Nonlinear Regression Functions

Ming-Ching Luoh

2022.3.14.

General Strategy

Nolinear in a Single Variable

- Polynomials
- Logarithm

Interactions Between Independent Variables

- Between Two Dummies
- Between Continuous and Binary Variables
- Between Continuous Variables

Application

- Everything so far has been linear in the X's.
- The approximation that the regression function is linear might be good for some variables, but not for others.
- The multiple regression framework can be extended to handle regression functions that are nonlinear in one or more X's.

The *Test Score* - *STR* relation looks approximately linear.

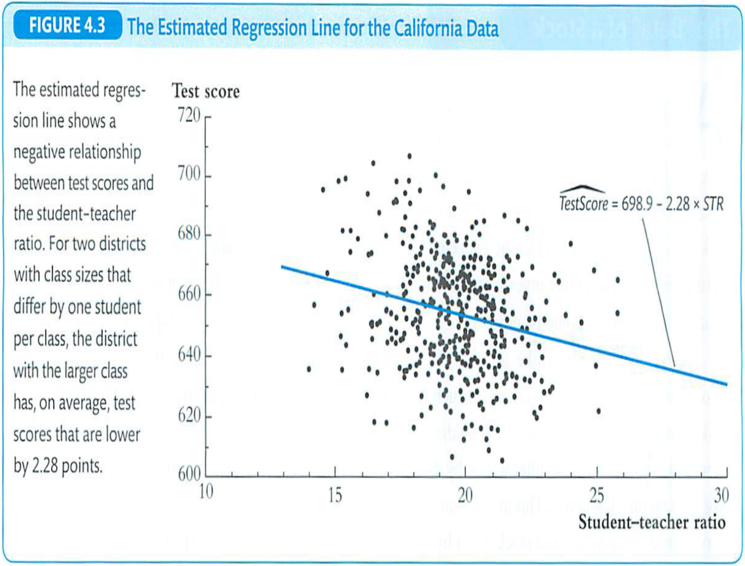


FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. For two districts with class sizes that differ by one student per class, the district with the larger class has, on average, test scores that are lower by 2.28 points.

But the *Test Score* - *Income* relation looks like it is nonlinear.

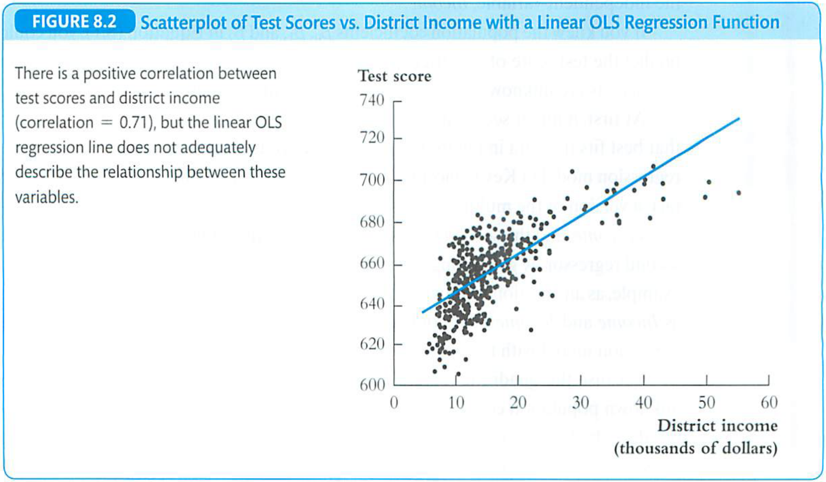


FIGURE 8.2 Scatterplot of Test Scores vs. District Income with a Linear OLS Regression Function

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.

A General Strategy for Modeling Nonlinear Regression Functions

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 need not even be right on average.
- The solution to this is to estimate a regression function that is nonlinear in X.

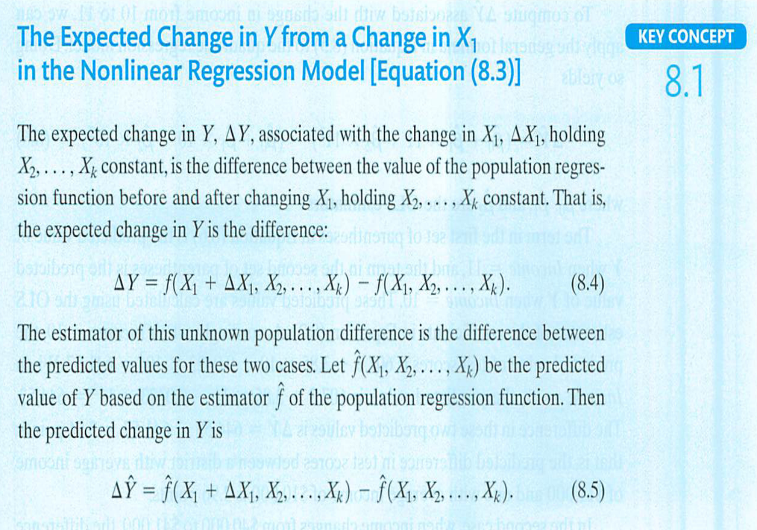
General Nonlinear Regression Function

The General Nonlinear Population Regression Function

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

Assumptions

1. $E(u_i | X_{1i}, X_{2i}, \cdots, X_{ki}) = 0$ (same); implies that f is the conditional expectation of Y given the X's.
2. $(X_{1i}, X_{2i}, \cdots, X_{ki})$ are $i.i.d.$ (same).
3. “enough" moments exist (same idea; the precise statement depends on specific f).
4. No perfect multicollinearity (same idea; the precise statement depends on the specific f).



KEY CONCEPT 8.1

The Expected Change in Y from a Change in X_1 in the Nonlinear Regression Model [Equation (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:

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

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

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

A General Approach to Modeling Nonlinearities Using Multiple Regression

1. Identify a possible nonlinear relationship.
2. Specify a nonlinear function and estimate its parameters by OLS.
3. Determine whether the nonlinear model improves upon a linear model.
4. Plot the estimated nonlinear regression function.
5. Estimate the effect of Y of a change in X.

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 "percentage" interpretation that makes sense in many applications.

1. Polynomials in X

Approximate the population regression function by a polynomial.

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

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

Example: the Test Score - Income relation

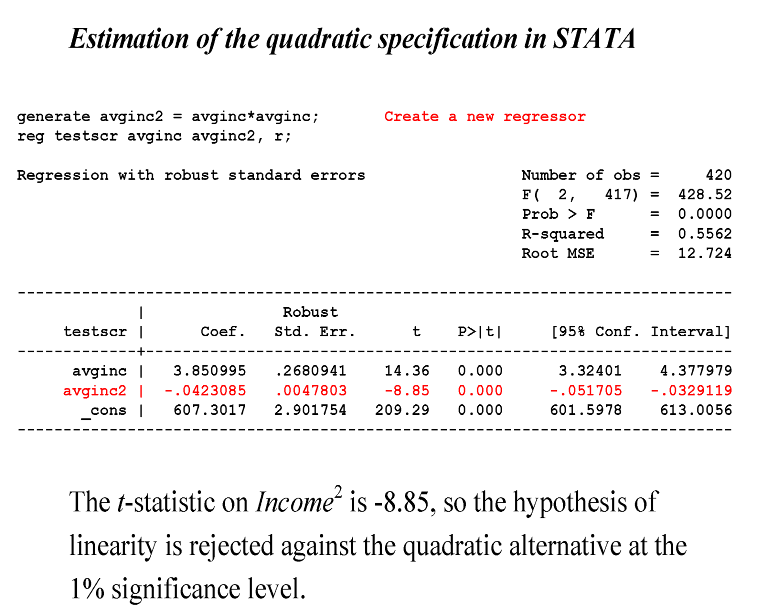
$Income_i$ = average district income in the ith district (thousand dollars per capita)

Quadratic specification:

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

Cubic specification:

$$\begin{aligned} Test\ Score_i &= \beta_0 + \beta_1 Income_i + \beta_2 (Income_i)^2 \\ &\quad + \beta_3 (Income_i)^3 + u_i \end{aligned}$$



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
------------------------------------------------------------------------------
```

The *t*-statistic on *Income*^{2} is -8.85, so the hypothesis of linearity is rejected against the quadratic alternative at the 1% significance level.

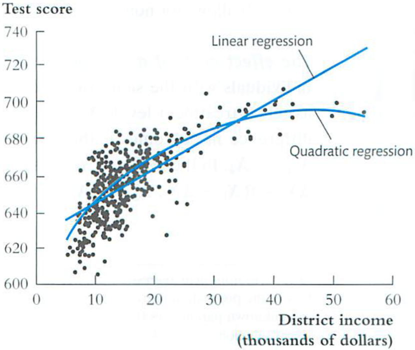
Interpreting the estimated regression function

(a) Plot the predicted values

$$
\begin{array}{rcl}
\widehat{TestScore}_i & = & 607.3 + 3.85 Income_i - 0.0423(Income_i)^2 \\
 & & (2.9) \quad (0.27) \qquad\qquad\quad (0.0048)
\end{array}
$$

FIGURE 8.3 Scatterplot of Test Scores vs. District Income with Linear and Quadratic Regression Functions

The quadratic OLS regression function fits the data better than the linear OLS regression function.



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

$$\begin{aligned}\widehat{TestScore_i} &= \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85} Income_i - \underset{(0.0048)}{0.0423}(Income_i)^2\end{aligned}$$

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

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

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

Predicted “effects" for different values of X

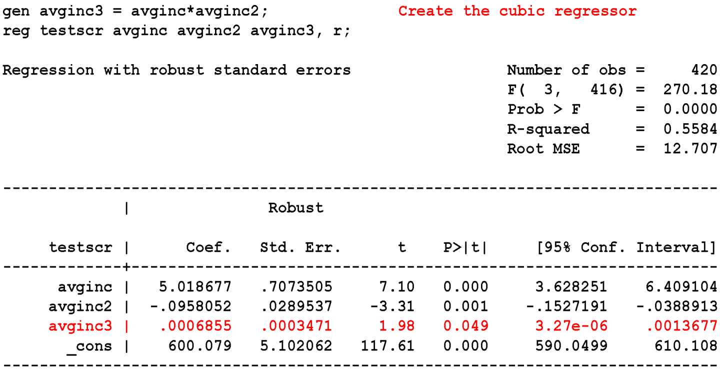
Change in $Income$ (th$ per capita)	$\Delta\widehat{Test\ Score}$
from 5 to 6	3.4
from 25 to 26	1.7
from 45 to 46	0.0

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 about a change from 65 to 66?

Don't extrapolate outside the range of the data.

Estimation of the cubic specification in STATA

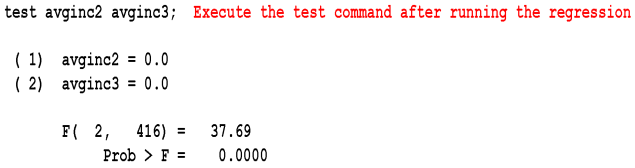


```
gen avginc3 = avginc*avginc2;            Create the cubic regressor
reg testscr avginc avginc2 avginc3, r;

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
------------------------------------------------------------------------------
```

The cubic term is statistically significant at the 5%, but not 1%, level

- 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: 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

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

- 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, then judgment.

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:

$$\begin{aligned} \ln(x+\Delta x)-\ln(x) &= \ln\left(1+\frac{\Delta x}{x}\right) \cong \frac{\Delta x}{x} \\ (\text{calculus}: \frac{d\ln(x)}{dx} &= \frac{1}{x}) \end{aligned}$$

$$\begin{aligned} \text{Numerically}: \ln(1.01) &= .00995 \cong .01 \\ \ln(1.10) &= .0953 \cong .10 \end{aligned}$$

Three cases:

1. linear-log: $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$.
2. log-linear: $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$.
3. 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)

1. Linear-log population regression function Compute Y "before" and "after" changing X:

$$\begin{aligned} Y &= \beta_0 + \beta_1 \ln(X) && (\text{"before"}) \\ Y + \Delta Y &= \beta_0 + \beta_1 \ln(X + \Delta X) && (\text{"after"}) \end{aligned}$$

Subtract ("after")-("before"):

$$\begin{aligned} \Delta Y = \beta_1[\ln(X + \Delta X) &- \ln(X)] \\ \ln(X + \Delta X) - \ln(X) &\cong \frac{\Delta X}{X} \\ \Delta Y &\cong \beta_1 \frac{\Delta X}{X} \\ \beta_1 &\cong \frac{\Delta Y}{\frac{\Delta X}{X}} (\text{ small} \Delta X) \end{aligned}$$

For Linear-log case,

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

for small ΔX,

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

- Now $100 \times \frac{\Delta X}{X}$ = percentage change in X, or **a 1% increase in X is associated with a $0.01\beta_1$ change in Y.**

Example: Test Score 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.

$$\begin{aligned}\widehat{Test\ Score}_i &= 557.8 + 36.42 \times \ln(Income)_i \\ &\ (3.8) \qquad (1.40)\end{aligned}$$

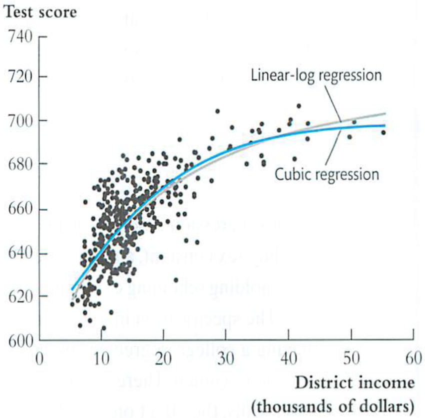
so a 1% increase in $Income$ is associated with an increase in $Test\ Score$ of 0.36 points on the test.

- 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 regression and cubic functions.

FIGURE 8.7 The Linear-Log and Cubic Regression Functions

The estimated cubic regression function [Equation (8.11)] and the estimated linear-log regression function [Equation (8.18)] are nearly identical in this sample.



2. Log-linear population regression function

$$\begin{aligned} \ln(Y) &= \beta_0 + \beta_1 X & (\text{“before"}) \\ \ln(Y+\Delta Y) &= \beta_0 + \beta_1 (X+\Delta X) & (\text{“after"}) \end{aligned}$$

Substract (“after")-(“before")

$$\begin{aligned} \ln(Y+\Delta Y) - \ln(Y) &= \beta_1 \Delta X \\ \frac{\Delta Y}{Y} &\cong \beta_1 \Delta X \\ \beta_1 &\cong \frac{\Delta Y / Y}{X} (\text{ small} \Delta X) \end{aligned}$$

For Log-linear case,

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

for small ΔX,

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

- Now $100 \times \frac{\Delta Y}{Y}$ = 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.

3. Log-log population regression function

$$
\begin{aligned}
\ln(Y) &= \beta_0 + \beta_1 \ln(X) && (\text{“before"}) \\
\ln(Y + \Delta Y) &= \beta_0 + \beta_1 \ln(X + \Delta X) && (\text{“after"})
\end{aligned}
$$

Substract (“after")-(“before")

$$
\begin{aligned}
\ln(Y + \Delta Y) - \ln(Y) &= \beta_1 [\ln(X + \Delta X) - \ln(X)] \\
\frac{\Delta Y}{Y} &\cong \beta_1 \frac{\Delta X}{X} \\
\beta_1 &\cong \frac{\Delta Y / Y}{\Delta X / X} (\text{ small}\Delta X)
\end{aligned}
$$

For 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 $100 \times \frac{\Delta Y}{Y}$ = 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: $\ln(Test\ Score)$ *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:

$$\begin{aligned} \ln \widehat{Test\ Score} &= \underset{(0.006)}{6.336} + \underset{(0.0021)}{0.0554} \times \ln(Income)_i \end{aligned}$$

An 1% increase in *Income* is associated with an increase of .0554% in *Test Score*

How does log-log compare to the log-linear model?

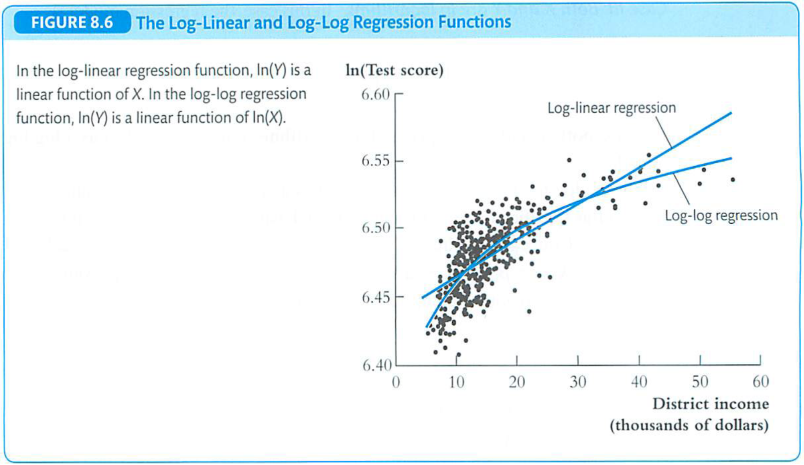


FIGURE 8.6 The Log-Linear and Log-Log Regression Functions

In the log-linear regression function, ln(*Y*) is a linear function of *X*. In the log-log regression function, ln(*Y*) is a linear function of ln(*X*).

Neither specification seems to fit as well as the linear-log and cubic regressions

Summary: Logarithmic transformations

- Three cases, differing in whether Y and/or X is transformed by taking logarithms.
- After creating the new variable(s) $\ln(Y)$ and/or $\ln(X)$, the regression is linear in the new variables and the coefficients can be estimated by OLS.
- Hypothesis tests and confidence intervals are now standard.
- 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.

Other nonlinear functions (and nonlinear least squares) (Appendix 8.1)

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

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

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

We want to estimate the parameters of

$$\begin{aligned} Y_i &= \beta_0 - \alpha e^{-\beta_1 X_i} + u_i \\ \text{or} \quad Y_i &= \beta_0 \left[1 - e^{-\beta_1 (X_i - \beta_2)} \right] + u_i \qquad (*) \end{aligned}$$

where $\alpha = \beta_0 e^{\beta_2}$.
Compare model (*) to linear-log or cubic models

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \ln(X_i) + u_i \\ Y_i &= \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + u_i \end{aligned}$$

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 $\cdots$
  - Implementation in STATA $\cdots$

. 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

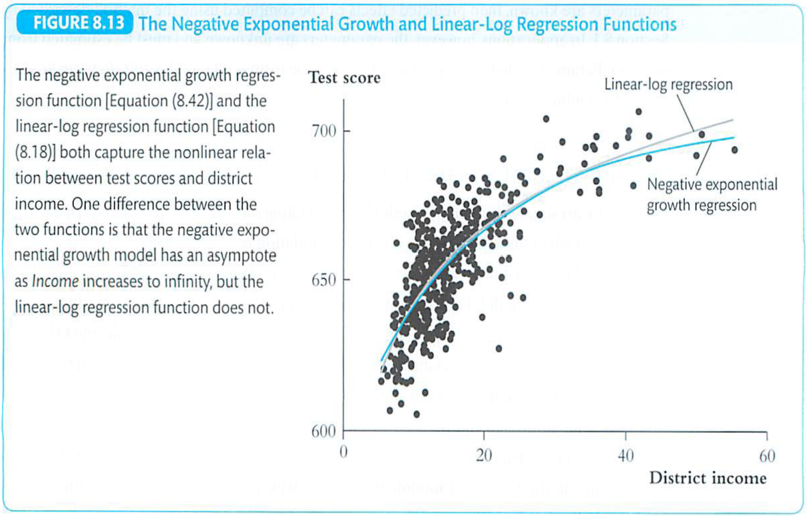


FIGURE 8.13 The Negative Exponential Growth and Linear-Log Regression Functions

The negative exponential growth regression function [Equation (8.42)] and the linear-log regression function [Equation (8.18)] both capture the nonlinear relation between test scores and district income. One difference between the two functions is that the negative exponential growth model has an asymptote as *Income* increases to infinity, but the linear-log regression function does not.

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, $\frac{\Delta Test\ Score}{\Delta STR}$ 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.

$$\begin{aligned} E(Y_i | D_{1i} = 0, D_{2i} = d_2) &= \beta_0 + \beta_2 d_2 \\ E(Y_i | D_{1i} = 1, D_{2i} = d_2) &= \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2 \end{aligned}$$

subtract:

$$\begin{aligned} E(Y_i | D_{1i} = 1, D_{2i} = d_2) &- E(Y_i | D_{1i} = 0, D_{2i} = d_2) \\ &= \beta_1 + \beta_3 d2 \end{aligned}$$

- The effect of D_1 depends on d_2.
- β_3 = increment to the effect of D_1, when $D_2 = 1$.

Example: TestScore, STR, English learners

Let

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

and

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

$$\begin{aligned} \widehat{Test\ Score} &= \underset{(1.4)}{664.1} - \underset{(2.3)}{18.2}HiEL - \underset{(1.9)}{1.9}HiSTR \\ &\quad - \underset{(3.1)}{3.5}(HiSTR \times HiEL) \end{aligned}$$

- “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.
- But, this interaction isn’t statistically significant: t = 3.5/3.1

(b) Interactions between continuous and binary variables

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

- 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.

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

Binary-continuous interactions: the two regression lines

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

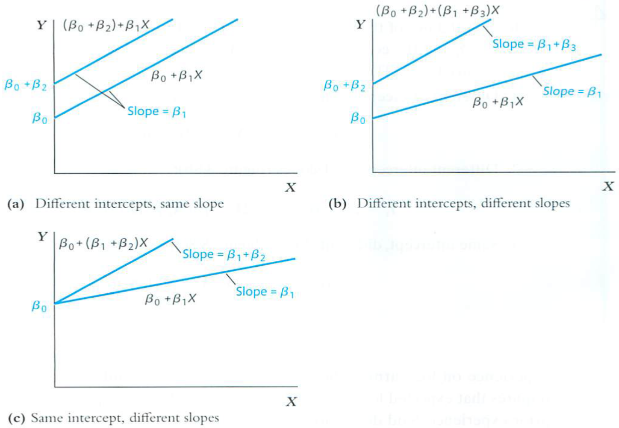
Observations with $D_i = 0$:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad \textbf{The D = 0 line}$$

Observations with $D_i = 1$:

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

FIGURE 8.8 Regression Functions Using Binary and Continuous Variables



(a) Different intercepts, same slope

(b) Different intercepts, different slopes

(c) Same intercept, different slopes

Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1 X + \beta_2 D$ allows for different intercepts but has the same slope, (b) $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$ allows for different intercepts and different slopes, and (c) $\beta_0 + \beta_1 X + \beta_2 (X \times D)$ has the same intercept but allows for different slopes.

Interpreting the coefficients

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i \\ \frac{\Delta Y}{\Delta X} &= \beta_1 + \beta_3 D \end{aligned}$$

- The effect of X depends on D.
- β_3 = increment to the effect of X, when $D = 1$.

Example:

$TestScore$, STR, $HiEL$ (=1 if $PctEL \geq 20$)

$$\begin{aligned}\widehat{Test\ Score} &= \underset{(11.9)}{682.2} - \underset{(0.59)}{0.97}STR + \underset{(19.5)}{5.6}HiEL \\ &\quad -\underset{(0.97)}{1.28}(STR \times HiEL)\end{aligned}$$

- When $HiEL = 0$

$$\widehat{Test\ Score} = 682.2 - 0.97STR$$

- When $HiEL = 1$,

$$\begin{aligned}\widehat{Test\ Score} &= 682.2 - 0.97STR + 5.6 \\ &\quad -1.28STR \\ &= 687.8 - 2.25STR\end{aligned}$$

- 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.

$$\begin{aligned}\widehat{Test\ Score} &= 682.2 - 0.97STR + 5.6HiEL \\ & \quad (11.9) \quad (0.59) \qquad\qquad (19.5) \\ & \quad -1.28(STR \times HiEL) \\ & \qquad (0.97)\end{aligned}$$

Testing various hypotheses:

- The two regression lines have the same slope ⇔ the coefficient on $STR \times HiEL$ is zero:

  $t = -1.28/0.97 = -1.32 \Rightarrow$ can't reject.

- The two regression lines have the same intercept ⇔the coefficient on $HiEL$ is zero: $t = -5.6/19.5 = 0.29 \Rightarrow$ can't reject

Example, ctd.

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

- **Joint** hypothesis that 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 $< .001$) !!

- Why do we reject the joint hypothesis but neither individual hypothesis?
- Consequence of high but imperfect multicollinearity: high correlation between $HiEL$ and $STR \times HiEL$.

(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.

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

Coefficients in continuous-continuous interactions

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

- The effect of X_1 depends on X_2.
- β_3= increment to the effect of X_1 from a unit change in X_2.

Example: TestScore, STR, PctEL

$$
\begin{aligned}
\widehat{Test\ Score} &= 686.3 - 1.12STR - 0.67PctEL \\
& \quad (11.8) \quad\; (0.59) \qquad\quad (0.37) \\
& \quad +.0012\,(STR \times PctEL) \\
& \qquad (0.019)
\end{aligned}
$$

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

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

$$
\begin{aligned}
\widehat{Test\ Score} &= 686.3 - 1.12STR - 0.67PctEL \\
&\quad (11.8) \quad (0.59) \qquad\quad (0.37) \\
&\quad +.0012(STR \times PctEL) \\
&\qquad (0.019)
\end{aligned}
$$

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

  $t = \frac{.0012}{.019} = .06 \Rightarrow$ can't reject at 5% level.

- Does population coefficient on $STR = 0$?

  $t = \frac{-1.12}{0.59} = -1.90 \Rightarrow$ can't reject null at 5% level.

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

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

Nonlinear Effects on Test Scores of the Student-Teacher Ratio

Nonlinear specifications let us examine more questions about the $Test\ Score$- STR relations, 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 $Test\ Score$-$Income$ 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 (nonlinear 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

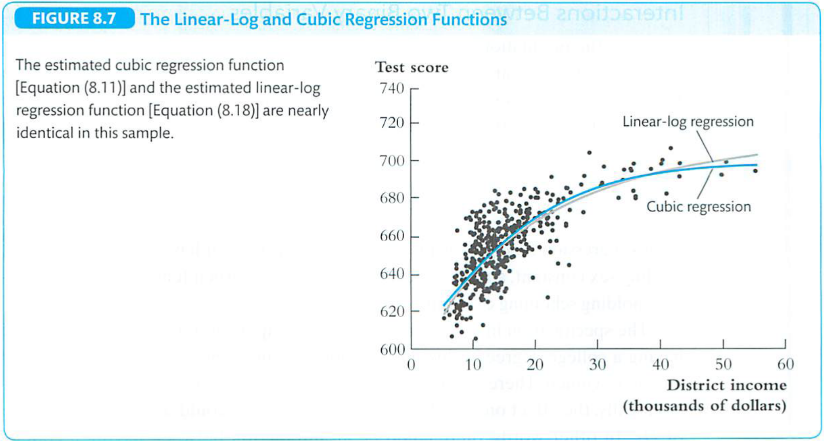


FIGURE 8.7 The Linear-Log and Cubic Regression Functions

The estimated cubic regression function [Equation (8.11)] and the estimated linear-log regression function [Equation (8.18)] are nearly identical in this sample.

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

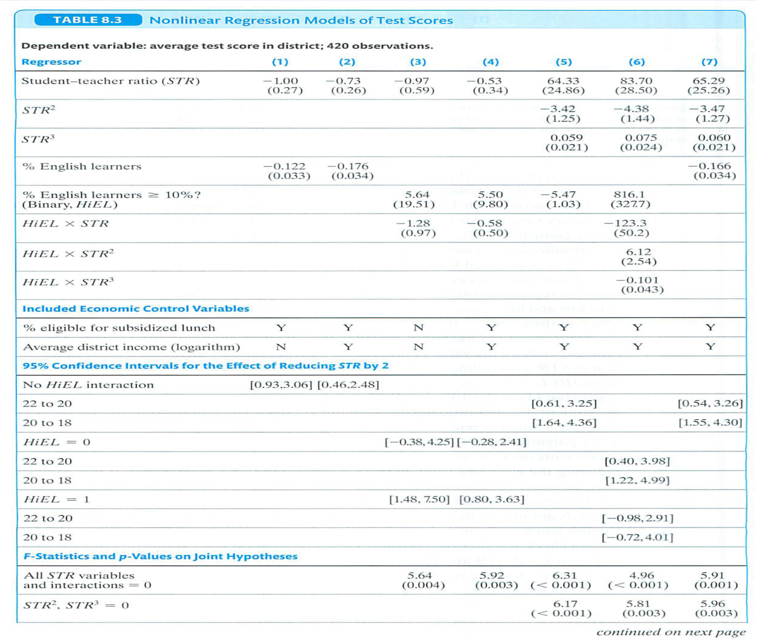
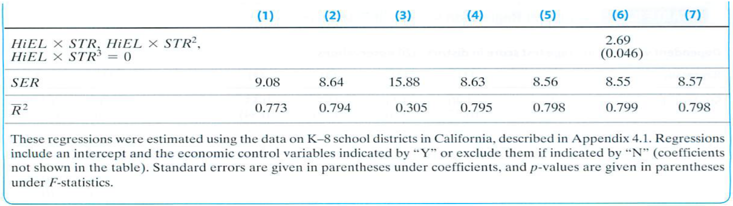


TABLE 8.3 Nonlinear Regression Models of Test Scores

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)	
Included Economic Control Variables							
% eligible for subsidized lunch	Y	Y	N	Y	Y	Y	Y
Average district income (logarithm)	N	Y	N	Y	Y	Y	Y
95% Confidence Intervals for the Effect of Reducing *STR* by 2							
No $HiEL$ interaction	[0.93,3.06]	[0.46,2.48]					
22 to 20					[0.61, 3.25]		[0.54, 3.26]
20 to 18					[1.64, 4.36]		[1.55, 4.30]
$HiEL = 0$			[−0.38, 4.25]	[−0.28, 2.41]			
22 to 20						[0.40, 3.98]	
20 to 18						[1.22, 4.99]	
$HiEL = 1$			[1.48, 7.50]	[0.80, 3.63]			
22 to 20						[−0.98, 2.91]	
20 to 18						[−0.72, 4.01]	
***F*-Statistics and *p*-Values on Joint Hypotheses**							
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)
STR^2, STR^3 = 0					6.17 (< 0.001)	5.81 (0.003)	5.96 (0.003)

continued on next page



	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$HiEL \times STR$, $HiEL \times STR^2$, $HiEL \times STR^3 = 0$						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. Regressions include an intercept and the economic control variables indicated by "Y" or exclude them if indicated by "N" (coefficients not shown in the table). Standard errors are given in parentheses under coefficients, and p-values are given in parentheses under F-statistics.

Tests of joint hypotheses:

Question #1:

Investigate by considering a polynomial in STR(column 5)

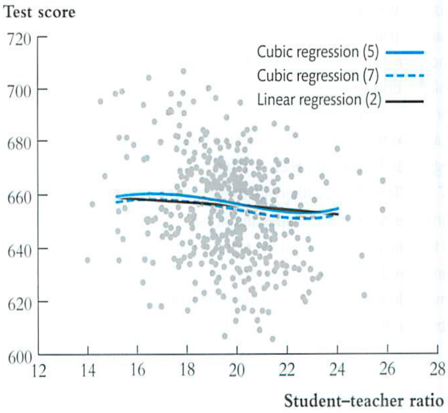
$$\begin{aligned} & \widehat{TestScore} \\ = & \underset{(163.6)}{252.0} + \underset{(24.86)}{64.33}STR - \underset{(1.25)}{3.42}STR^2 + \underset{(.021)}{.059}STR^3 \\ & -\underset{(1.03)}{5.47}HiEL - \underset{(.029)}{.420}LunchPCT + \underset{(1.78)}{11.75}\ln(Income) \end{aligned}$$

Interpreting the regression function via plots

(preceding regression is labeled (5) in this figure)

FIGURE 8.10 Three Regression Functions Relating Test Scores and Student-Teacher Ratio

The cubic regressions from columns (5) and (7) of Table 8.3 are nearly identical. They indicate a small amount of nonlinearity in the relation between test scores and student-teacher ratio.



Are the higher order terms in STR statistically significant?

$$
\begin{aligned}
&\widehat{TestScore} \\
= \quad & \underset{(163.6)}{252.0} + \underset{(24.86)}{64.33}STR - \underset{(1.25)}{3.42}STR^2 + \underset{(.021)}{.059}STR^3 \\
& -\underset{(1.03)}{5.47}HiEL - \underset{(.029)}{.420}LunchPCT + \underset{(1.78)}{11.75}\ln(Income)
\end{aligned}
$$

(a) H_0: quadratic in STR v. H_1: cubic in STR?

$$t = .059/.021 = 2.86(p = .005)$$

(b) H_0: linear in STR v. H_1: nonlinear/up to cubic in STR?

$$F = 6.17(p = .002)$$

Question #2:

$STR - PctEL$ interactions (column 4)

(to simplify things, ignore STR^2, STR^3 terms for now)

$$\begin{aligned} & \widehat{Test\ Score} \\ = \quad & 653.6 - .53 STR + 5.50 HiEL - .58 HiEL \times STR \\ & (9.9) \quad (.34) \qquad\qquad (9.80) \qquad\qquad (.50) \\ & -.411 LunchPCT + 12.12 \ln(Income) \\ & \quad (.029) \qquad\qquad\qquad (1.80) \end{aligned}$$

$$\begin{aligned}
&\widehat{Test\ Score} \\
= \quad & 653.6 - .53STR + 5.50HiEL - .58HiEL \times STR \\
& (9.9) \quad (.34) \qquad\qquad (9.80) \qquad\qquad\qquad (.50) \\
& -.411LunchPCT + 12.12\ln(Income) \\
& \quad (.029) \qquad\qquad\qquad\qquad (1.80)
\end{aligned}$$

“Real-world" importance of the interaction term:

$$\frac{\Delta \widehat{Test\ Score}}{\Delta STR} = -.53 - .58HiEL$$

It is -1.12 if $HiEL = 1$ and is -.53 if $HiEL = 0$.

- The difference in the estimated effect of reducing the STR is substantial; class size reduction is more effective in districts with more English learners.

Is the interaction effect statistically significant?

$$
\begin{aligned}
&\widehat{Test\ Score} \\
= \quad & \underset{(9.9)}{653.6} - \underset{(.34)}{.53}STR + \underset{(9.80)}{5.50}HiEL - \underset{(.50)}{.58}HiEL \times STR \\
& -\underset{(.029)}{.411}LunchPCT + \underset{(1.80)}{12.12}\ln(Income)
\end{aligned}
$$

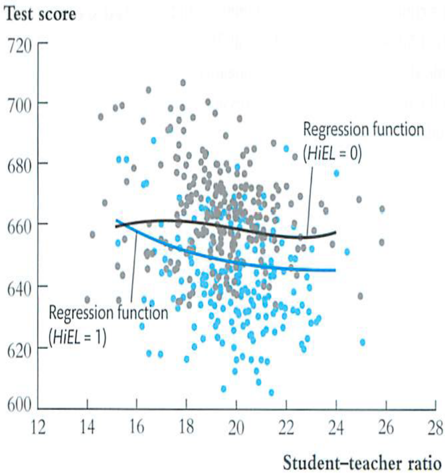
(a) H_0: coeff. on interaction=0 v. H_1: nonzero interaction $t = -0.58/0.50 = -1.17 \Rightarrow$ not significant at the 10% level. (b) H_0: both coeffs involving $STR = 0$ vs. H_1: at least one coefficient is nonzero (STR enters)

$$F = 5.92(p = .003)$$

Interpreting the regression functions via plots:

FIGURE 8.11 Regression Functions for Districts with High and Low Percentages of English Learners

Districts with low percentages of English learners (*HiEL* = 0) are shown by gray dots, and districts with *HiEL* = 1 are shown by colored dots. The cubic regression function for *HiEL* = 1 from regression (6) in Table 8.3 is approximately 10 points below the cubic regression function for *HiEL* = 0 for $17 \le STR \le 23$, but otherwise the two functions have similar shapes and slopes in this range. The slopes of the regression functions differ most for very large and small values of *STR*, for which there are few observations.



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 proceeds 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?