Nonlinear Regression [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

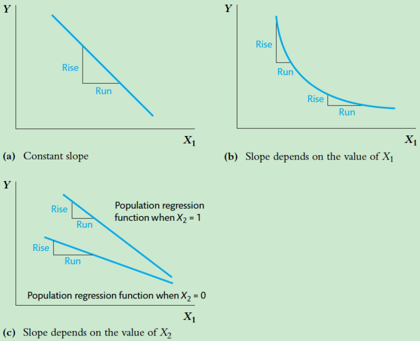
September 9, 2021

[1] This section is based on Stock and Watson (2020), Chapter 8.

Motivation I

Goal: two groups of methods for detecting and modeling nonlinear population regression functions.

FIGURE 8.1 Population Regression Functions with Different Slopes



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.

Motivation II

1. the relationship between Y and an independent variable, X_1, depends on the value of X_1 itself.
   - E.g. reducing class sizes by one student per teacher might have a greater effect if class sizes are already manageably small.
   - If so, the test score (Y) is a nonlinear function of the student–teacher ratio (X_1), where this function is steeper when X_1 is small.

Motivation III

2. the effect on Y of a change in X_1 depends on the value of another independent variable—say, X_2.
   - For example, students still learning English might especially benefit from having more one-on-one attention;
   - if so, the effect on test scores of reducing the student–teacher ratio will be greater in districts with many students still learning English than in districts with few English learners.

- If CEF is nonlinear function of the X's and of the parameters.
- If so, the parameters cannot be estimated by OLS,
- Can be estimated using nonlinear least squares.

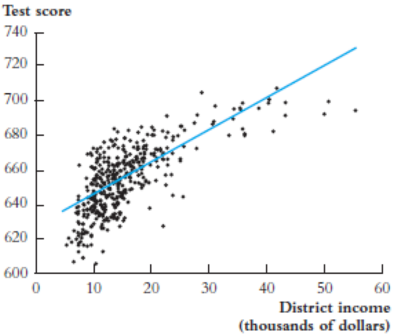
General Strategy I

- the nonlinear models are extensions of the multiple regression model and (can be estimated and tested using OLS).
- Consider the relationship between (*Test Scores* and *District Income*)
- Measure ratio of students with poor family background:
  - the percentage of students qualifying for a subsidized lunch
  - the percentage of students whose families qualify for income assistance.
  - Alternatively: the average annual per capita income in the school district ("district income").

General Strategy II

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.



- Data: income measured in thousands of 1998 dollars.
- median district income is 13.7($13,700 per person).
- ranges from 5.3 to 55.3.

General Strategy III

- Test scores and district income are strongly positively correlated, with a correlation coefficient of 0.71;
- There seems to be some curvature in the relationship between test scores and district income that is not captured by the linear regression.

General Formula for Nonlinear Regression

I

- The nonlinear population regression models considered in this chapter are of the form

$$Y_i = f(X_{1i}, X_{1i}, \ldots, X_{ki}) + u_i, i = 1, \ldots, n, (8.3)$$

where $f(X_{1i}, X_{1i}, \ldots, X_{ki})$ is the population nonlinear regression function, a possibly nonlinear function of the independent variables $X_{1i}, X_{1i}, \ldots, X_{ki}$ and u_i is the error term.

- E.g., in the quadratic regression model

$$f(Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2.$$

Partial Effect I

- Experiment: on individuals with the same values of $X_2, \ldots, X_k$, and participants are randomly assigned treatment levels $X_1 = x_1$ or $X_1 + \Delta X_1 = x_1 + \Delta x_1$.
- The expected difference in Y is the causal effect of the treatment, holding constant $X_2, \ldots, X_k$.
- In the quadratic regression model, this effect on Y is $\Delta Y = f(X_1 + \Delta X_1, \ldots, X_k) - f(X_1, X_2, \ldots, X_k)$.
- Δ is the predicted difference in with difference $X_1 + \Delta X_1$ and X_1.

Partial Effect II

- The regression function f is unknown, this population causal effect is also unknown.
  - First estimate the regression function f.
  - At a general level, denote this estimated function by f_n;
  - e.g. estimated quadratic regression function in Equation (8.2).
- $\hat{Y} = \hat{f}(X_1 + \Delta X_1, \ldots, X_k) - \hat{f}(X_1, X_2, \ldots, X_k)$.

Application example I

$$\widehat{TestScore} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85}\,Income - \underset{(0.0048)}{0.0423} Income^2, R^2 = 0.554, (8.2) \tag{1}$$

- **What is the predicted change in test cores associated with a change in district income of \$1000, based on the estimated quadratic regression function?**
- this effect depends on the initial district income. We therefore consider two cases: an increase in district income from 10 to 11 (i.e., from \$10,000 per capita to \$11,000 per capita) and an increase in district income from 40 to 41 (i.e., from \$40,000 per capita to \$41,000 per capita).
- $$\Delta \hat{Y} = (\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) \\ -(\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2).$$

Application example II

- when $Income = 10$, the predicted value of test scores is $607.3 + 3.85 * 10 - 0.0423 * 102 = 641.57$. When $Income = 11$, the predicted value is $607.3 + 3.85 * 11 - 0.0423 * 112 = 644.53$. The difference in these two predicted values is $Y_n = 644.53 - 641.57 = 2.96$ points.
- when income changes from \$40,000 to \$41,000, the difference in the predicted values in Equation (8.6) is $Y_n = (607.3 + 3.85 * 41 - 0.0423 * 41^2) - (607.3 + 3.85 * 40 - 0.0423 * 40^2) = 694.04 - 693.62 = 0.42$ points.
- Thus a change of income of \$1000 is associated with a larger change in predicted test scores if the initial income is \$10,000 than if it is from \$40,000 to \$41,000.

Standard errors of estimated effects I

- **Sampling error** The estimate of the effect on Y of changing X depends on $\hat{f}$, which varies from one sample to the next.
- Example:

$$SE(\Delta Y) = SE(\hat{\beta}_1 + 21\hat{\beta}_2)$$

  - if we can compute the standard error of $\hat{\beta}_1 + 21\hat{\beta}_2$ then we have computed the standard error of Yn.

Standard errors of estimated effects II

- If not directly computable, there are two other ways to compute it; these correspond to the two approaches for testing a single restriction on multiple coefficients.
  1. The first method is to use approach 1 of Section 7.3, which is to compute the F-statistic testing the hypothesis that $\beta_1 + 21\beta_2 = 0$. The standard error of Y_n is then given by

     $$SE(\Delta\hat{Y}) = \frac{|\Delta\hat{Y}|}{\sqrt{F}}.(8.8)$$

     the F-statistic testing the hypothesis that $\beta_1 + 21\beta_2 = 0$ is F $=$ 299.94. With $\Delta\hat{Y} = 2.96$, $SE(\Delta\hat{Y}) = 2.96/\sqrt{299.94} = 0.17$.
  2. The second approach is to compute the standard error using $var(\hat{\beta})$.

Comment on interpreting coefficients I

- In the multiple regression model, the regression coefficients had a natural interpretation.
- For example, β_1 is the expected change in Y associated with a change in X_1, holding the other regressors constant.
- But as we have seen, this is not generally the case in a nonlinear model. e.g. $\hat{\beta}_1$
- In nonlinear models, the regression function is best interpreted by graphing it and by calculating the predicted effect on Y of changing one or more of the independent variables.

General Approach I

The general approach to modeling nonlinear regression functions taken in this chapter has five elements:

1. Identify a possible nonlinear relationship.
   - Use economic theory and what you know about the application to suggest a possible nonlinear relationship.
   - Why might such nonlinear dependence exist? What nonlinear shapes does this suggest?
   - e.g. classroom dynamics with 11-year-olds suggests that cutting class size from 18 students to 17 could have a greater effect than cutting it from 30 to 29.
2. Specify a nonlinear function and estimate its parameters by OLS.
3. Determine whether the nonlinear model improves upon a linear model.
   - Use t-statistics and F-statistics to test the null hypothesis that the population regression function is linear against the alternative that it is nonlinear.
4. Plot the estimated nonlinear regression function. Does
5. Estimate the effect on Y of a change in X.

Polynomials I

- One way to specify a nonlinear regression function is to use a polynomial in X. In general, let r denote the highest power of X that is included in the regression. The polynomial regression model of degree r is

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

- When r = 2, the model is called **quadratic regression** model.
- When r = 3, the model is called **cubic regression model**.

Testing the null hypothesis that the population regression function is linear I

- If the population regression function is linear, then the quadratic and higher-degree terms do not enter the population regression function.
- The null hypothesis H_0 that the regression is linear and the alternative H_1 that it is a polynomial of degree up to r

$$H0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_r = 0 \text{ vs. } H1 : \text{ at least one } \beta_j \neq 0, j = 2,$$

- Can be tested against the alternative that it is a polynomial of degree up to r by testing H_0 against H_1 using joint null hypothesis with $q = r - 1$ restrictions.

Determine the order of the polynomial I

- how many powers of X should be included in a polynomial regression? trade-off between flexibility and statistical precision.
- Increasing the degree r introduces more flexibility.
- reduce the precision of the estimated coefficients.
- Include enough to model the nonlinear regression function adequately—but no more.

Determine the order of the polynomial II

Sequential Test:

1. Pick a maximum value of r, and estimate the polynomial regression for that r
2. Test $H_0 : \beta_r = 0$, if reject, test for $r + 1$.
3. If fail to reject: test $r - 1$.
4. continue this procedure until the coefficient on the highest power in your polynomial is statistically significant.

Application to district income and test scores I

The estimated cubic regression function relating district income to test scores is

$$\widehat{TestScore} = \underset{(5.1)}{600.1} + \underset{(0.71)}{5.02}\,Income - \underset{(0.029)}{0.096}\,Income^2 + \underset{(0.00035)}{0.00069}\,Income^3,$$
$$\bar{R}^2 = 0.555.$$

- The t-statistic on $Income^3$ is 1.97, so the null hypothesis that the regression function is a quadratic is rejected against the alternative that it is a cubic at the 5% level.
- the F-statistic testing the joint null hypothesis that the coefficients on $Income^2$ and $Income^3$ are both 0 is 37.7, with a p-value less than 0.01%.
- null hypothesis that the regression function is linear is rejected against the alternative that it is either a quadratic or a cubic.

Interpretation of coefficients in polynomial regression models I

- The coefficients in polynomial regressions do not have a simple interpretation.
- The best way to interpret polynomial regressions is to plot the estimated regression function and calculate the estimated effect on Y associated with a change in X for one or more values of X.

Logarithms I

- Another way to specify a nonlinear regression function is to use the natural logarithm of Y and/or X.
- Logarithms convert changes in vairables into percentage changes.

Example

- SW Ch 3, "Social Class or Education? Childhood Circumstances and Adult Earnings Revisited," examined the household earnings gap by socioeconomic classification.
  - Easier to compare wage gaps across professions and over time when they are expressed in percentage terms.
- In SW 8.1, we found that district income and test scores were nonlinearly related.
  - might it be that a change in district income of 1%—rather than \$1000—is associated with a change in test scores that is approximately constant.
- In the economic analysis of consumer demand, it is often assumed that a 1% increase in price leads to a certain percentage decrease in the quantity demanded(price **elasticity**).

Logarithm Function

The logarithm function has the following useful properties

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

Logarithms and percentages I

The link between the logarithm and percentages relies on a key fact: When x is small, the difference between the logarithm of $x + \Delta x$ and the logarithm of x is approximately $\Delta x / x$, the percentage change in x divided by 100. That is,

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}.(\text{ when } \frac{\Delta x}{x} \text{ is small.})$$

Proof: using taylor expansion.

The three logarithmic regression models I

1. linear-log model. $Y_i = \beta_0 + \beta_1 \ln(X_i)$. In the linear-log model, a 1% change in X is associated with a change in Y of $0.01\beta 1$.
2. log-linear model. $\ln Y_i = \beta_0 + \beta_1 X_i$. a one-unit change in $\Delta X_1 = 1$ is associated with a $(100 * \beta_1)\%$ change in Y.
3. log-log model. $\ln Y_i = \beta_0 + \beta_1 \ln X_i$. In the log-log model, a 1% change in X is associated with a β_1 change in Y.

A difficulty with comparing logarithmic specifications I

- the R^2 can be used to compare the log-linear and log-log models; as it happened, the log-log model had the higher R^2.
- the R^2 can be used to compare the linearlog regression in and the linear regression of Y against X.

In the test score and district income regression, the linear-log regression has an R^2 of 0.561, while the linear regression has an R^2 of 0.508, so the linear-log model fits the data better.

A difficulty with comparing logarithmic specifications II

How can we compare the linear-log model and the log-log model?

- Unfortunately, the R^2 cannot be used to compare these two regressions because their dependent variables are different.
- Recall that the R^2 measures the fraction of the variance of the dependent variable explained by the regressors.
- The dependent variables in the log-log and linear-log models are different, it does not make sense to compare their R^2's.

Use economic theory to decide.

Computing predicted values of Y when Y is in logarithms I

Consider the log-linear regression model,

$$Y_i = \exp(\beta_0 + \beta_1 X_i + u_i).$$

Then $\hat{Y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_i)$.

Polynomial and Logarithmic Models have of Test Scores and District Income I

- economic theory or expert judgment might suggest a functional form to use,
- the true form of the population regression function is unknown.
- Need to decide which method or combination of methods works best.

Polynomial specifications. Because the coefficient on $Income^3$ was significant at the 5% level, select the cubic model as the preferred polynomial specification.
Logarithmic specifications. The logarithmic specification seemed to provide a good fit to these data.
One way to test: to augment it with higher powers of the logarithm of income. If not statistically different from 0, then we can conclude

Polynomial and Logarithmic Models have of Test Scores and District Income II

that the specification is adequate in the sense that it cannot be rejected against a polynomial function of the logarithm.

$$\widehat{TestScore} = \frac{(79.4)}{486.1} + \frac{(87.9)}{113.4} ln(Income) - \frac{(31.7)}{26.93} ln(Income)^2 + \frac{(3.74)}{3.063} ln(Income)^3, R^2 = 0.560.$$

he F-statistic testing the joint hypothesis that the true coefficients on the quadratic and cubic term are both 0 is 0.44, with a p-value of 0.64.

Interactions Between Two Binary Variables I

- **Binary variable interaction regression model** Introducing another regressor, the product of the two binary variables, $D_{1i} * D_{2i}$. The resulting regression is

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

- The new regressor, the product $D_{1i} * D_{2i}$, is called an interaction term or an **interacted regressor**.
-

$$E[Y_i|D_{1i} = 1, D_{2i} = d_2] - E[Y_i|D_{1i} = 0, D_{2i} = d_2] = \beta_1 + \beta_3 d_2.$$

- The effect of acquiring a college degree depends on the person's sex.
- The binary variable interaction regression allows the effect of changing one of the binary independent variables to depend on the value of the other binary variable.

Application to the student–teacher ratio and the percentage of English learners I

- $HiSTR_i$ be a binary variable that equals 1 if the student–teacher ratio is 20 or more and that equals 0 otherwise.
- $HiEL_i$ be a binary variable that equals 1 if the percentage of English learners is 10% or more and that equals 0 otherwise.

$$\widehat{TestScore} = \underset{(1.4)}{664.1} - \underset{(1.9)}{1.9}\,HiSTR - \underset{(2.3)}{18.2}HiEL - \underset{(3.1)}{3.51}HiSTR * HiEL,$$
$$\bar{R}^2 = 0.29$$

-
- The predicted effect of moving from a district with a low student–teacher ratio to one with a high student–teacher ratio is $-1.9 - 3.5HiEL$.
- if the fraction of English learners is low $HiEL = 0$, then the effect on test scores of moving from $HiSTR = 0$ to $HiSTR = 1$ is for test scores to decline by 1.9 points.

Application to the student–teacher ratio and the percentage of English learners II

- If the fraction of English learners is high, then test scores are estimated to decline by $1.9 + 3.5 = 5.4$ points.
- Can use to estimate the mean test scores for each of the four possible combinations of the binary variables.
- $HiSTR_i = 0$ (low student–teacher ratios) and $HiEL_i = 0$ (low fractions of English learners) is 664.1.
- $HiSTR_i = 1$ (high student–teacher ratios) and $HiEL_i = 0$ (low fractions of English learners), the sample average is 662.2
- When $HiSTR_i = 0$ and $HiEL_i = 1$, the sample average is 645.9
- $HiSTR_i = 1$ and $HiEL_i = 1$, the sample average is 640.5.

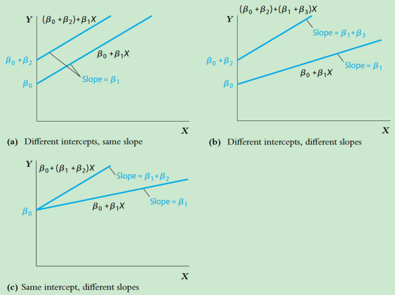
Interactions Between a Continuous and a Binary Variable I

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i * D_{i2}) + u_i,$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i,$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_3 (X_i * D_{i2}) + u_i.$$

where $X_i * D_i$ is a new variable, the product of X_i and D_i.

Interactions Between a Continuous and a Binary Variable II

FIGURE 8.8 Regression Functions Using Binary and Continuous Variables



Interactions of binary variables and continuous variables can produce three different population regression functions: (a) $\beta_0 + \beta_1X + \beta_2D$ allows for different intercepts but has the same slope, (b) $\beta_0 + \beta_1X + \beta_2D + \beta_3(X \times D)$ allows for different intercepts and different slopes, and (c) $\beta_0 + \beta_1X + \beta_2(X \times D)$ has the same intercept but allows for different slopes.

All three specifications are versions of the multiple regression model, and once the new variable $X_i * D_i$ is created, the coefficients of all three can be estimated by OLS.

Application to the student–teacher ratio and the percentage of English learners I

Research Question: Does the effect on test scores of cutting the student–teacher ratio depend on whether the percentage of students still learning English is high or low?
Test if the slope/intercept is different.

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

- reducing the student–teacher ratio by 1 is predicted to increase test scores by **0.97** points in districts with low fractions of English learners but by **2.25** points in districts with high fractions of English learners.
- The difference between these two effects, 1.28 points, is the coefficient on the interaction term.
- nuanced policy counterfactuals.

Application to the student–teacher ratio and the percentage of English learners II

- Test
  1. the two lines are the same. This F-statistic is 89.9, significant at the 1% level.
  2. the hypothesis that two lines have the same slope can be tested. (Cannot reject at 10 % level)
  3. Third, the hypothesis that the lines have the same intercept. (Cannot reject at 5 % level)
- Contradictory results.
- The reason is that the regressors, $HiEL$ and $STR * HiEL$, are highly correlated.
- results in **large standard errors** on the individual coefficients.
- impossible to tell which of the coefficients is nonzero, there is strong evidence against the hypothesis that both are 0.

Application to the student–teacher ratio and the percentage of English learners III

- Finally, the hypothesis that the student–teacher ratio does not enter this specification can be tested by computing the F-statistic for the joint hypothesis that the coefficients on STR and on the interaction term are both 0. This F-statistic is 5.64(p-value: 0.004).

Application to the student–teacher ratio and the percentage of English learners I

$$\widehat{TestScore} = \underset{(11.8)}{686.3} - \underset{(0.59)}{1.12}\,STR - \underset{(0.37)}{0.67}\,PctEL + \underset{(0.019)}{0.0012}(STR * PctEL),$$
$$\bar{R}^2 = 0.422.(8.37)$$

- When the percentage of English learners is at the median ($PctEL = 8.85$), the slope of the line relating test scores and the student–teacher ratio is estimated to be -1.11 $(= -1.12 + 0.0012 * 8.85)$.
- at the 75th percentile ($PctEL = 23.0$), slope of -1.09 $(= -1.12 + 0.0012 * 23.0)$.
- The difference between these estimated effects is not statistically significant ($t = 0.0012/0.019 = 0.06$).

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.