Nonlinear Regression Functions

Introduction to Econometrics,Fall 2017

Zhaopeng Qu

Nanjing University

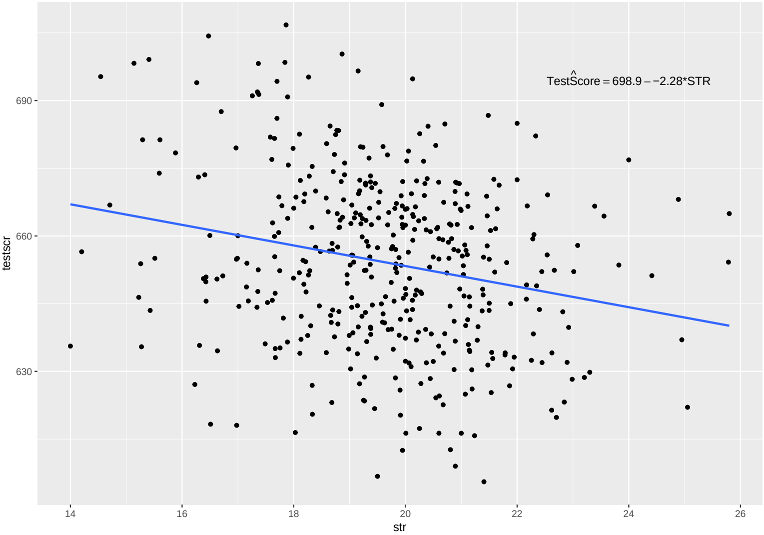
11/03/2017

Nonlinear Regression Functions:

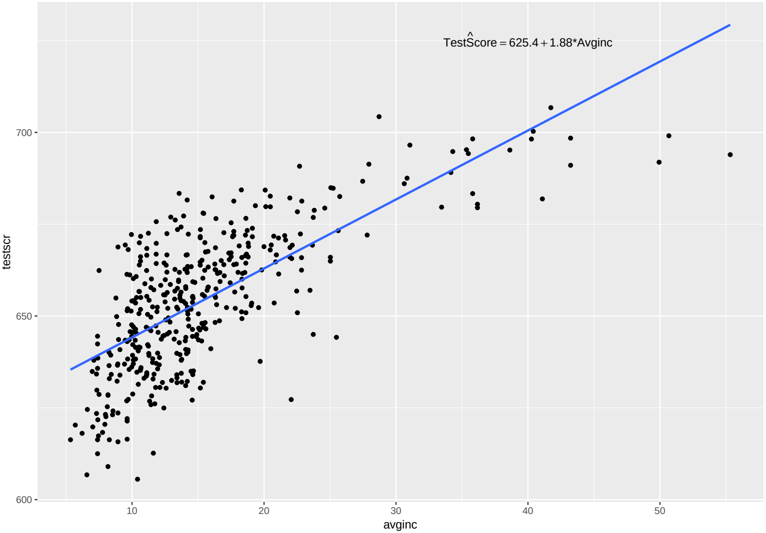
Introduction

- 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.

The TestScore – STR relation looks linear (maybe)



But the TestScore – Income relation looks nonlinear



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

$$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_k X_{k,i} + u_i$$

- The effect of a change in X_j by 1 is constant and equals βj :

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 misspecified – the functional form is wrong
- The estimator of the effect on Y of X is biased(a special case of OVB)
- The solution to this is to estimate a regression function that is nonlinear in X.

What are nonlinear regression functions: 2 Types

- There are 2 types of *nonlinear* regression models
  - Regression model that is a nonlinear function of the independent variables,

  $X_{1,i}, X_{2,i}, ..., X_{k,i}$ which is another version of multiple regression model and can be estimated by OLS.
  - Regression model that is a nonlinear function of the unknown coefficients, which can't be estimated by OLS, requires different estimation method.
- This lecture we will only consider first type of nonlinear regression models.

OLS Assumptions Still Hold

General formula for a nonlinear population regression model:

$$Y_i = f(X_{1,i}, X_{2,i}, ..., X_{k,i}) + u_i$$

Assumptions:

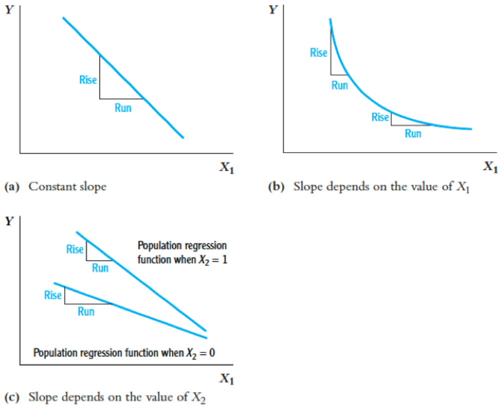
1. $E[u_i|X_{1,i}, X_{2,i}, ..., X_{k,i}] = 0$ implies that f is the conditional expectation of Y given the X's.
2. $(X_{1,i}, X_{2,i}, ..., X_{k,i})$ are i.i.d.
3. Large outliers are rare.
4. No perfect multicollinearity.

Two Cases:

- Two cases:
- The effect of change in X1 on Y depends on X1
- for example: the effect of a change in class size is bigger when initial class size is small
- The effect of change in X1 on Y depends on another variable X2
- For example: the effect of class size depends on the percentage of disadvantaged pupils in the class
- We start with case 1 using a regression model with only 1 independent variable

Different Slops

FIGURE 8.1 Population Regression Functions with Different Slopes



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

Figure 1: pic

The Effect on Y of a Change in X in a Nonlinear Specifications

The Expected Change 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:

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

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

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

A General Approach to Modeling Nonlinearities Using Multiple Regression

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

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

Polynomials in X

Polynomials in X

- Approximate the population regression function by a polynomial:

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

- 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

Testing the null hypothesis that the population regression function is linear

$$H_0 : \beta_2 = 0, \beta_3 = 0, ..., \beta_r = 0 \; and \; H_1 : at \; least \; one \; \beta_j \neq 0$$

- it can be tested using the F-statistic

Which degree polynomial should I use?

- how many powers of X should be included in a polynomial regression? The answer balances a trade-off between flexibility and statistical precision.
- In many applications involving economic data, the nonlinear functions are smooth, that is, they do not have sharp jumps, or "spikes."
- If so, then it is appropriate to choose a small maximum degree for the polynomial, such as 2, 3, or 4.

Example: the TestScore – Income relation

- 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 R

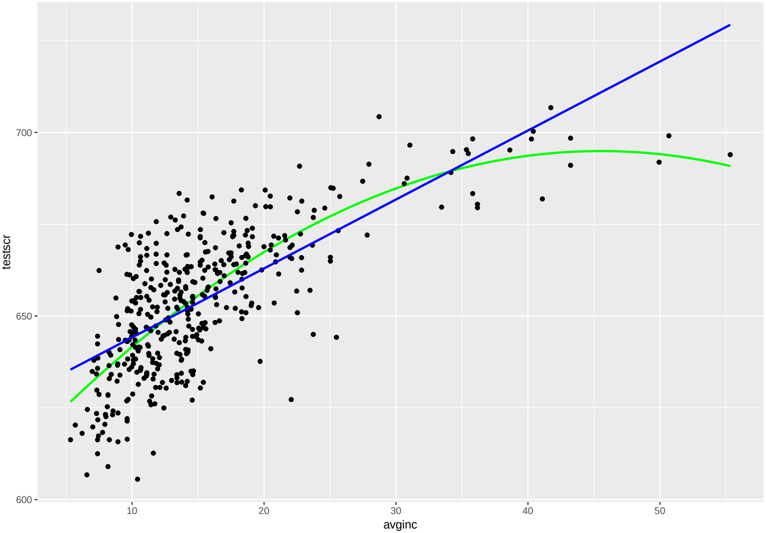
```
##
## Call:
##   felm(formula = testscr ~ avginc + I(avginc^2), data = ca)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -44.416  -9.048   0.440   8.348   31.639
##
## Coefficients:
##              Estimate Robust s.e t value Pr(>|t|)
## (Intercept) 607.30174    2.90175 209.288   <2e-16 ***
## avginc        3.85100    0.26809  14.364   <2e-16 ***
## I(avginc^2)  -0.04231    0.00478  -8.851   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Interpreting the estimated regression function

- The OLS regression yields

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

Linear and Quadratic Regression in figure



Quadratic vs Linear

- Is the quadratic model better than the linear model?
- We can test the null hypothesis that the regression function is linear against the alternative hypothesis that it is quadratic:

$$H_0 : \beta_2 = 0 \; and \; H_1 : \beta_2 \neq 0$$

- the t-statistic

$$t = \frac{(\hat{\beta}_2 - 0)}{SE(\hat{\beta}_2)} = \frac{-0.0423}{0.0048} = -8.81$$

- Since $8.81 > 2.58$ we reject the null hypothesis (the linear model) at a 1% significance level.

Interpreting the estimated regression function

- Predict Change in TestScore for a change in income
- from \$10,000 per capita to \$11,000 per capita:

$$\begin{aligned}\Delta TestScore &= 607.3 + 3.85 \times 11 - 0.0423 \times (11)^2 \\ &- [607.3 + 3.85 \times 10 - 0.0423 \times (10)^2] \\ &= 2.96\end{aligned}$$

- from \$40,000 per capita to \$41,000 per capita:

$$\begin{aligned}\Delta TestScore &= 607.3 + 3.85 \times 41 - 0.0423 \times (41)^2 \\ &- [607.3 + 3.85 \times 40 - 0.0423 \times (40)^2] \\ &= 0.42\end{aligned}$$

Logarithms

Logarithmic functions of Y and/or X

- Another way to specify a nonlinear regression model is to use the natural logarithm 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.

Review of the Logarithmic functions

$$
\begin{aligned}
ln(1/x) &= -ln(x) \\
ln(ax) &= ln(a) + ln(x) \\
ln(x/a) &= ln(x) - ln(a) \\
ln(x^a) &= aln(x)
\end{aligned}
$$

Logarithms and percentages

- Because

$$ln(x+\Delta x) - ln(x) = ln\left(\frac{x+\Delta x}{x}\right)$$
$$\cong \frac{\Delta x}{x}\ (when \frac{\Delta x}{x} is\ small)$$

- for example

$$ln(1+0.01) = ln(101) - ln(100) = 0.00995 \cong 0.01$$

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 population regression function

- the regression model is

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

- Change X:

$$\begin{aligned} \Delta Y &= [\beta_0 + \beta_1 ln(X + \Delta X)] - [\beta_0 + \beta_1 ln(X)] \\ &= \beta_1 [ln(X + \Delta X) - ln(X)] \\ &\cong \beta_1 \frac{\Delta X}{X} \end{aligned}$$

- Now $100\frac{\Delta X}{X} = percentage\ change\ in\ X$,so a 1% increase in X (multiplying X by 1.01) is associated with a $0.01\beta_1$ change in Y.

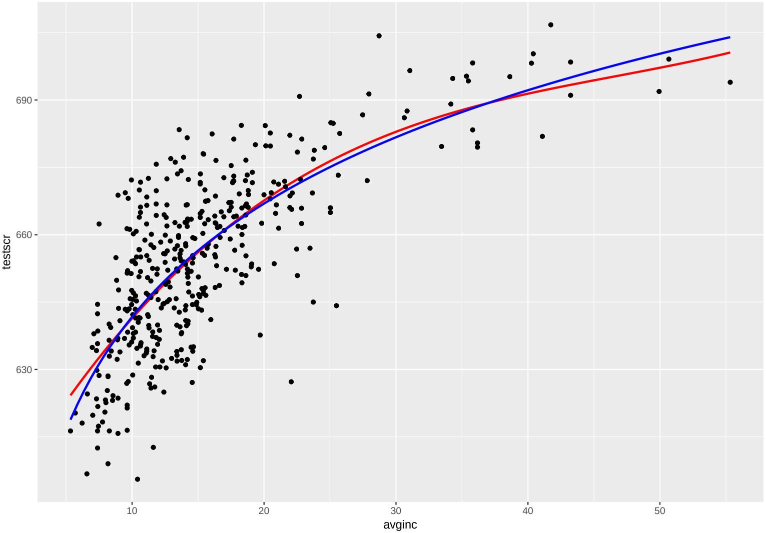
Example:the TestScore – log(Income) relation

- The OLS regression of ln(Income) on Testscore yields

$$\widehat{TestScore} = \underset{(3.8)}{557.8} + \underset{(1.4)}{36.42} \times ln(Income)$$

- so a 1% increase in Income is associated with an increase in TestScore of 0.36 points on the test.

Test scores: linear-log and cubic regression functions



Case II. Log-linear population regression function

- the regression model is

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

- Change X:

$$ln(\Delta Y + Y) - ln(Y) = [\beta_0 + \beta_1(X + \Delta X)] - [\beta_0 + \beta_1 X]$$
$$ln(1 + \frac{\Delta Y}{Y}) = \beta_1 \Delta X$$

- then

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

- Now $100\frac{\Delta Y}{Y} = percentage\ change\ in\ Y$,so a change in X by one unit is associated with a β_1% change in Y.

Case III. Log-linear population regression function

- the regression model is

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

- Change X:

$$\begin{aligned} ln(\Delta Y + Y) - ln(Y) &= [\beta_0 + \beta_1 ln(X + \Delta X)] - [\beta_0 + \beta_1 ln(X)] \\ ln(1 + \frac{\Delta Y}{Y}) &= ln(1 + \frac{\Delta X}{X}) \\ \frac{\Delta Y}{Y} &\cong \beta_1 \frac{\Delta X}{X} \end{aligned}$$

- Now $100\frac{\Delta Y}{Y} = percentage\ change\ in\ Y$ and $100\frac{\Delta X}{X} = percentage\ change\ in\ X$
- so a 1% change in X by one unit is associated with a β_1% change in Y,thus $\beta_1$has the interpretation of an **elasticity**.

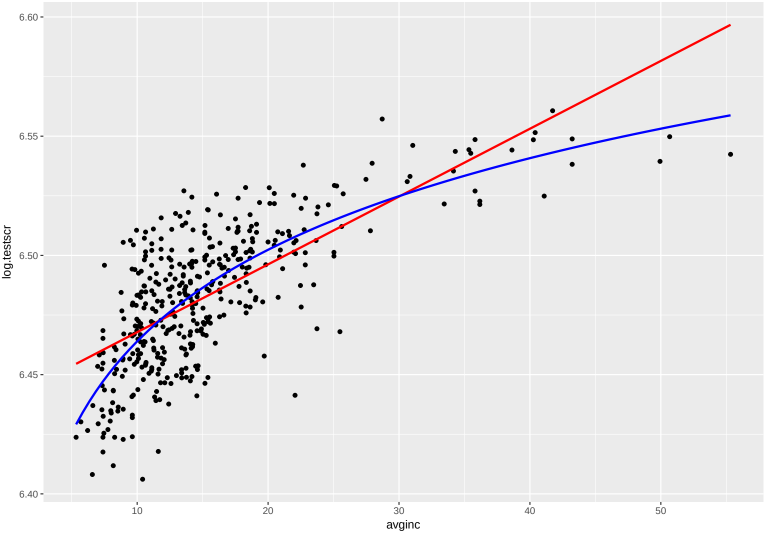
Test scores and income: log-log specifications

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 6.3363494  0.0059105 1072.056 < 2.2e-16 ***
## loginc      0.0554190  0.0021395   25.903 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

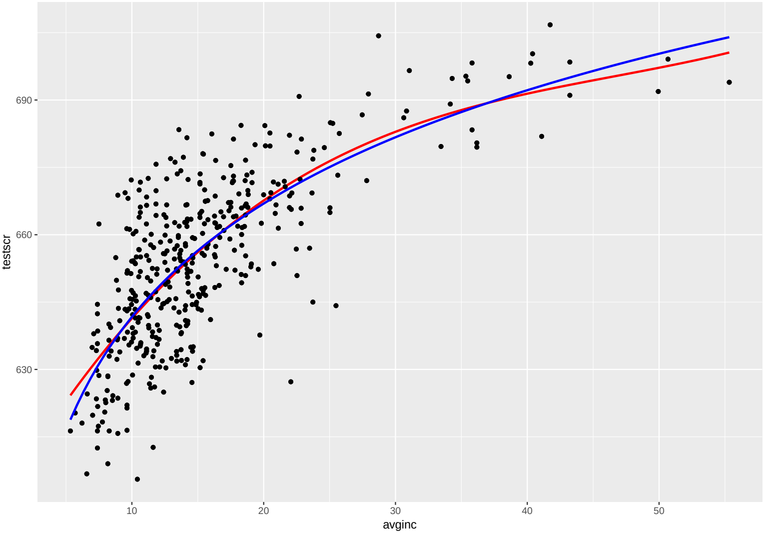
$$\begin{aligned} ln(\widehat{TestScore}) =& 6.336 + 0.055 \times ln(Income) \\ & (0.006) \quad (0.002) \end{aligned}$$

- An 1% increase in Income is associated with an increase of .0554% in TestScore.

Test scores: The log-linear and log-log specifications:



linear-log and cubic regression functions



Choice of specification should be guided

- by Economic logic or theories(which interpretation makes the most sense in your application?),
- formal tests(seldom use in reality)
- and plotting predicted values

Summary

- We already have a very powerful tool for detecting misspecified functional form: the F test for joint exclusion restrictions.
- We can add quadratic terms of any significant variables to a model and to perform a joint test of significance. If the additional quadratics are significant, they can be added to the model.
- It can be difficult to pinpoint the precise reason for functional form misspecification.
- Fortunately, using logarithms of certain variables and adding quadratic functions are sufficient for detecting many important nonlinear relationships in economics.