Lecture Notes #6
(Chapter 8)
Economics 120B
Econometrics

Prof. Dahl
UC San Diego

Nonlinear Regression Functions

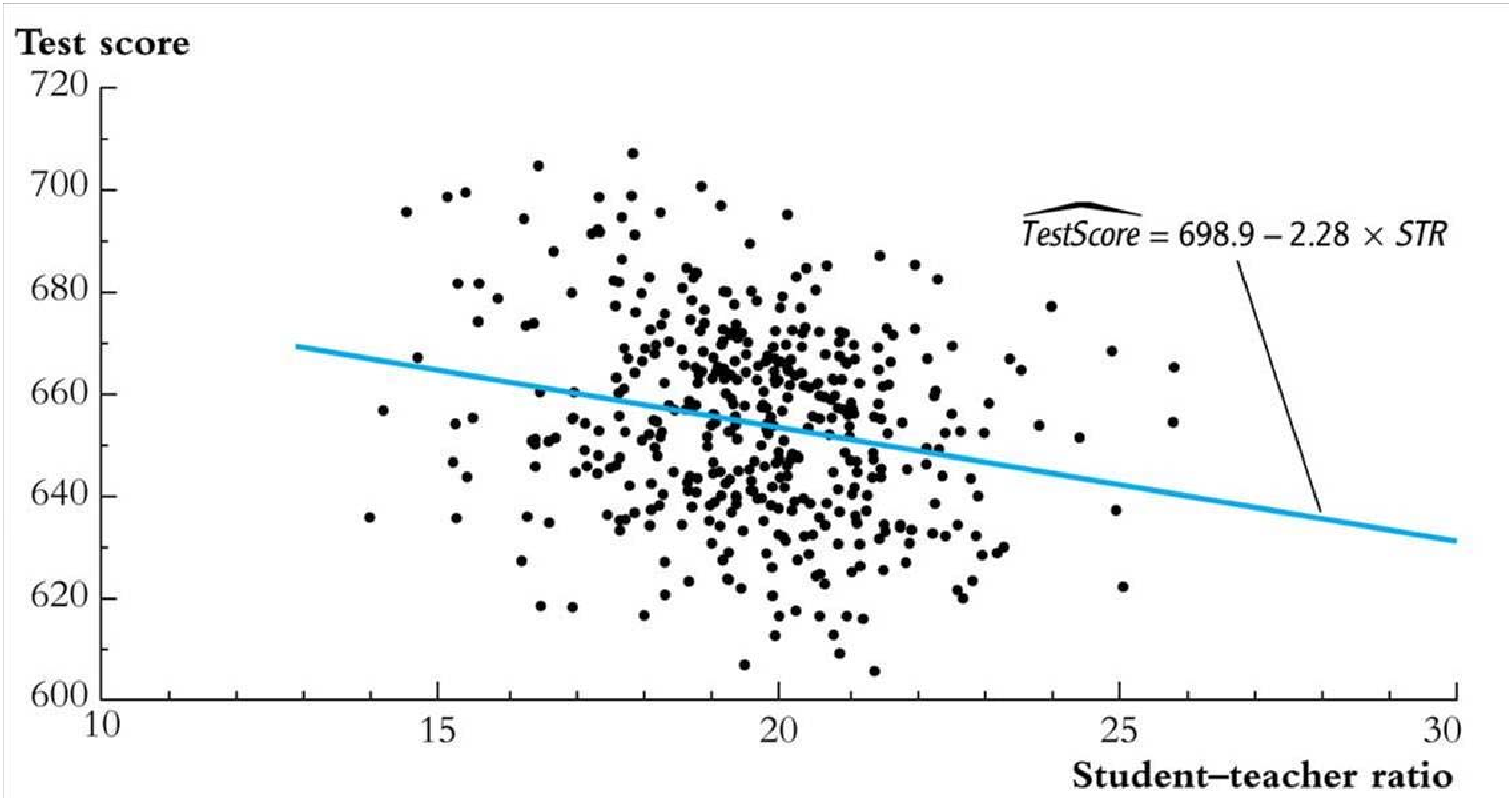
(SW Chapter 8)

- Everything so far has been linear in the *X*'s
- But the linear approximation is not always a good one
- The multiple regression framework can be extended to handle regression functions that are nonlinear in one or more *X*.

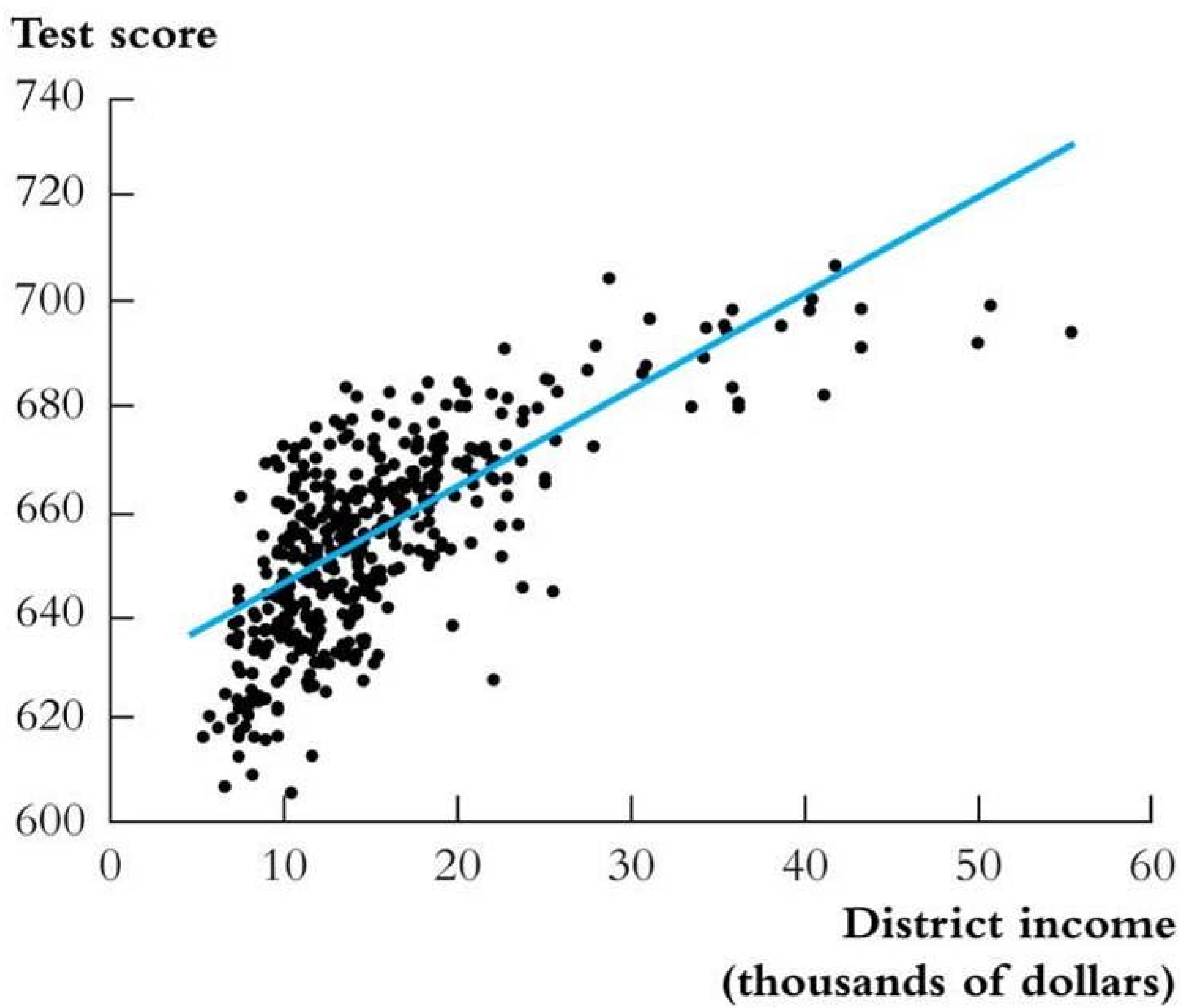
<u>Outline</u>

1. Nonlinear regression functions – general comments
2. Nonlinear functions of one variable
3. Nonlinear functions of two variables: interactions

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



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



Nonlinear Regression Population Regression Functions – General Ideas (SW Section 8.1)

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 – it needn't even be right on average.
- The solution to this is to estimate a regression function that is nonlinear in X

Nonlinear Functions of a Single Independent Variable (SW Section 8.2)

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:

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

- 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

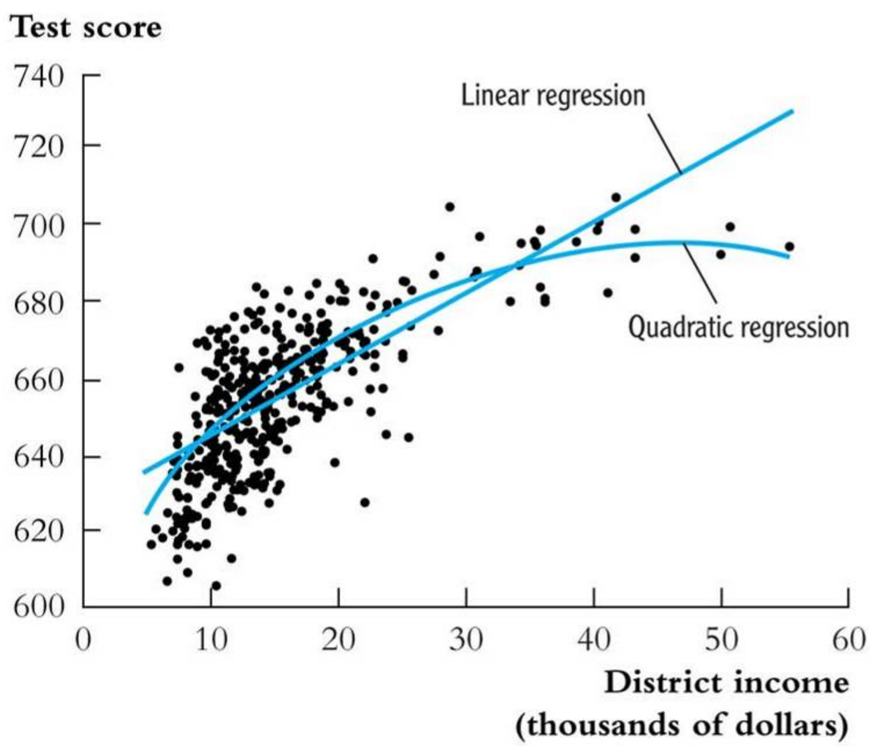
```
generate avginc2 = avginc*avginc;        Create a new regressor
reg testscr avginc avginc2, r;

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

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



Interpreting the estimated regression function, ctd:

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

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

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

$$\begin{aligned}\Delta\widehat{TestScore} &= 607.3 + 3.85 * 6 - 0.0423 * 6^2 \\ &\quad -(607.3 + 3.85 * 5 - 0.0423 * 5^2) \\ &= 3.4\end{aligned}$$

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: 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.

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*)

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

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

Numerically:

$$\ln(1.01) = .00995 \cong .01;$$

$$\ln(1.10) = .0953 \cong .10 \text{ (sort of)}$$

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*."

I. Linear-log case

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

for small ΔX,

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

Now $100 \times \dfrac{\Delta X}{X}$ = 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 $X \Rightarrow$ .01 increase in $\ln(X)$

$\Rightarrow .01\beta_1$ increase in Y)

Example: TestScore vs. ln(Income)

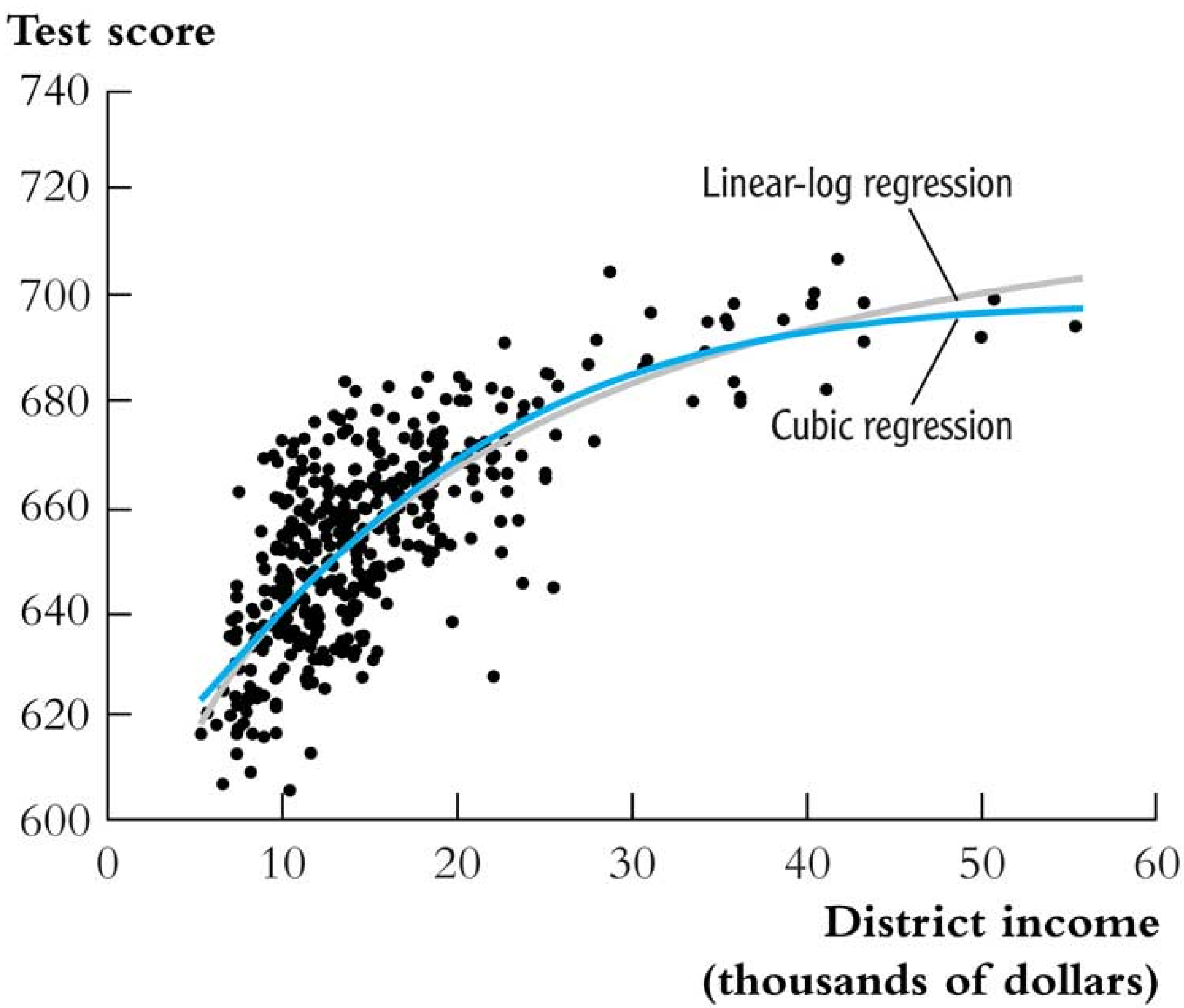
- First defining the new regressor, ln(*Income*)
- The model is now linear in ln(*Income*), so the linear-log model can be estimated by OLS:

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

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?

The linear-log and cubic regression functions



II. Log-linear case

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

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

- Now $100 \times \frac{\Delta Y}{Y}$ = percentage change in Y, so ***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 case

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

for small ΔX,

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

Now $100 \times \dfrac{\Delta Y}{Y}$ = percentage change in Y, and $100 \times \dfrac{\Delta X}{X}$ = 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***.

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.
- Choice of specification should be guided by judgment (which interpretation makes the most sense in your application?), tests, and plotting predicted values

Interactions Between Independent Variables (SW Section 8.3)

- 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, $\dfrac{\Delta TestScore}{\Delta STR}$ might depend on $PctEL$
- More generally, $\dfrac{\Delta Y}{\Delta X_1}$ 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

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

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

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

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

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

- "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

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

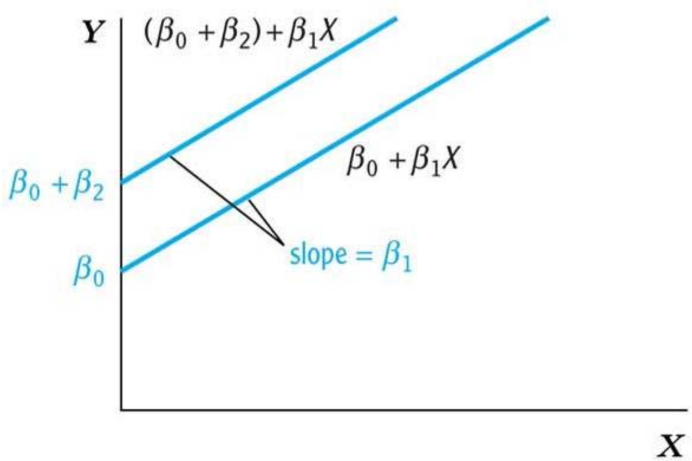
Observations with D_i= 0 (the "D = 0" group):

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

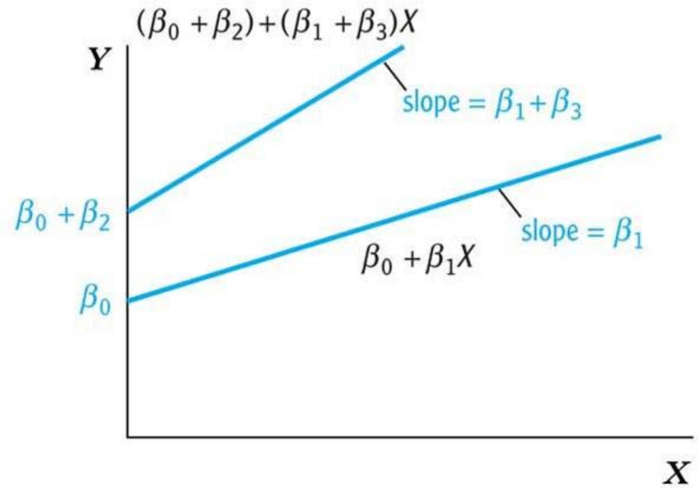
Observations with D_i= 1 (the "D = 1" group):

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

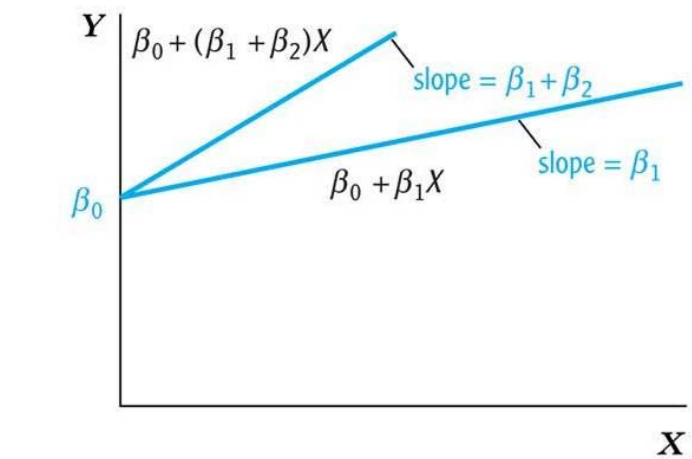
Binary-continuous interactions, ctd.



(a) Different intercepts, same slope



(b) Different intercepts, different slopes



(c) Same intercept, different slopes

Interpreting the coefficients

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

General rule: compare the various cases

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

Now change X:

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

subtract (a) – (b):

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

- 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, = 0 otherwise)

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

- When $HiEL = 0$:

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

- When $HiEL = 1$,

$$\begin{aligned}\widehat{TestScore} &= 682.2 - 0.97 * STR + 5.6 - 1.28 * STR \\ &= 687.8 - 2.25 * STR\end{aligned}$$

- 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

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

- 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:

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

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):

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

- 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

Application: Nonlinear Effects on Test Scores of the Student-Teacher Ratio (SW Section 8.4)

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 y 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 $HiEL \times STR^3$.

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?

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?