Nonlinear Regression Functions

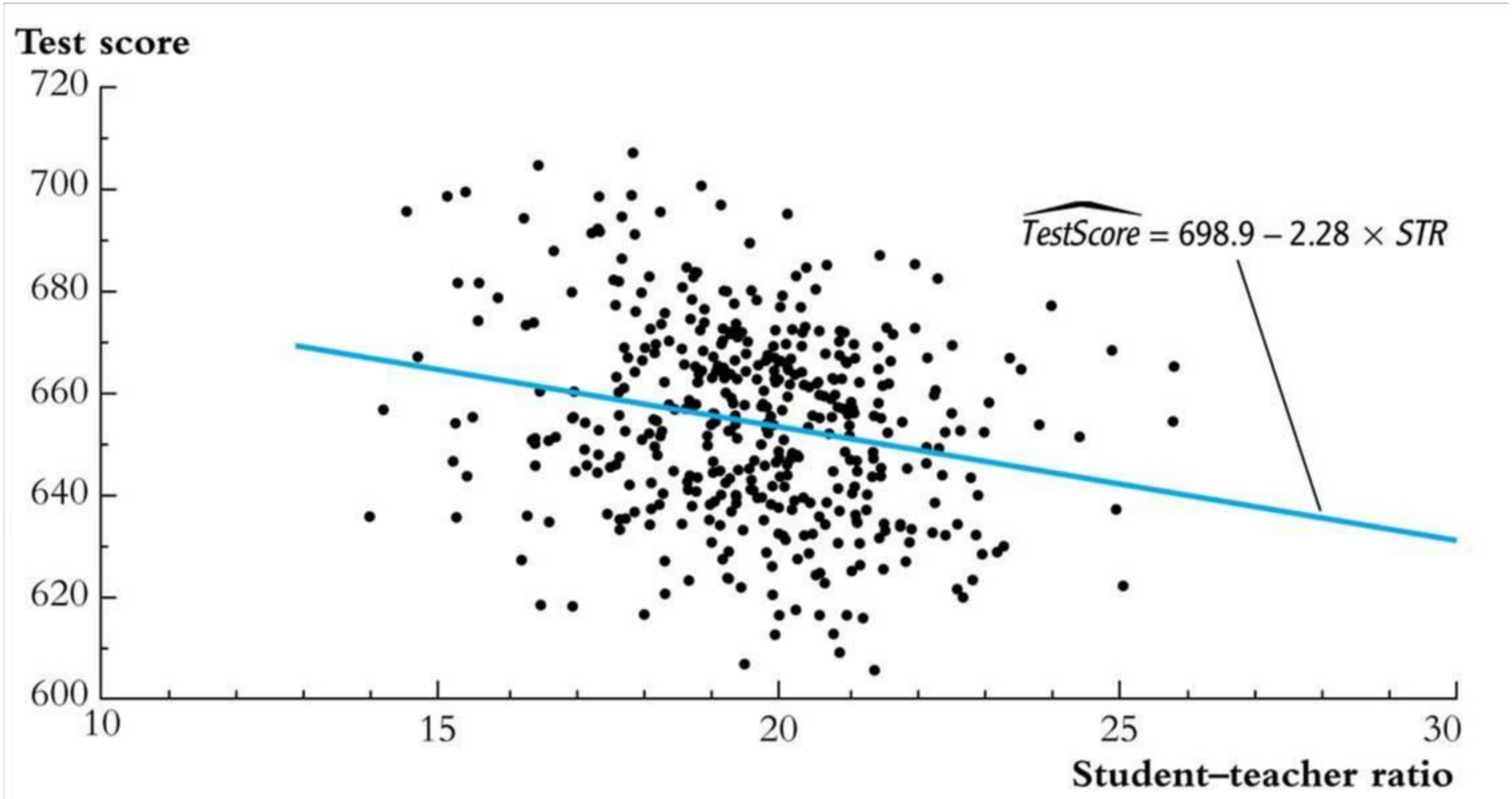
(SW Chapter 8)

- 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*.

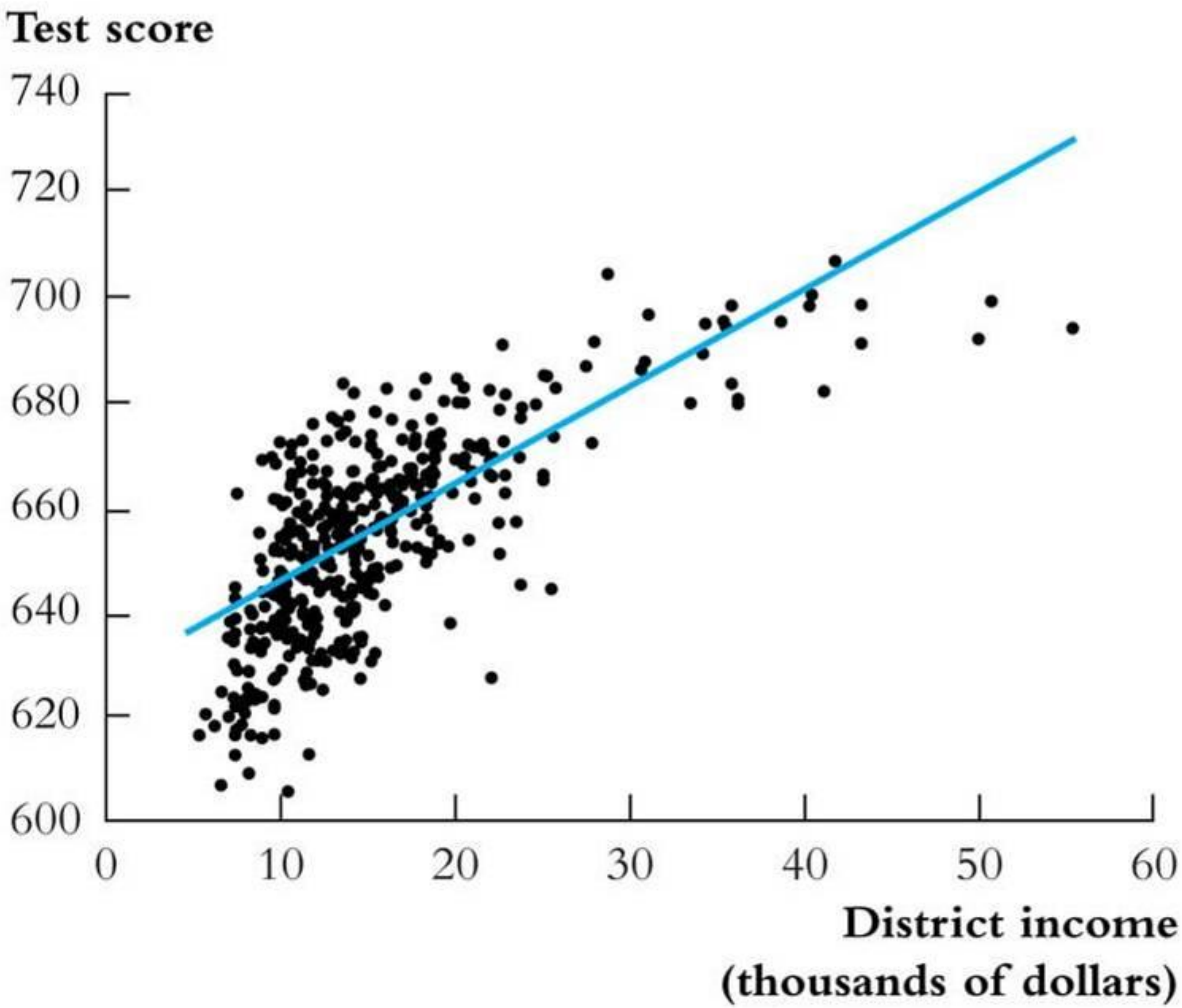
<u>Outline</u>

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

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

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:

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

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

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

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

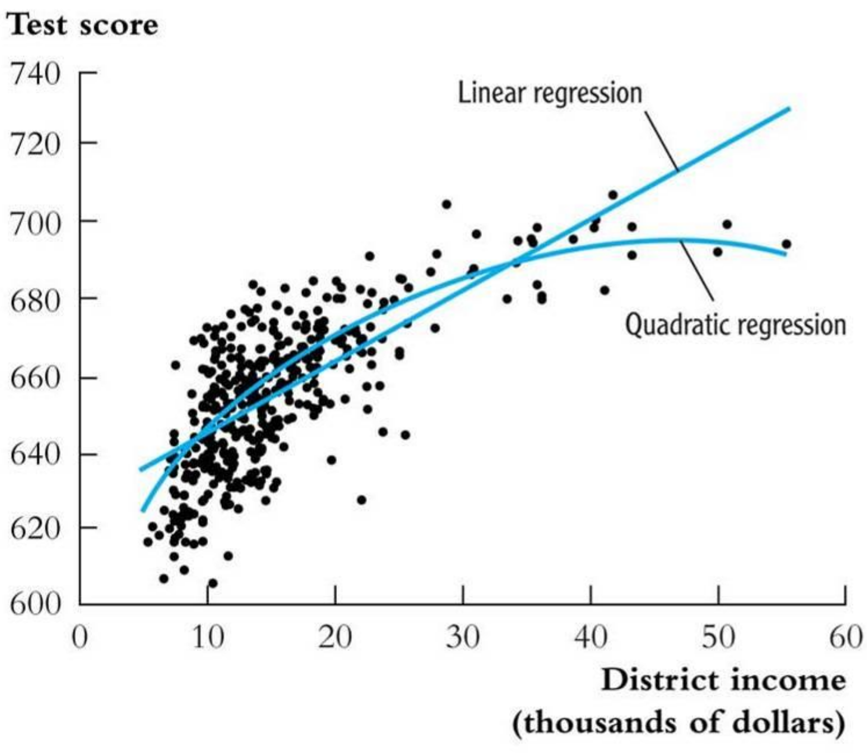
```
generate avginc2 = avginc*avginc;        Create a new regressor
reg testscr avginc avginc2, r;

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

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



Interpreting the estimated regression function, ctd:

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

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

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: 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*)