

6 – NONLINEAR REGRESSION FUNCTIONS



University of
Massachusetts
Amherst BE REVOLUTIONARY™

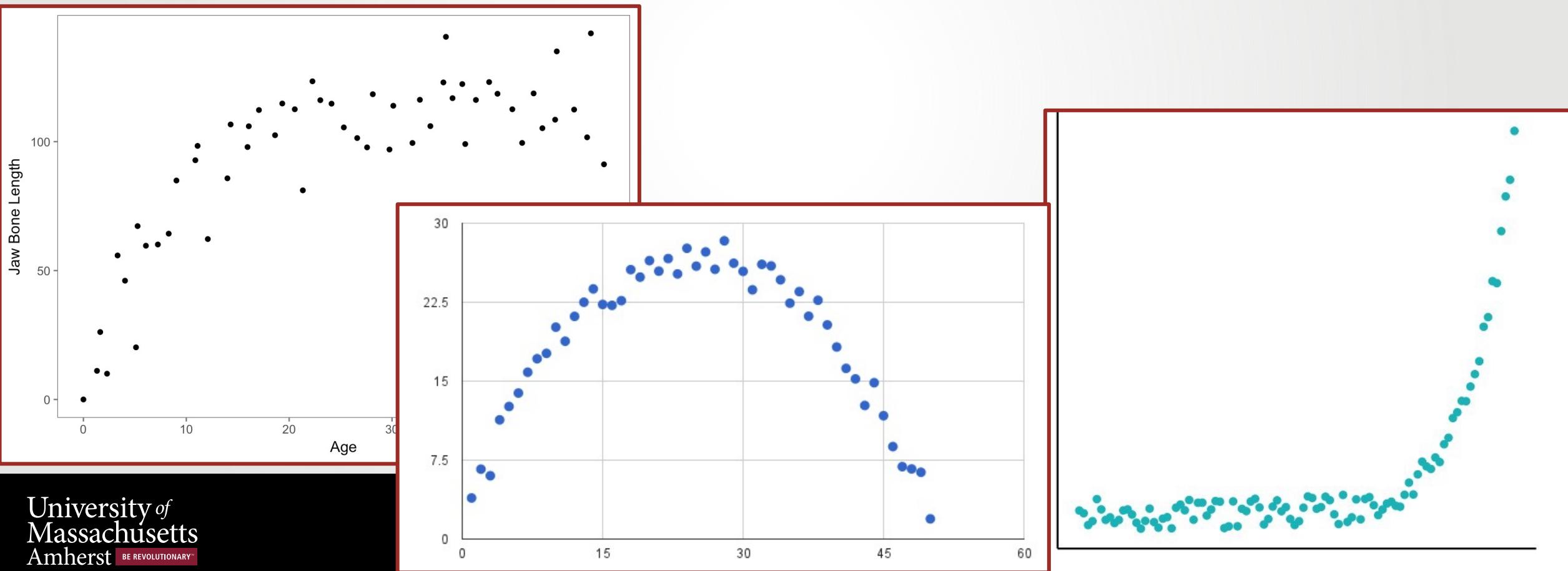
SECTION 6 – NONLINEAR REGRESSION FUNCTIONS

THE PLAN

1. Nonlinear functions of a single independent variable.
2. Polynomial regression functions.
3. Logarithmic regression functions.
4. Interactions between regressors.

OVERVIEW

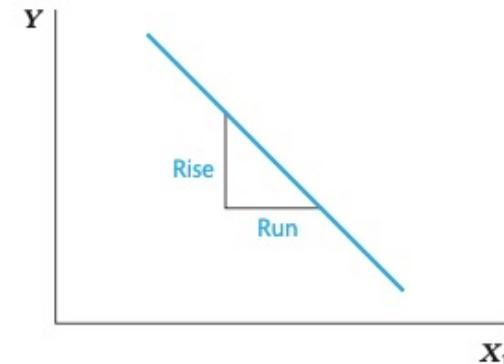
We considered *linear* regression functions so far (constant slope)...
...but what if your data looks like one of these?



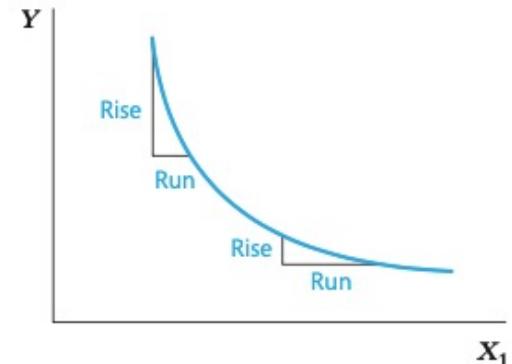
Two main ways of being nonlinear:

1. The “effect” of one regressor is nonlinear.
2. Interaction: the “effect” of a regressor depends on the value taken by the other.

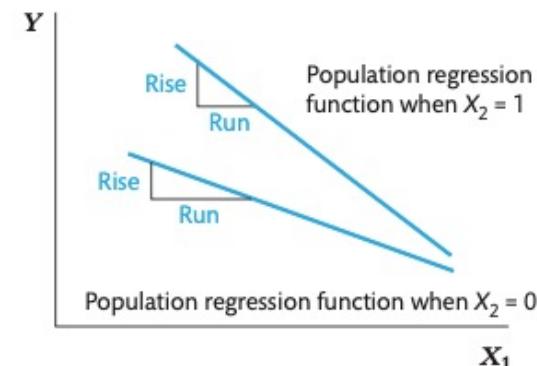
FIGURE 8.1 Population Regression Functions with Different Slopes



(a) Constant slope



(b) Slope depends on the value of X_1



(c) Slope depends on the value of X_2

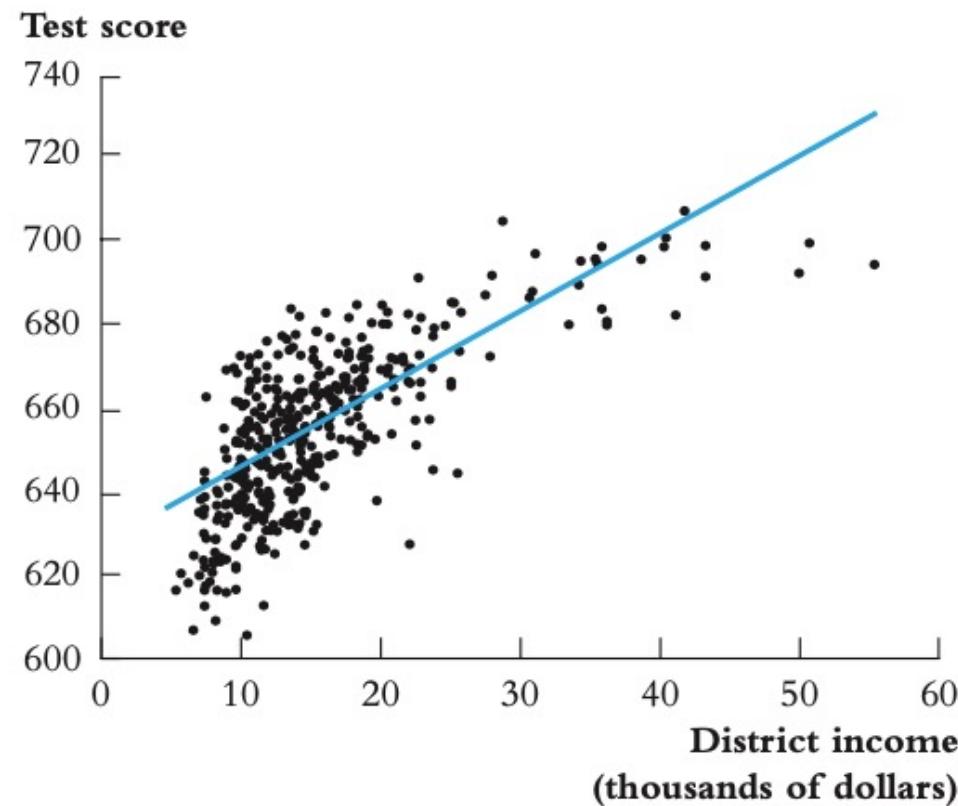
In Figure 8.1(a), the population regression function has a constant slope. In Figure 8.1(b), the slope of the population regression function depends on the value of X_1 . In Figure 8.1(c), the slope of the population regression function depends on the value of X_2 .

6.1 NONLINEAR FUNCTIONS OF A SINGLE INDEPENDENT VARIABLE

TEST SCORES & DISTRICT INCOME

FIGURE 8.2 Scatterplot of Test Scores vs. District Income with a Linear OLS Regression Function

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



TEST SCORES & DISTRICT INCOME

- Nonlinear function: the slope is not constant.

- Quadratic regression function:

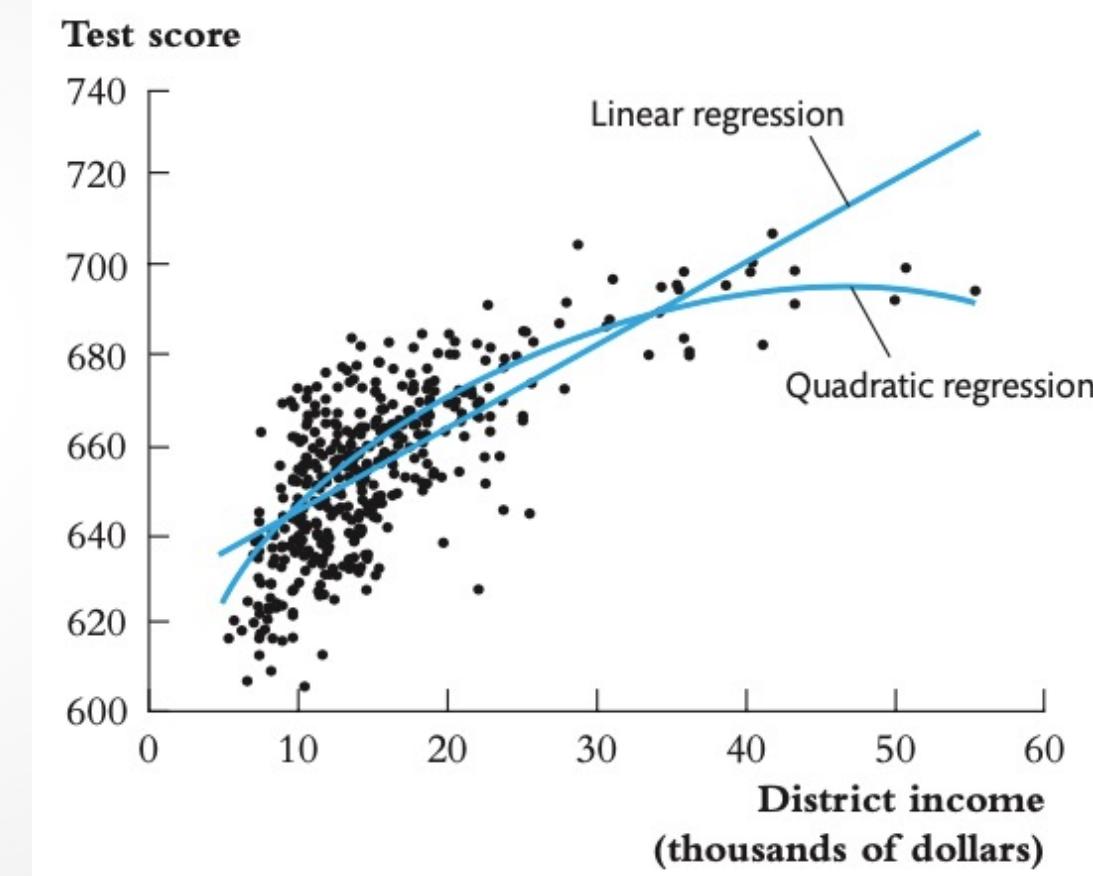
$$TestScore = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$$

- Technically, just a multiple regression with $Income_i^2$ as an additional regressor.

- OLS estimate:

$$\widehat{TestScores}_i = 607.4 + 3.85 Income - 0.04 Income^2$$

(2.9) (0.27) (0.005)



NONLINEAR FUNCTIONS

$$E(Y_i) = f(X_{1i}, X_{2i}, \dots, X_{ki})$$



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

- Effect of a change in X_1 by ΔX_1 units:

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

- If X_1 is continuous and ΔX_1 small, can use the partial derivative:

$$E(\Delta Y) \approx \Delta X_1 \frac{df(X_1, X_2, \dots, X_k)}{dX_1}$$

TEST SCORES & DISTRICT INCOME

$$TestScore = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$$

$$\hat{\beta}_0 = 607.4; \quad \hat{\beta}_1 = 3.85; \quad \hat{\beta}_2 = -0.0423$$

- Effect of district income going from 10 to 11 (thousand dollars):

$$\begin{aligned}\Delta Y &= (\beta_0 + \beta_1(11) + \beta_2(11)^2) - (\beta_0 + \beta_1(10) + \beta_2(10)^2) = \\ &= (\beta_1(11 - 10) + \beta_2(11^2 - 10^2)) = \\ &= (\beta_1 + 21\beta_2) = 3.85 - 0.89 = 2.96\end{aligned}$$

TEST SCORES & DISTRICT INCOME

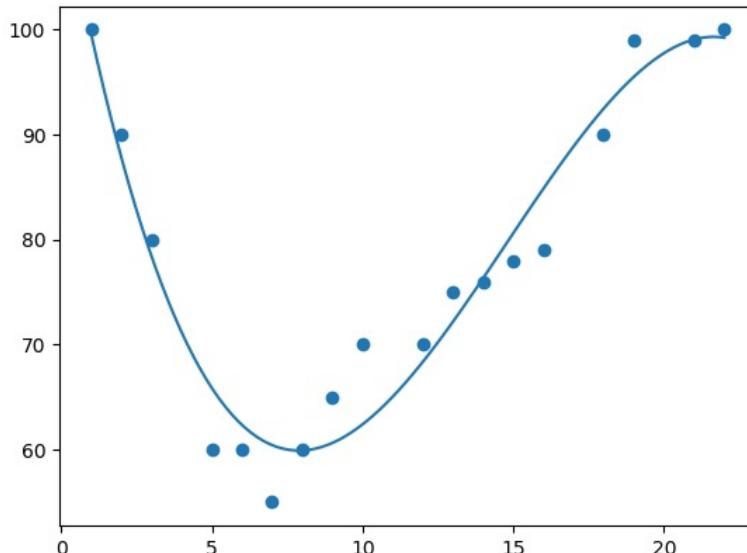
- $TestScore = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$
- Effect of district income going from 10 to 11 (thousand dollars):

$$\Delta \hat{Y} = (\widehat{\beta}_1 + 21\widehat{\beta}_2) = 3.85 - 0.89 = 2.96$$

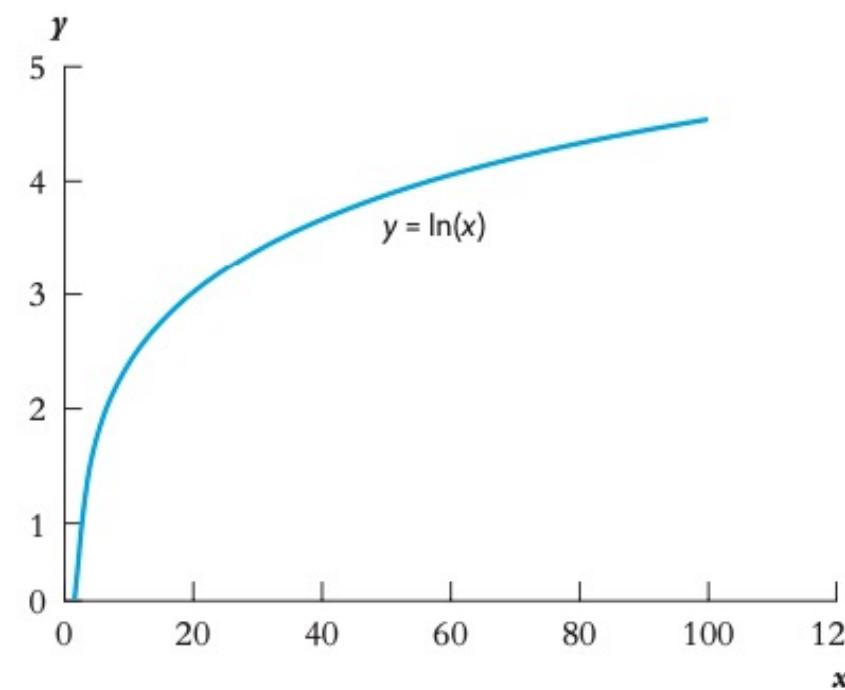
- $SE(\Delta \hat{Y}) = SE(\beta_1 + 21\beta_2)$
- STATA will compute it for you (if you know how to ask!)
- → *SE for predicted effect can be computed through a test of a single restriction involving multiple coefficients.*

NONLINEAR FUNCTIONS OF A SINGLE INDEPENDENT VARIABLE

1. Polynomial regression model.
2. Logarithmic regression model(s).



ometrics (



niele Girardi

6.2 POLYNOMIAL REGRESSION FUNCTIONS

POLYNOMIALS

- Polynomial regression model of degree r :

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

- $r = 2 \rightarrow$ quadratic regression model
- $r = 3 \rightarrow$ cubic regression model
- Estimation & inference: just like a multivariate regression.

POLYNOMIALS

- Polynomial regression model of degree r:

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

- Testing the null hypothesis of linearity:
 - $H_0: \beta_2, \dots, \beta_r = 0$
 - $H_1: \text{at least one of these coefficients is not } 0$
- How to choose the degree r?
 - *Theory; Plot the data; F- & t-tests.*

POLYNOMIALS: INCOME & TEST SCORES

- Quadratic specification:

$$TestScore = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i$$

- Cubic specification:

$$TestScore = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + \beta_3 Income_i^3 + u_i$$

QUADRATIC SPECIFICATION IN STATA

```
generate avginc2 = avginc*avginc  
reg testscr avginc avginc2, r
```

Create a new regressor

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]
<hr/>						
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

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

testscr	Robust					
	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 OF LINEARITY AGAINST A CUBIC ALTERNATIVE

test avginc2 avginc3 **Use 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 up to degree 3.

6.3 LOGARITHMIC REGRESSION FUNCTIONS

LOGARITHMIC REGRESSION FUNCTIONS

- Y and /or X transformed in its *natural logarithm*.

- **Three types:**

1. Linear-log model

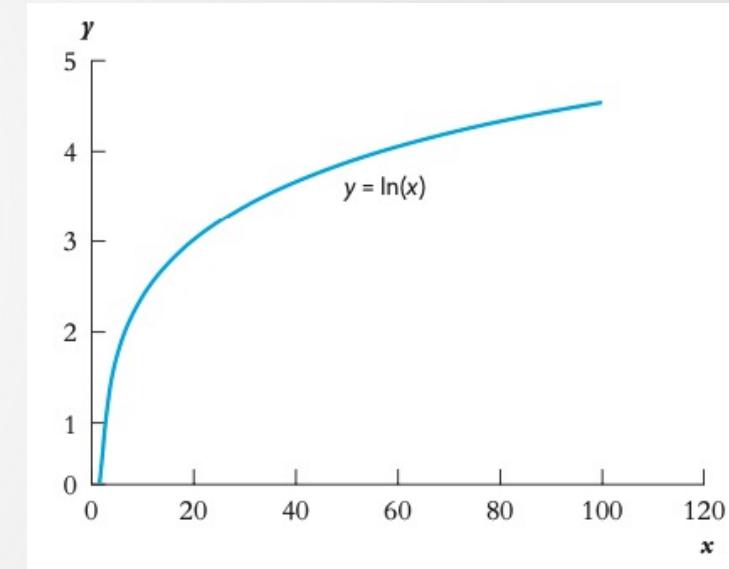
$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$$

2. Log-linear model

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

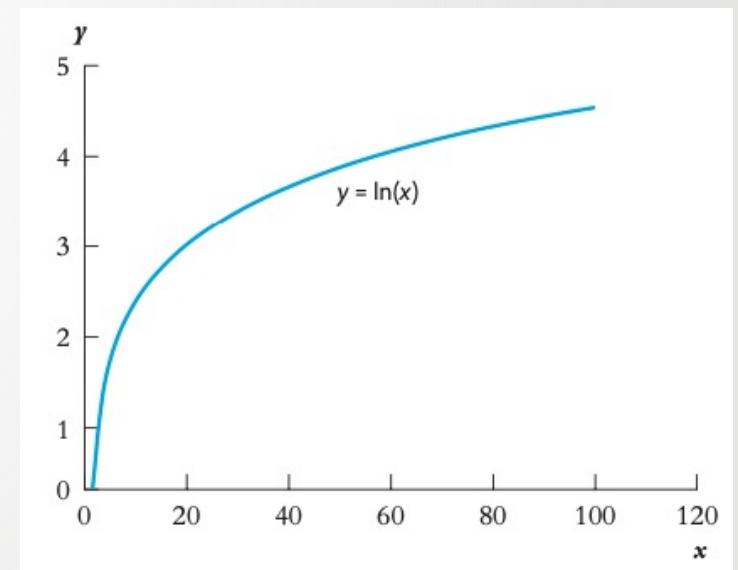
3. Log-log model

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



LOGS & PERCENTAGES

- Log changes are approximately equal to percentage changes.
- $\ln(x_2) - \ln(x_1) \approx \frac{x_2 - x_1}{x_1}$
- The approximation gets worse as $x_2 - x_1$ gets larger.



1) LINEAR-LOG MODEL

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

- Constant effect of a % change in X.
- 1% increase in X → $0.01 \times \beta_1$ change in E(Y).
- $\Delta \hat{Y} = \beta_1 [\ln(X + \Delta X) - \ln(X)] \approx \beta_1 \frac{\Delta X}{X}$
- Test score example:

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

- A 1% increase in income increases test scores by 0.36 points.

2) LOG-LINEAR MODEL

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

- The % change in $E(Y)$ caused by a unit change in X is constant.
- **unit change in $X \rightarrow (100 \times \beta_1)\%$ change in $E(Y)$.**
- Age-earnings example:

$$\ln(\widehat{Earnings}) = 2.876 + 0.0095 \text{ Age}$$

- *Earnings increase by 0.95% for each additional year of age*

3) LOG-LOG MODEL

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

- Constant *elasticity* of Y with respect to X.
- 1% change in X $\rightarrow \beta_1$ % change in E(Y).
- Test score example:

$$\ln(\widehat{\text{TestScore}}) = 6.4 + 0.06 \ln(\text{Income})$$

- A 1% increase in income increases test scores by 0.06%.

HOW TO CHOOSE A LOG SPECIFICATION?

- **Warning:** Can't compare R^2 if the dependent variable is different
 - $\ln(Y)$ vs. Y
- Does it make sense to think in terms of % changes?
 - *Usually it does for income or wages.*
- Does it make results easier to interpret?
 - *Percentage changes don't depend on the unit of measure.*

6.4 INTERACTIONS BETWEEN REGRESSORS

INTERACTION TERMS

- What if the effect of X_j on Y is different in different circumstances?
- Example with binary regressors:

$$\ln(Earnings_i) = \beta_0 + \beta_1 College_i + \beta_2 Female_i + u_i$$

- Effect of College assumed to be the same for men and women.
- Add an *interaction term*:

$$\begin{aligned}\ln(Earnings_i) \\ = \beta_0 + \beta_1 College_i + \beta_2 Female_i + \beta_3 College_i \times Female_i + u_i\end{aligned}$$

INTERACTION TERMS

$$\ln(Earnings_i) = \beta_0 + \beta_1 College_i + \beta_2 Female_i + \beta_3 College_i \times Female_i + u_i$$

- $E(Earnings | College = 0, Female = 0) = \beta_0$
- $E(Earnings | College = 1, Female = 0) = \beta_0 + \beta_1$
- $E(Earnings | College = 0, Female = 1) = \beta_0 + \beta_2$
- $E(Earnings | College = 1, Female = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$

Effect of college for a man: β_1

Effect of college for a woman: $\beta_1 + \beta_3$

INTERACTION BETWEEN A CONTINUOUS AND A BINARY VARIABLE

- Regression with 1 continuous & 1 binary variable:

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

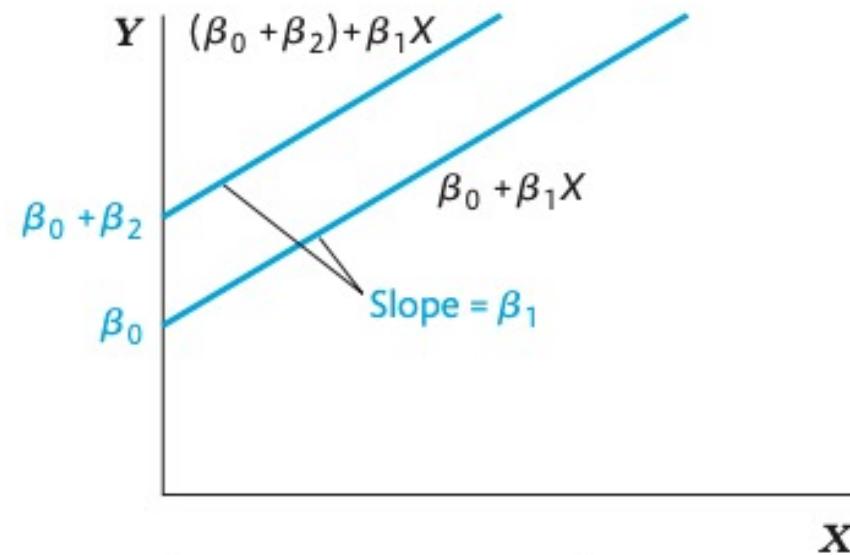
- With interaction term:

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

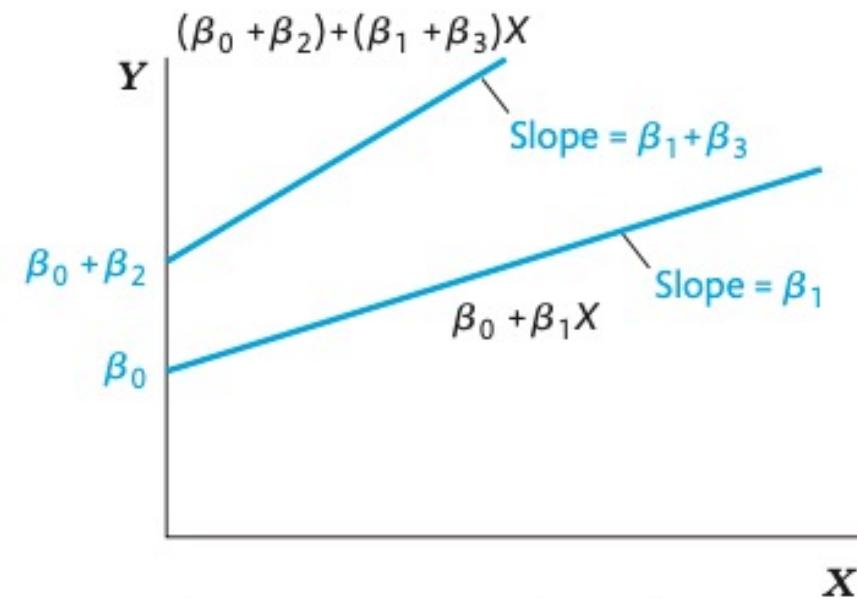
- Slope if D=0: β_1
- Slope if D=1: $\beta_1 + \beta_3$

FIGURE 8.8

Regression Functions Using Binary and Continuous Variables



(a) Different intercepts, same slope



(b) Different intercepts, different slopes

INTERACTION: TWO CONTINOUS VARIABLES

- Regression with two continuous variables:

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

- With interaction:

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

- $\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2$
- $\frac{\Delta Y}{\Delta X_2} = \beta_2 + \beta_3 X_1$

EXAMPLE: CALIFORNIA SCHOOLS DATASET

```
. reg testscr str el_pct str_el_pct, robust
```

Linear regression

Number of obs = 420
F(3, 416) = 155.05
Prob > F = 0.0000
R-squared = 0.4264
Root MSE = 14.482

testscr	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
str	-1.117018	.5875135	-1.90	0.058	-2.271884	.0378468
el_pct	-.6729116	.3741231	-1.80	0.073	-1.408319	.0624958
str_el_pct	.0011618	.0185357	0.06	0.950	-.0352736	.0375971
_cons	686.3385	11.75935	58.37	0.000	663.2234	709.4537

The interaction term (str*el_pct) is not significantly different from zero - effect of class sizes does not depend on share of English language learners.

EXAMPLE: AGE AND EARNINGS (CPS DATA)

```
. reg ln_incwage female age female_age, r
```

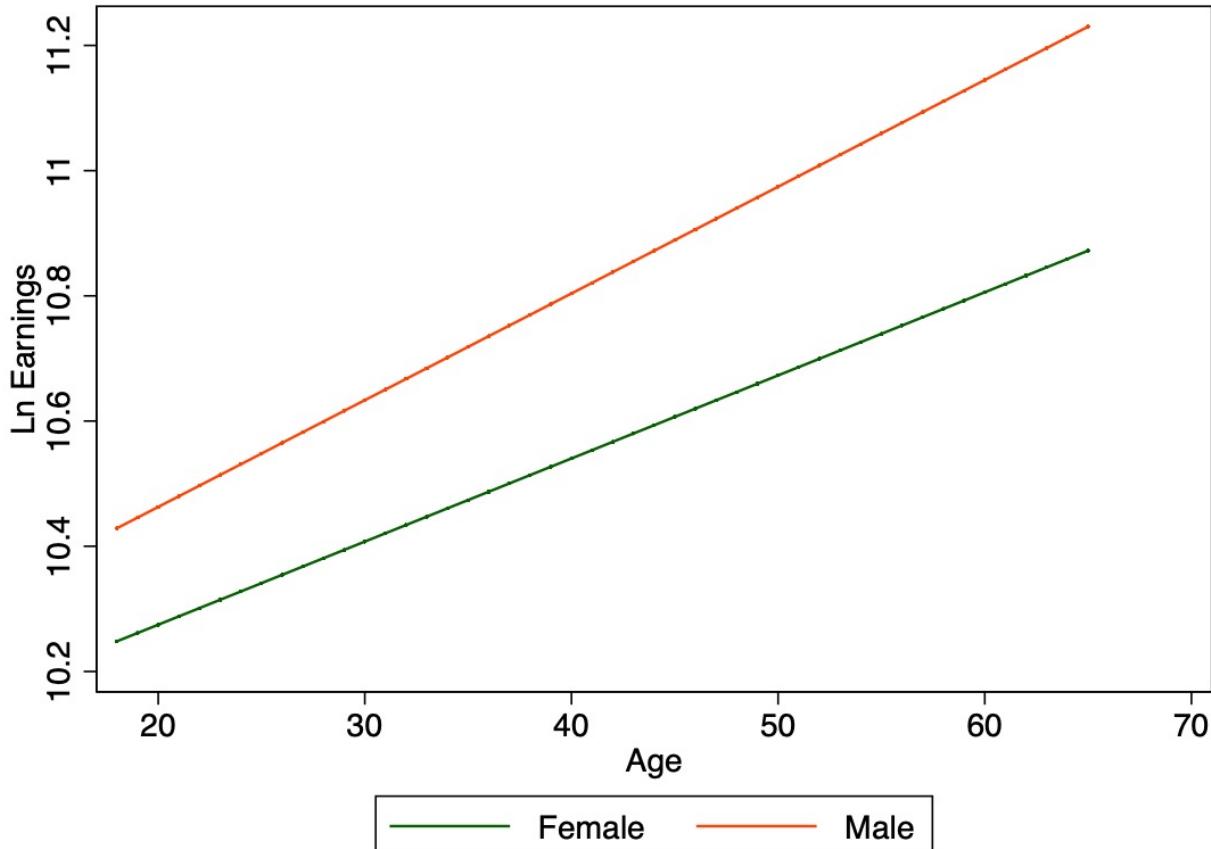
Linear regression

Number of obs = 55,527
F(3, 55523) = 1315.58
Prob > F = 0.0000
R-squared = 0.0726
Root MSE = .81175

ln_incwage	Robust					
	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.1126975	.0265112	-4.25	0.000	-.1646596	-.0607355
age	.017048	.0004219	40.40	0.000	.016221	.017875
female_age	-.0037716	.0006105	-6.18	0.000	-.0049682	-.0025751
_cons	10.12187	.0181315	558.25	0.000	10.08633	10.1574

The interaction term (female*age) is negative and significantly different from zero - effect of age on earnings does depend on gender.

EXAMPLE: AGE AND EARNINGS (CPS DATA)



The line for men is steeper than for women - women get less of an increase in earnings each year they get older compared to men.