied by 1.18.

Poisson Regression - Applications

Data on the relation between income and whether one possesses a credit card. At each level of annual income in millions of lira the data indicates the number of subjects sampled and the number of these subjects possessing at least one credit card.

- What is the expected number of credit cards a person may have, given his/her income?, or
- What is the sample rate of possession of credit cards?
- In the group of six people that earn 65 million lira, the expected number is
- $\log(\mu/t) = -2.3866 + 0.0208 \times \text{Income} = -2.3866 + 0.0208 \times 65$
- $\log(\mu) = -2.3866 + 0.0208 \times 65 + \log(t)$
- $\log(\mu) = -2.3866 + 0.0208 \times 65 + 1.79176$
- $\mu = 2.12641$
- **Question**: How many people would we expect to have at least one travel credit card in a group of 10 people who earn about 120 million lira?

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 $X \rightarrow \infty$ is β_0.

Negative exponential growth

We want to estimate the parameters of,

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

or

$$Y_i = \beta_0 \left[1 - e^{-\beta_1 (X_i - \beta_2)} \right] + u_i \qquad (*)$$

where $\alpha = \beta_0 e^{\beta_2}$ (why would you do this???)

Compare model (*) to linear-log or cubic models:

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

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

The linear-log and polynomial models are *linear in the parameters* β_0 and β_1 – but the 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)
- The NLS problem for the proposed specification:

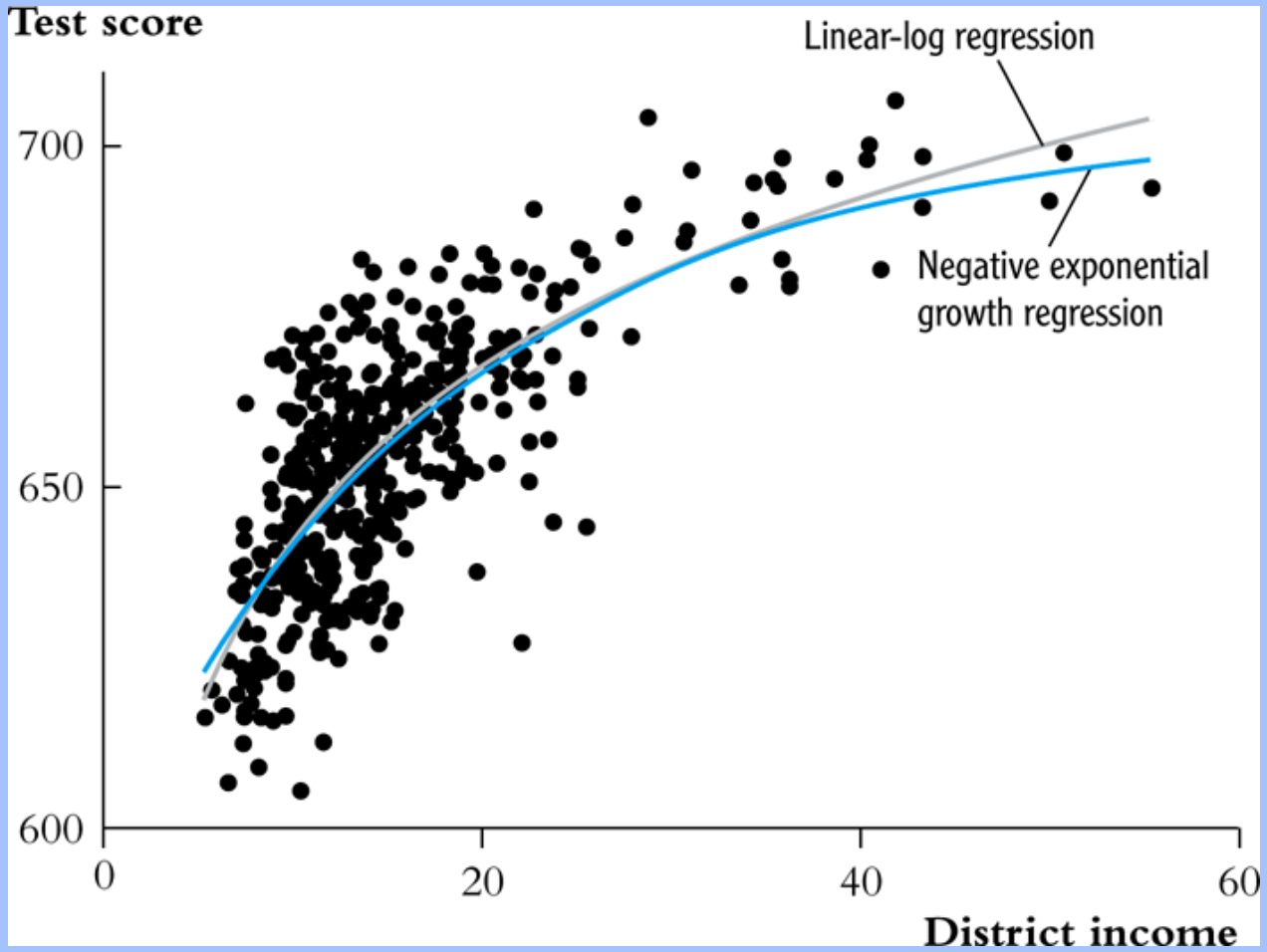
$$\min_{\beta_0,\beta_1,\beta_2} \sum_{i=1}^{n} \left\{ Y_i - \beta_0 \left[1 - e^{-\beta_1 (X_i - \beta_2)} \right] \right\}^2$$

This is a nonlinear minimization problem (a "hill-climbing" problem). How could you solve this?

- Guess and check
- There are better ways…

```
. nl (testscr = {b0=720}*(1 - exp(-1*{b1}*(avginc-{b2})))), r

(obs = 420)
Iteration 0:  residual SS =  1.80e+08                        .
Iteration 1:  residual SS =  3.84e+07                        .
Iteration 2:  residual SS =   4637400                        .
Iteration 3:  residual SS =  300290.9          STATA is "climbing the hill"
Iteration 4:  residual SS =  70672.13          (actually, minimizing the SSR)
Iteration 5:  residual SS =  66990.31                        .
Iteration 6:  residual SS =   66988.4                        .
Iteration 7:  residual SS =   66988.4                        .
Iteration 8:  residual SS =   66988.4

Nonlinear regression with robust standard errors      Number of obs =       420
                                                      F(  3,   417) = 687015.55
                                                      Prob > F      =    0.0000
                                                      R-squared     =    0.9996
                                                      Root MSE      =  12.67453
                                                      Res. dev.     =  3322.157
------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          b0 |   703.2222   4.438003   158.45   0.000     694.4986    711.9459
          b1 |   .0552339   .0068214     8.10   0.000     .0418253    .0686425
          b2 |  -34.00364    4.47778    -7.59   0.000    -42.80547    -25.2018
------------------------------------------------------------------------------
 (SEs, P values, CIs, and correlations are asymptotic approximations)
```

Negative exponential growth; *RMSE* = 12.675
Linear-log; *RMSE* = 12.618 (oh well…)



Interactions Between Independent Variables

- 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

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

General rule: compare the various cases

$$E(Y_i | D_{1i}=0, D_{2i}=d_2) = \beta_0 + \beta_2 d_2 \qquad \text{(b)}$$

$$E(Y_i | D_{1i}=1, D_{2i}=d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \qquad \text{(a)}$$

subtract (a) – (b):

$$E(Y_i | D_{1i}=1, D_{2i}=d_2) - E(Y_i | D_{1i}=0, D_{2i}=d_2) = \beta_1 + \beta_3 d_2$$

- 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

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

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

- "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

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

Let

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

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

- Can you relate these coefficients to the following table of group ("cell") means?

	Low STR	*High STR*
Low EL	664.1	662.2
High EL	645.9	640.5

(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

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

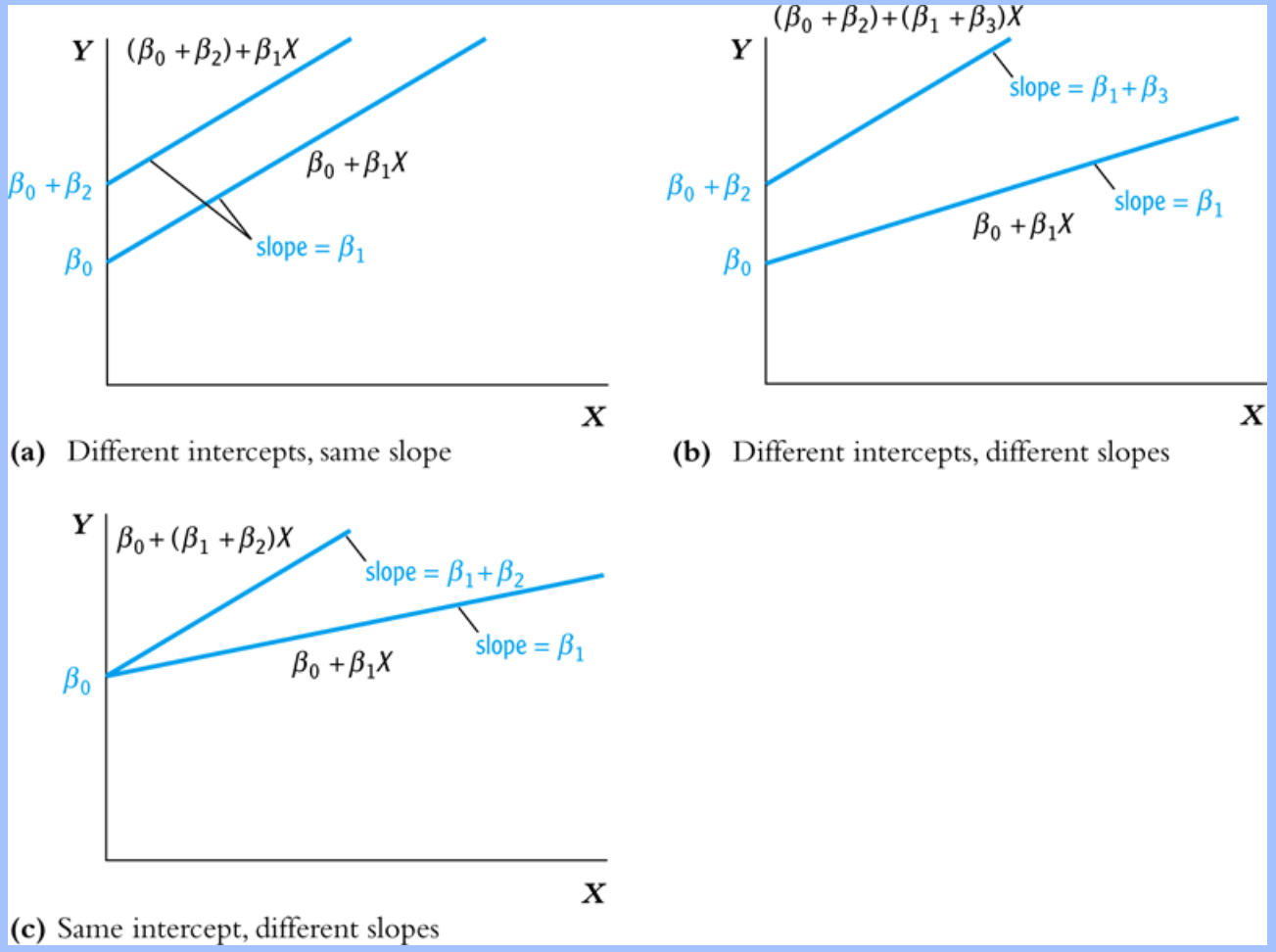
Observations with $D_i = 0$ (the "$D = 0$" group):

$$Y_i = \beta_0 + \beta_2 X_i + u_i$$ ***The D=0 regression line***

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

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

Binary-continuous interactions, ctd.



Interpreting the coefficients

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

General rule: compare the various cases

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) \quad \text{(b)}$$

Now change X:

$$Y + \Delta Y = \beta_0 + \beta_1 D + \beta_2 (X + \Delta X) + \beta_3 [D \times (X + \Delta X)] \quad \text{(a)}$$

subtract (a) – (b):

$$\Delta Y = \beta_2 \Delta X + \beta_3 D \Delta X \quad \text{or} \quad \frac{\Delta Y}{\Delta X} = \beta_2 + \beta_3 D$$

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

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

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

- When *HiEL* = 0:

  $\widehat{TestScore} = 682.2 - 0.97STR$
- When *HiEL* = 1,

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

  $\widehat{TestScore} = 687.8 - 2.25STR$
- 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

$$\underset{}{\textit{Test Score}} = 682.2 - 0.97\textit{STR} + 5.6\textit{HiEL} - 1.28(\textit{STR} \times \textit{HiEL})$$
$$\qquad\qquad (11.9) \quad (0.59) \qquad (19.5) \qquad (0.97)$$

- The two regression lines have the same **slope** ← → the coefficient on *STR* × *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* × *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:

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

General rule: compare the various cases

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) \qquad \text{(b)}$$

Now change X_1:

$$Y + \Delta Y = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2 + \beta_3 [(X_1 + \Delta X_1) \times X_2] \qquad \text{(a)}$$

subtract (a) – (b):

$$\Delta Y = \beta_1 \Delta X_1 + \beta_3 X_2 \Delta X_1 \quad \text{or} \quad \frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$$

- 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*

$$\text{Test Score} = 686.3 - 1.12STR - 0.67PctEL + .0012(STR \times PctEL),$$
$$(11.8)\ \ (0.59)\qquad (0.37)\qquad (0.019)$$

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%	$-1.12+.0012 \times 20 = -1.10$

Example, ctd: hypothesis tests

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

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

$t = .0012/.019 = .06$ → can't reject null at 5% level

- Does population coefficient on $STR = 0$?

$t = -1.12/0.59 = -1.90$ → can't reject null at 5% level

- Do the coefficients on ***both*** *STR* ***and*** $STR \times PctEL = 0$?

$F = 3.89$ (p-value = .021) → reject null at 5% level(!!) (Why? high but imperfect multicollinearity)

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.

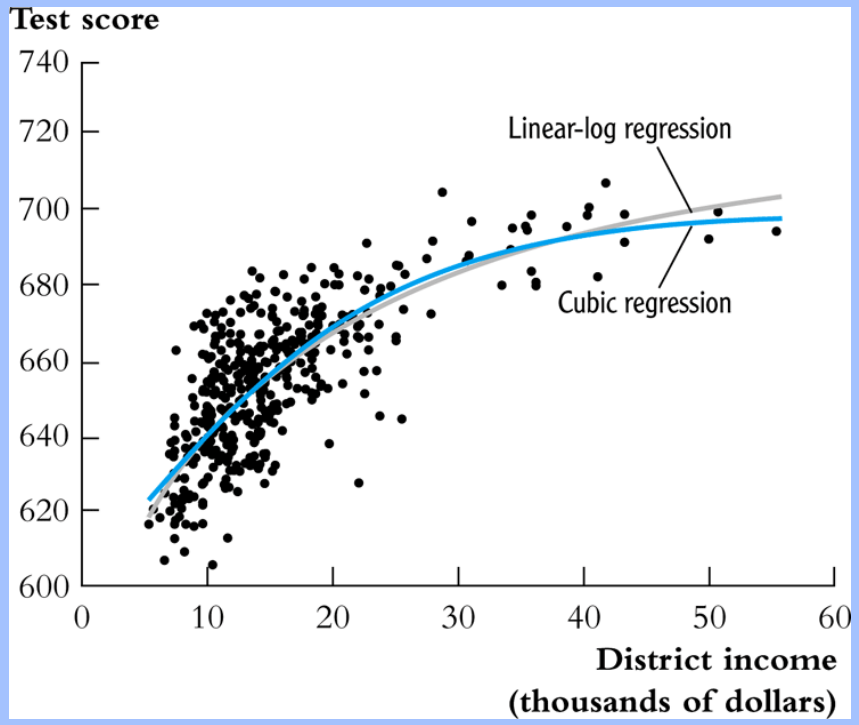


TABLE 8.3 Nonlinear Regression Models of Test Scores

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

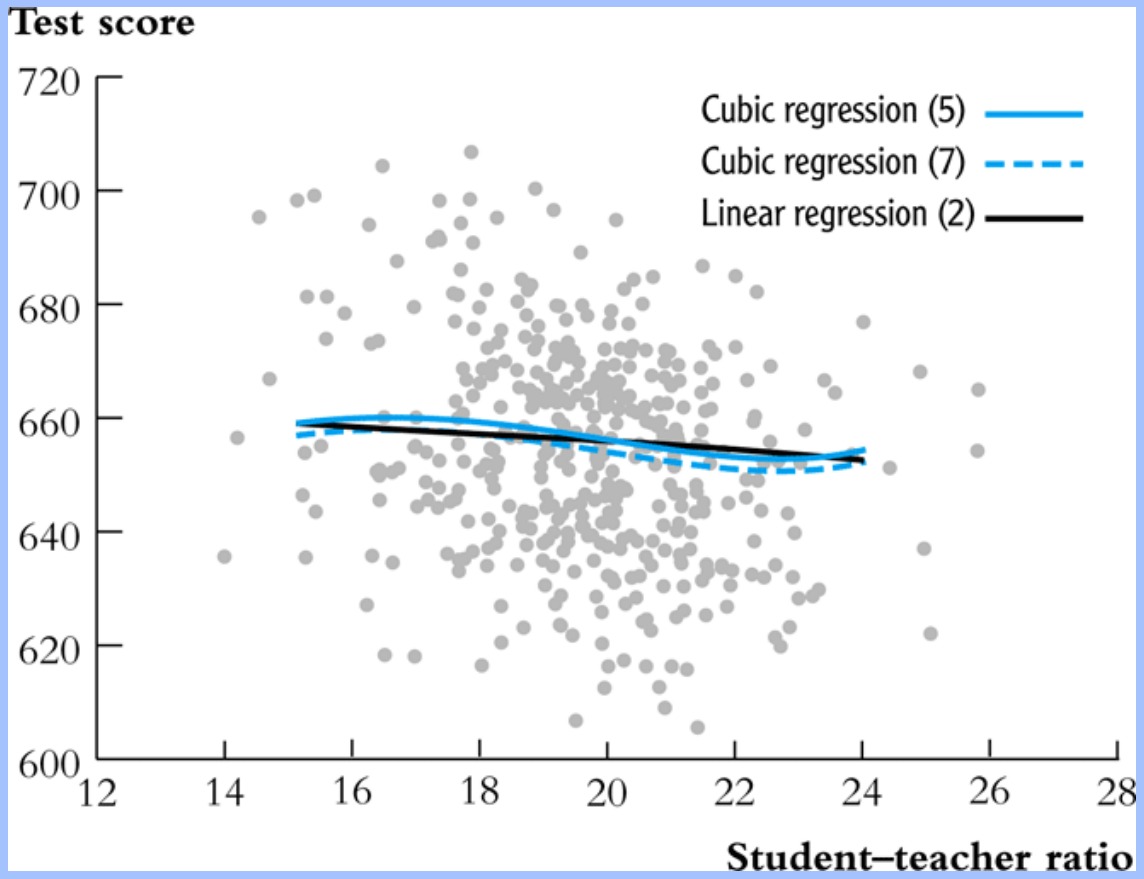
These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Standard errors are given in parentheses under coefficients, and *p*-values are given in parentheses under *F*-statistics. Individual coefficients are statistically significant at the *5%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

What can you conclude about question #1?

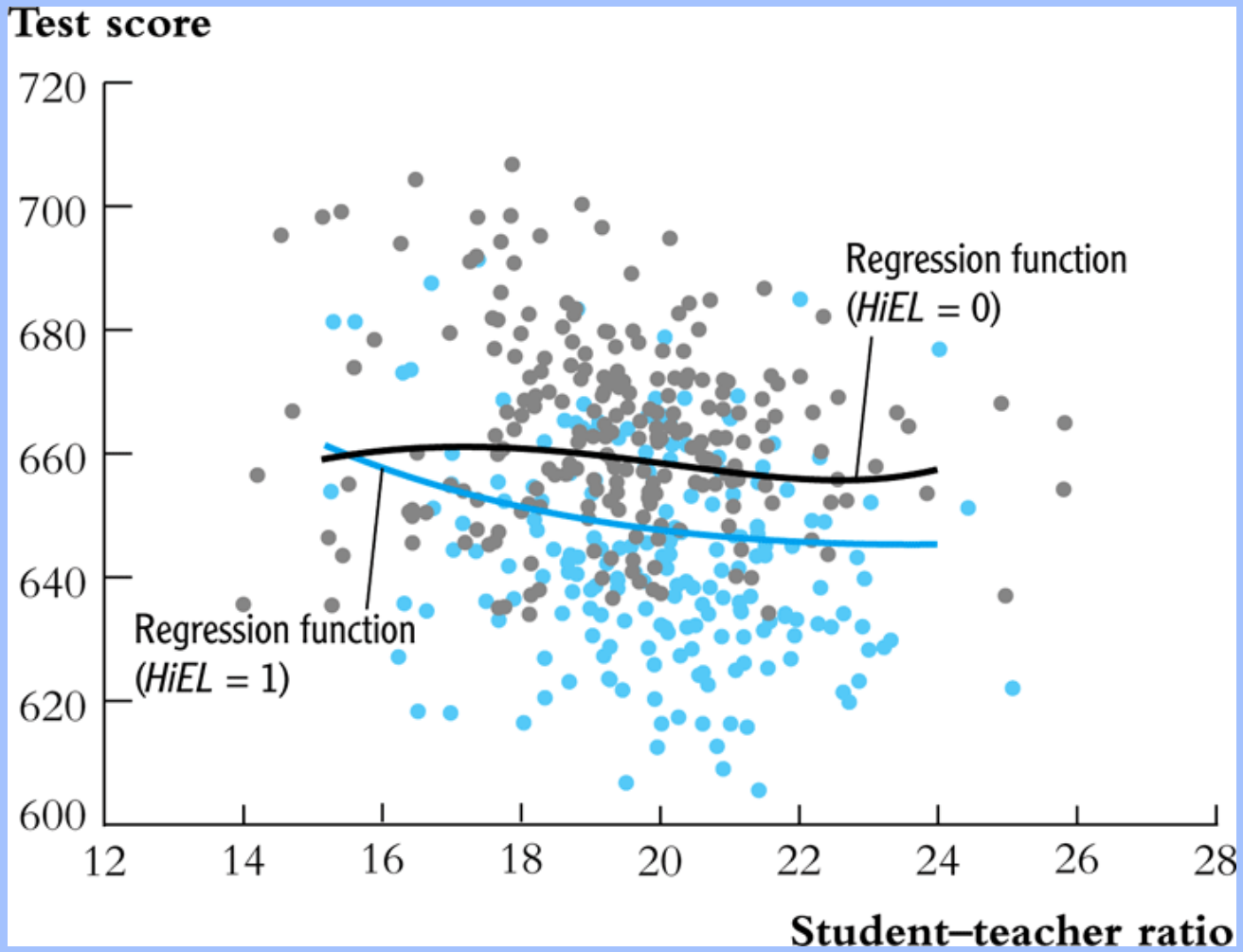
About question #2?

Interpreting the regression functions via plots:

First, compare the linear and nonlinear specifications:



Next, compare the regressions with interactions:



Summary: Nonlinear Regression Functions

- Using functions of the independent variables such as ln(X) or $X_1 \times X_2$, allows recasting a large family of nonlinear regression functions as multiple regression.
- Estimation and inference proceed 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?