Lecture 10: Nonlinear Regression Functions

Zheng Tian

Contents

1 Introduction

1.1 Overview

In the previous lectures, the population regression function is assumed to be linear. In other words, the slope of the population regression function was constant, so the effect on Y of a unit change in X does not itself depend on the value of X, neither on other independent variables. In some cases, the linearity cannot capture the feature of the data and may also contradict with economic theories or common sense.

This lecture introduces nonlinear population regression functions that are nonlinear with respect to the regressors but linear with respect to the parameters. Such regression models can still be considered as being "linear" due to the linearity in the parameters so that they can be estimated using the OLS method we have known. Specifically, we will introduce three types of nonlinear regression functions, the polynomial function, the logarithmic function, and the function with the interaction terms consisting of two independent variables.

Since these types of models can be estimated using the OLS, the estimation and inference are not the foci of this lecture. Instead, we put emphasis on the correct interpretation of the coefficients in each type of models.

1.2 Reading materials

- Chapter 8 in *Introduction to Econometrics* by Stock and Watson.

2 A General Strategy For Modeling Nonlinear Regression Functions

2.1 Test Scores and district income

In the application of California elementary schools, we know that test scores can be determined by average district income, along with student-teacher ratios that we have included in regression. For simplicity, let's just consider the relationship between test scores and district income using a scatterplot. As shown in Figure 1, test scores and district income are indeed positively correlated, with a correlation coefficient of 0.71.

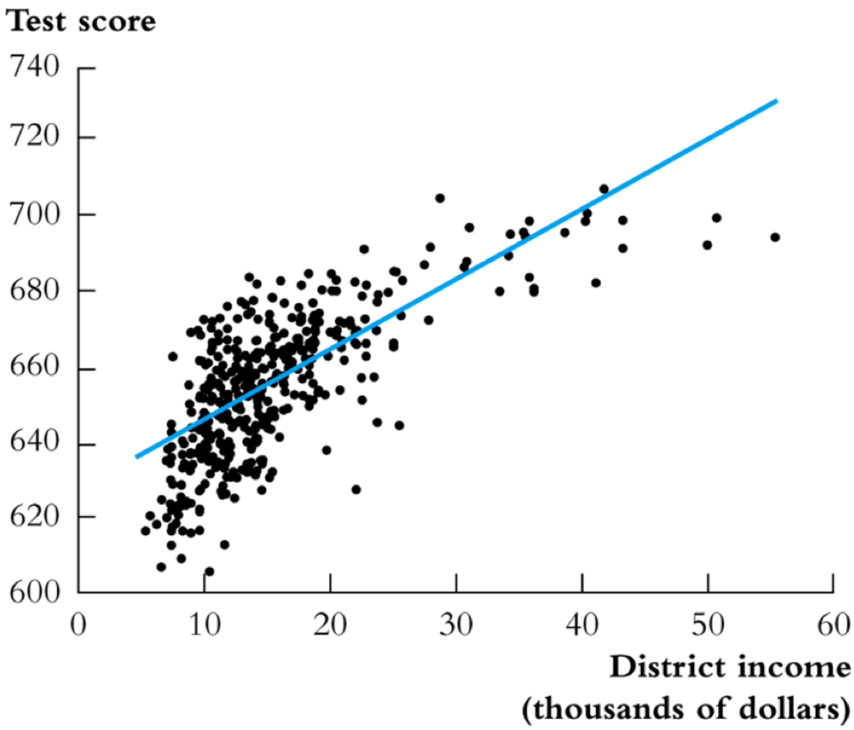


Figure 1: Scatterplot of test score vs district income and a linear regression line

How does a simple linear regression model fit the data?

We estimate a simple linear regression model

$$TestScore = \beta_0 + \beta_1 Income + u$$

The sample regression line is superimposed on the scatterplot. However, we can observe some problems of using the linear regression line to fit the data

- Data points are below the regression line when income is very low (under \$10,000) or very high (over \$40,000), and are above the line when income is between \$15,000 and \$30,000.
- The scatterplot may imply a curvature in the relationship between test scores and income. That is, a unit increase in income may have larger effect on test scores when income is very low than when income is very high.
- The linear regression line cannot capture the curvature because the effect of district income on test scores is constant over all the range of income since $\Delta TestScore/\Delta Income = \beta_1$ is constant.

Estimate a quadratic regression model

Instead of estimating a linear regression model, we can set up a quadratic regression model as

$$TestScore = \beta_0 + \beta_1 Income + \beta_2 Income^2 + u \tag{1}$$

- This model is nonlinear, specifically quadratic, with respect to $Income$ because we include the squared income.
- The population regression function is

$$E(TestScore|Income) = \beta_0 + \beta_1 Income + \beta_2 Income^2$$

- But it is linear with respect to β. So we can still use the OLS estimation method to estimate the model, and use R^2, t and F statistics for inference as we do in multiple regression. We simply treat $Income^2$ as another regressor in a multiple regression model.
- The quadratic regression line is added onto the scatterplot, which fits the points better than the linear regression line, as shown in Figure 2.

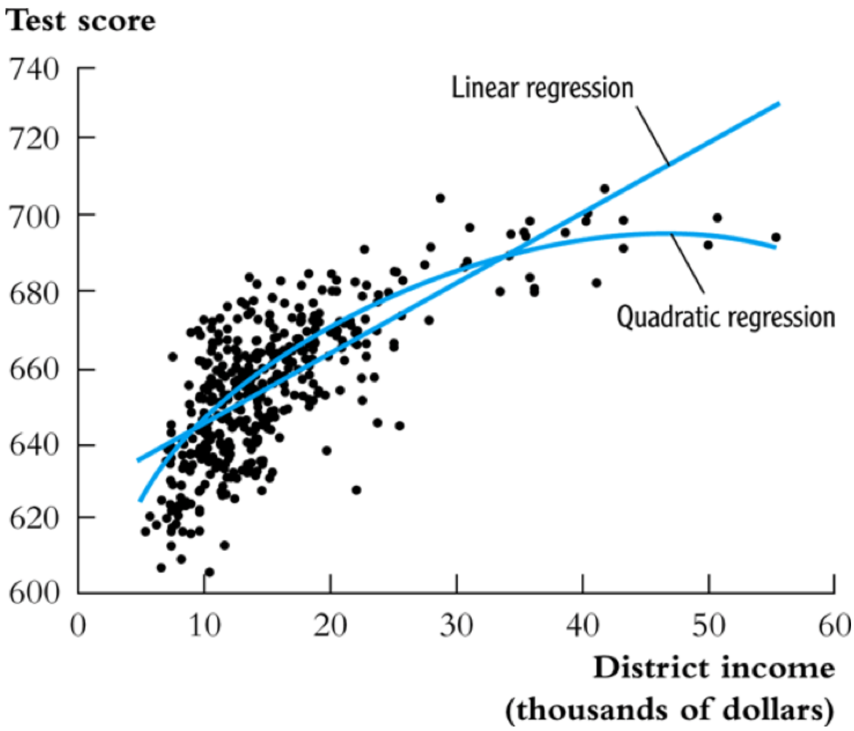


Figure 2: Scatterplot of test score vs district income and a quadratic regression line

2.2 A general formula for a nonlinear population regression function

A general nonlinear regression model is

$$Y_i = f(X_{i1}, X_{i2}, \ldots, X_{ik}; \beta_1, \beta_2, \ldots, \beta_m) + u_i \tag{2}$$

where $f(X_{i1}, X_{i2}, \ldots, X_{ik}; \beta_1, \beta_2, \ldots, \beta_m)$ is the **population nonlinear regression function**. Note that the number of regressors and the number of parameters are not necessarily equal in the nonlinear regression model.

We can use $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})$ to represent all regressors for the i$^{\text{th}}$ observation, and use $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$ to represent the parameters to be estimated. Then, the nonlinear regression model in Equation (2) can be re-written as

$$Y_i = f(\mathbf{X}_i; \boldsymbol{\beta}) + u_i \tag{3}$$

In this lecture, we focus on the nonlinear regression models such that $f(\cdot)$ is nonlinear with $\mathbf{X}_i$ but linear with $\boldsymbol{\beta}$. So this type of models are also consider as being "linear" and estimated by the OLS method.

2.3 The effect on Y of a change in a regressor

For the general nonlinear model in Equation (2), the effect on Y of a change in one regressor, say X_1, holding other things constant, can be computed as

$$\Delta Y = f(X_1 + \Delta X_1, X_2, \ldots, X_k; \boldsymbol{\beta}) - f(X_1, X_2, \ldots, X_k; \boldsymbol{\beta}) \tag{4}$$

When X_1 and Y are continuous variables and $f(\cdot)$ is differentiable, the marginal effect of X_1 is the partial derivative of f with respect to X_1, that is, holding other things constant

$$\mathrm{d}Y = \frac{\partial f(X_1, \ldots, X_k; \boldsymbol{\beta})}{\partial X_i} \mathrm{d}X_i$$

because $\mathrm{d}X_j = 0$ for $j \neq i$.

2.4 Application to test scores and income

We estimate the quadratic regression model for test scores and district income (Equation 1) by OLS, resulting in the following

$$\widehat{TestScore} = \underset{(2.9)}{607.3} + \underset{(0.27)}{3.85}\ Income - \underset{(0.0048)}{0.0423}\ Income^2,\ \bar{R}^2 = 0.554 \tag{5}$$

We can test whether the squared income has a significant coefficient. That is, we test $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$. In other words, we test the quadratic regression mode against the linear regression model. For this two-sided test, we can as usual compute the t-statistic

$$t = \frac{-0.0423}{0.0048} = -8.81 > -1.96$$

Thus, we can reject the null at the 1%, 5% and 10% significance levels.

From Equation (5), we can compute the effect of change in district average income on test scores.

A change in income from \$10 thousand to \$20 thousand

$$\begin{aligned}
\Delta \hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2 - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2) \\
&= \hat{\beta}_1(11 - 10) + \hat{\beta}_2(11^2 - 10^2) \\
&= 3.85 - 0.0423 \times 21 = 2.96
\end{aligned}$$

Thus, the predicted difference in test scores between a district with average income of \$11,000 and one with average income of \$10,000 is 2.96 points.

A change in income from \$40 thousand to \$41 thousand

$$\begin{aligned}\Delta\hat{Y} &= \hat{\beta}_0 + \hat{\beta}_1 \times 41 + \hat{\beta}_2 \times 41^2 - (\hat{\beta}_0 + \hat{\beta}_1 \times 40 + \hat{\beta}_2 \times 40^2) \\ &= \hat{\beta}_1(41 - 40) + \hat{\beta}_2(41^2 - 40^2) \\ &= 3.85 - 0.0423 \times 81 = 0.42\end{aligned}$$

Thus, the predicted difference in test scores between a district with average income of \$41,000 and one with average income of \$40,000 is 0.42 points. Thus, a change of income of \$1,000 is associated with a larger change in predicted test scores if the initial income is \$10,000 than if it is \$40,000.

2.5 A general approach to modeling nonlinearities using multiple regression

1. Identify a possible nonlinear relationship.
   - Economic theory
   - Scatterplots
   - Your judgment and experts' opinions
2. Specify a nonlinear function and estimate its parameters by OLS.
   - The OLS estimation and inference techniques can be used as usual when the regression function is linear with respect to β.
3. Determine whether the nonlinear model can improve a linear model
   - Use t- and/or 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.
5. Compute the effect on Y of a change in X and interpret the results.

3 Nonlinear functions of a single independent variable

3.1 Polynomials

A polynomial regression model of degree r

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i \tag{6}$$

- $r = 2$: a **quadratic** regression model

- $r = 3$: a **cubic** regression model
- A polynomial regression model is just regression of Y_i on regressors $X_i, X_i^2, \dots, X_i^r$. So use the OLS method to estimate $\beta_1, \beta_2, \dots, \beta_r$.

Testing the null hypothesis that the population regression function is linear

$$H_0: \beta_2 = 0, \beta_3 = 0, ..., \beta_r = 0 \text{ vs. } H_1: \text{ at least one } \beta_j \neq 0, j = 2, \dots, r$$

- Use F statistic to test this joint hypothesis. The number of restriction is $q = r - 1$.

What is $\Delta Y/\Delta X$ in a polynomial regression model?

- Consider a cubic model and continuous X and Y

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + u$$

- Then, we can calculate

$$\frac{\mathrm{d}Y}{\mathrm{d}X} = \beta_1 + 2\beta_2 X + 3\beta_3 X^2$$

- The effect of a unit change in X on Y depends on the value of X at evaluation.

Which degree polynomial should I use?

- Balance a trade-off between flexibility and statistical precision.
  - Flexibility. Relate Y to X in more complicated way than simple linear regression.
  - Statistical precision. $X, X^2, X^3, \dots$ are correlate so that there is the problem of imperfect multicollinearity.
- Follow a sequential hypothesis testing procedure
  1. Pick a maximum value of r and estimate the polynomial regression for that r.
  2. Use the t-statistic to test the hypothesis that the coefficient on X^r is zero. If you reject this hypothesis, then X^r belongs in the regression, so use the polynomial of degree r.
  3. If you do not reject $\beta_j = 0$ in step 2, eliminate X^r from the regression and estimate a polynomial regression of degree $r - 1$. Test whether the coefficient on X^{r-1} is zero. If you reject, use the polynomial of degree $r - 1$.

4. If you do not reject $\beta_{r-1} = 0$ in step 3, continue this procedure until the coefficient on the highest power in your polynomial is statistically significant.

- Quite often, we use a maximum degree of 2, 3, or 4.

Application to district income and test scores

We estimate a cubic regression model relating test scores to district income as follows

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

- Test whether it is a cubic model. We can test the null $H_0 : \beta_3 = 0$ using the t-statistic, which is 1.97 so that we can reject the null at the 5% level but fail to reject the null at the 1% level.

- Test whether it is a nonlinear model. In this case, we test the null $H_0 : \beta_2 = \beta_3 = 0$ using the F-statistic, which is 37.7 with the p-value less than 0.01% so that we reject the null at the 1% level.

- Interpretation of coefficients. The coefficients in polynomial regressions do not have a simple interpretation. The best way to interpret 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.

3.2 Logarithms

A natural logarithmic function $y = \ln(x)$

- Properties of $\ln(x)$

$$\ln(1/x) = -\ln(x),\ \ln(ax) = \ln(a) + \ln(x)$$
$$\ln(x/a) = \ln(x) - \ln(a), \text{ and } \ln(x^a) = a\ln(x)$$

- The derivative of $\ln(x)$ is

$$\frac{\mathrm{d}\ln(x)}{\mathrm{d}x} = \lim_{\Delta x \to 0} \frac{\ln(x + \Delta x) - \ln(x)}{\Delta x} = \frac{1}{x}.$$

  It follows that $\mathrm{d}\ln(x) = \mathrm{d}x/x$, representing the percentage change in x.

- We can also denote the "percentage-change" form using the Δ operator, that is,

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

We can reach the above equation by the Taylor expansion of $\ln(x + \Delta x)$ at x, which is

$$\begin{aligned}\ln(x + \Delta x) &= \ln(x) + \frac{\mathrm{d}\ln(x)}{\mathrm{d}}(x + \Delta x - x) + \frac{1}{2!}\frac{\mathrm{d}^2\ln(x)}{\mathrm{d}x^2}(x + \Delta x - x)^2 + \cdots \\ &= \ln(x) + \frac{\Delta x}{x} - \frac{\Delta x^2}{2x^2} + \cdots\end{aligned}$$

When Δx is very small, we can omit the terms with $\Delta x^2, \Delta x^3$, etc. Thus, we have $\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$ when Δx is small.

The three logarithmic regression models

There are three types of logarithmic regression models. Differences in logarithmic transformation of X and/or Y lead to differences in interpretation of the coefficient.

- **Case I: linear-log model**

  In this case X is in logarithms, Y is not.

  $$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i, i = 1, \ldots, n \tag{7}$$

  In the linear-log model, a 1% change in X is associated with a change in Y of $0.01\beta_1$ because

  $$\Delta Y = \beta_1 \ln(X + \Delta X) - \beta_1 \ln(X) \approx \beta_1 \frac{\Delta X}{X}$$

  If X changes by 1%, then $\Delta X/X = 0.01$ and $\Delta Y = 0.01\beta_1$. Using the derivative of $ln(x)$, we can easily see that $\beta_1 = \mathrm{d}Y/\mathrm{d}\ln(X) = \mathrm{d}Y/(\mathrm{d}X/X)$.

  e.g. Suppose that we have the estimated model as

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

  Then it implies that 1% increase in average district income results in an increase in test scores by $0.01 \times 36.42 = 0.36$ point.

- **Case II: log-linear model**

  In this case Y is in logarithms, X is not.

  $$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i \tag{8}$$

  In the log-linear model, a one-unit change in X is associated with a $100 \times \beta_1\%$ change in Y because

  $$\frac{\Delta Y}{Y} \approx \ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X$$

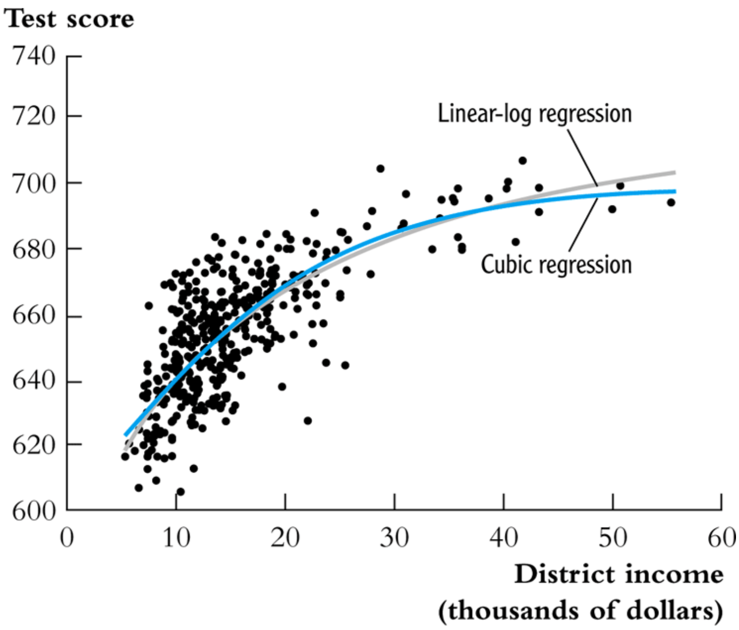


Figure 3: The linear-log and cubic regression function

If $\Delta X = 1$, then $\Delta Y/Y = \beta_1$. Expressed in percentage, we say that Y change by $100\beta_1\%$. With the derivative, $\beta_1 = \mathrm{d}\ln(Y)/\mathrm{d}X = (\mathrm{d}Y/Y)/X$.

e.g. In a regression of earnings on age, we have the estimated model as

$$\ln(\widehat{Earnings}) = 2.805 + 0.0087Age$$

So in this regression, earnings are predicted to increase by 0.87% for each additional year of age.

- **Case III: log-log model**

  In this case both X and Y are in logarithms.

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i \tag{9}$$

In the log-log model, 1% change in X is associated with a β_1% change in Y because

$$\frac{\Delta Y}{Y} \approx \ln(Y + \Delta Y) - \ln(Y) = \beta_1(\ln(X + \Delta X) - \ln(X)) \approx \beta_1 \frac{\Delta X}{X}$$

Thus, β_1 is the elasticity of Y with respect to X, that is

$$\beta_1 = \frac{100 \times (\Delta Y/Y)}{100 \times (\Delta X/X)} = \frac{\text{percentage change in } Y}{\text{percentage change in } X}.$$

With the derivative, $\beta_1 = \mathrm{d}\ln(Y)/\mathrm{d}\ln(X) = (\mathrm{d}Y/Y)/(\mathrm{d}X/X)$.

e.g. The log-log model of the test score application is estimated as

$$\ln(\widehat{TestScore}) = 6.336 + 0.0544 \ln(Income)$$

This implies that a 1% increase in income corresponds to a 0.0544% increase in test scores.

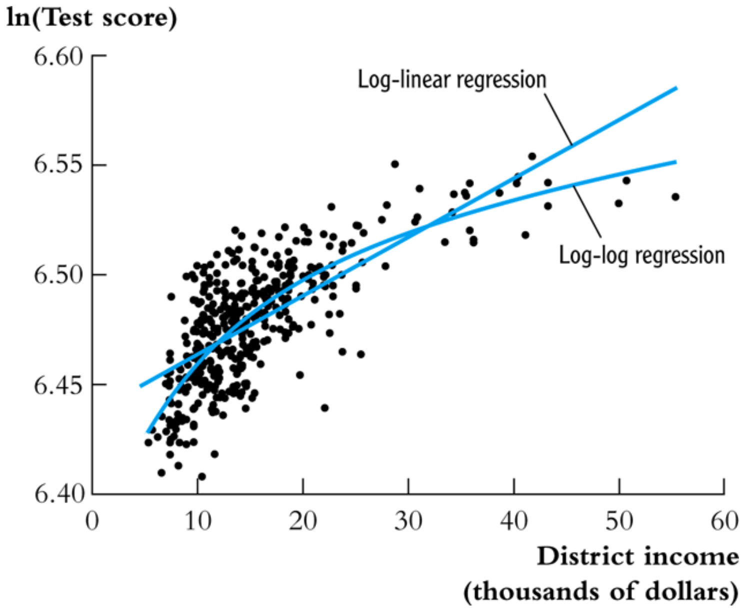


Figure 4: The log-linear and log-log regression functions

Table 1: Interpretation β_1 in three logarithmic specifications

Case	Regression specification	Interpretation of β_1
I	$Y = \beta_0 + \beta_1 \ln(X) + u$	A 1% change in X is associated with a change in Y of $0.01\beta_1$
II	$\ln(Y) = \beta_0 + \beta_1 X + u$	A change in X by one unit is associated with a $100\beta_1\%$ change in Y
III	$\ln(Y) = \beta_0 + \beta_1 \ln(X) + u$	A 1% change in X is associated with a $\beta_1\%$ change in Y, so β_1 is the elasticity of Y with respect to X

4 Interactions between independent variables

4.1 Interactions between two binary variables

The regression model with interaction between two binary variables

Consider the population regression of earnings (Y_i, where $Y_i = Earnings_i$) against two binary variables: a worker has a college degree (D_i, where $D_{1i} = 1$ if the i^{th} person graduated from college) and a worker's gender (D_{2i}, where $D_{2i} = 1$ if the i^{th} person is female).

Then the population regression model with an interaction term of two binary variables is

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3(D_{1i} \times D_{2i}) + u_i \tag{10}$$

in which $D_{1i} \times D_{2i}$ is the **interaction term**.

The method of interpreting coefficients in regressions with interacted binary variables

We follow a general rule for interpreting coefficients in Equation (10):

> First compute the expected values of Y for each possible case described by the set of binary variables. Next compare these expected values. Each coefficient can then be expressed either as an expected value or as the difference between two or more expected values.

- **Compute the expected values of Y for each possible combinations of D_1 and D_2**

  Case 1 $E(Y_i|D_{1i} = 0, D_{2i} = 0) = \beta_0$: the average income of male non-college graduates is β_0.

  Case 2 $E(Y_i|D_{1i} = 1, D_{2i} = 0) = \beta_0 + \beta_1$: the average income male college graduates is $\beta_0 + \beta_1$.

  Case 3 $E(Y_i|D_{1i} = 0, D_{2i} = 1) = \beta_0 + \beta_2$: the average income of female non-college graduates is $\beta_0 + \beta_2$.

  Case 4 $E(Y_i|D_{1i} = 1, D_{2i} = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$: the average income of female college graduates is $\beta_0 + \beta_1 + \beta_2 + \beta_3$.

- **Compute the difference between a pair of cases**

  Case 1 vs. Case 2 $E(Y_i|D_{1i} = 1, D_{2i} = 0) - E(Y_i|D_{1i} = 0, D_{2i} = 0) = \beta_1$. Thus, the average income difference between college graduates and non-college graduates among male workers is β_1.

  Case 1 vs. Case 3 $E(Y_i|D_{1i} = 0, D_{2i} = 1) - E(Y_i|D_{1i} = 0, D_{2i} = 0) = \beta_2$. Thus, the average income difference between female and male workers who are not college graduates is β_2.

  Case 1 vs. Case 4 $E(Y_i|D_{1i} = 1, D_{2i} = 1) - E(Y_i|D_{1i} = 0, D_{2i} = 0) = \beta_1 + \beta_2 + \beta_3$. Thus, The average income difference between female college graduates and male non-college graduates is $\beta_1 + \beta_2 + \beta_3$.

Case 2 vs. Case 3 $E(Y_i|D_{1i} = 0, D_{2i} = 1) - E(Y_i|D_{1i} = 1, D_{2i} = 0) = \beta_2 - \beta_1$. Thus, the average income difference between female non-college graduates and male college graduates is $\beta_2 - \beta_1$.

Case 2 vs. Case 4 $E(Y_i|D_{1i} = 1, D_{2i} = 1) - E(Y_i|D_{1i} = 1, D_{2i} = 0) = \beta_2 + \beta_3$. Thus, the average income difference between female college graduates and male college graduates is $\beta_2 + \beta_3$.

Case 3 vs. Case 4 $E(Y_i|D_{1i} = 1, D_{2i} = 1) - E(Y_i|D_{1i} = 0, D_{2i} = 1) = \beta_1 + \beta_3$. Thus, the average income difference between female college graduates and female non-college graduates is $\beta_1 + \beta_3$.

We can use t-statistic or F-statistic to test whether the differences between different cases are statistically significant. For example, if we want to test whether the average income of male college graduates differs from that of male non-college graduates, the null hypothesis is $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$. Then, we can use a t-statistic for this test. In another case, if we want to test whether the average income of female college graduates differs from that of female non-college graduates, the null hypothesis is $H_0 : \beta_1 + \beta_3 = 0$ vs. $H_1 : \beta_1 + \beta_3 \neq 0$. Then, we need to use an F-statistic for this test.

4.2 Interactions between a continuous and a binary variable

Consider the population regression of earnings (Y_i) against one continuous variable, individual's years of work experience (X_i), and one binary variable, whether the worker has a college degree (D_i, where $D_i = 1$ if the i^{th} person is a college graduate).

As shown in Figure 5, the population regression line relating Y and X depends on D in three different ways.

Different intercept, same slope.

As shown in Figure 5 (a), the corresponding population regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i \tag{11}$$

- From Equation (11), we have the population regression functions as $E(Y_i|D_i = 1) = \beta_0 + \beta_1 X_i + \beta_2$ and $E(Y_i|D_i = 0) = \beta_0 + \beta_1 X_i$. Thus, $E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \beta_2$.

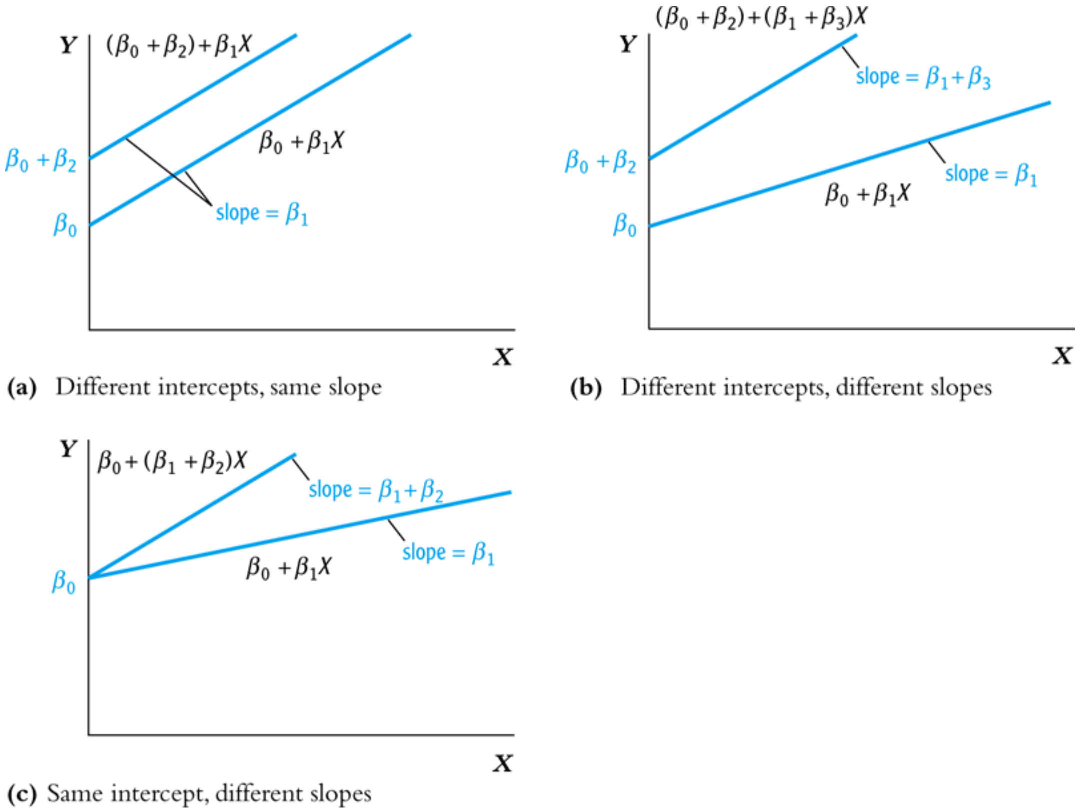


Figure 5: Regression Functions Using Binary and Continuous Variables

- The average initial salary of college graduates is higher than non-college graduates by β_2, and this gap persists at the same magnitude regardless of how many years a worker has been working.

Different intercepts and different slopes.

As shown in Figure 5 (b), to allow for different slopes, we need to add an interaction term to Equation (11) as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i \tag{12}$$

- The population regression functions for the two cases are $E(Y_i|D_i = 1) = (\beta_0+\beta_2)+(\beta_1 + \beta_3)X_i$ and $E(Y_i|D_i = 0) = \beta_0 + \beta_1 X_i$. Thus, β_2 is the difference in intercepts and β_3 is the difference in slopes.

- The average initial salary of college graduates is higher than non-college graduates by β_2, and this gap will widen (or narrow) depending on the effect of the years of work experience on earnings.

Different intercepts and same intercept.

As shown in Figure 5 (c), there is no difference in the intercept between the two regression line, but they have different slopes. The corresponding regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i \tag{13}$$

4.3 Interactions between two continuous variables

Now we consider the regression of earnings against two continuous variables, one for the years of work experience (X_1) and another for the years of schooling (X_2). The interaction term $X_{1i} \times X_{2i}$ can be included to account for (1) the effect of working experience on earnings, depending on the years of schooling, and (2) conversely, the effect of the years of schooling on earnings, depending on working experience.

The regression model with the interaction between X_1 and X_2 is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i \tag{14}$$

- The effect of a change in X_1, holding X_2 constant, is

  $$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2 \text{ for continuous variables}$$

  or generally,

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

- Similarly, the effect of a change in X_2, holding X_1 constant, is

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

5 Regression Functions That Are Nonlinear in the Parameters

All the regression models that we have discussed in this lecture are nonlinear in the regressors but linear in parameters so that we can still treat them as linear regression models and estimate using the OLS. However, there exist regression models that are nonlinear in parameters. For these models, we can either transform them to the "linear" type of models or estimate using the **nonlinear least squares** (NLS) estimators.

5.1 Transform a nonlinear model to a linear one

Suppose we have a nonlinear regression model as follows

$$Y_i = \alpha X_{1i}^{\beta_1} X_{2i}^{\beta_2} \cdots X_{ki}^{\beta_k} e^{u_i} \tag{15}$$

which is nonlinear in both X and β. The Cobb-Douglas utility (or production) function takes the form as in Equation (15).

Although Equation (15) is nonlinear in β, we can easily transform it to be linear in β by taking the natural logarithmic function on both sides of the equation, yielding the following equation:

$$\ln(Y_i) = \ln(\alpha) + \beta_1 \ln(X_{1i}) + \beta_2 \ln(X_{2i}) + \cdots + \beta_k \ln(X_{ki}) + u_i \tag{16}$$

Let $\beta_0 = \ln(\alpha)$. Equation (15) becomes a log-log regression model, which is linear in all parameters and can be estimated using the OLS. β_i for $i = 1, 2, \ldots, k$ are the elasticities of Y with respect to X_i.

5.2 Nonlinear models that cannot be linearized

There are nonlinear models that cannot be linearized by any transformation. We introduce two examples.

Logistic curve

Sometimes we may have a dependent variable taking values between 0 and 1, such as the fraction of students who get the test scores lower than 70. The linear regression model generally cannot guarantee the predicted dependent variable to be bounded between 0 and 1. For this reason, we can use the logistic function to set up a nonlinear regression function.

The logistic regression model with k regressors is

$$Y_i = \frac{1}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \cdots \beta_k X_{ki})} + u_i \tag{17}$$

The logistic function with a single X is graphed in Figure 6(a). The logistic function has an elongated "S" shape.

- For small values of X, the value of the function is nearly 0 and the shape is flat.
- For large values of X, the function approaches 1 and the slope is flat again.

Negative exponential growth function

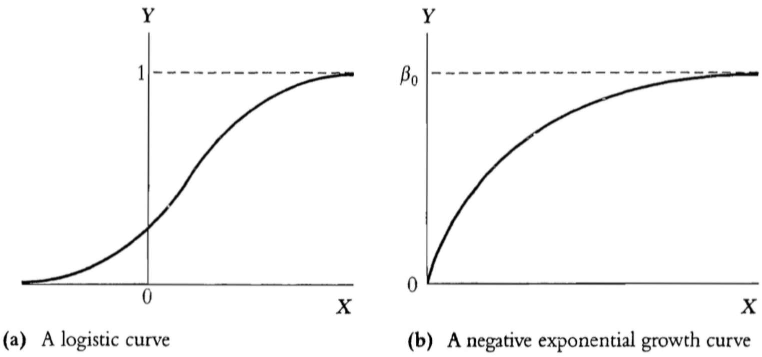
Sometimes the effect of X on Y must be positive and the effect is bounded by a upper bound. For this case, we can use a negative exponential growth function to set up a regression model as follows

$$Y_i = \beta_0[1 - \exp(-\beta_1(X_i - \beta_2))] + u_i \tag{18}$$

The negative exponential growth function is graphed in Figure 6(b), which has the desired properties:

- The slope is positive for all values of X.
- The slope is greatest at low values of X and decreases as X increases.
- There is an upper bound, that is, a limit of Y as X goes to infinity, β_0.

FIGURE 8.12 Two Functions That Are Nonlinear in Their Parameters



Part (a) plots the logistic function of Equation (8.38), which has predicted values that lie between 0 and 1. Part (b) plots the negative exponential growth function of Equation (8.39), which has a slope that is always positive and decreases as X increases, and an asymptote at β_0 as X tends to infinity.

Figure 6: The logistic and negative exponential growth functions

The nonlinear least squares estimators

For a nonlinear regression function

$$Y_i = f(X_1, \ldots, X_k; \beta_1, \ldots, \beta_m) + u_i$$

which is nonlinear in both X and β, we can obtain the estimated parameters by **nonlinear least squares** (NLS) estimation. The essential idea of NLS is the same as OLS, which is to minimize the sum of squared prediction mistakes. That is

$$\min_{b_1,\ldots,b_m} S(b_1,\ldots,b_m) = \sum_{i=1}^{n} [Y_i - f(X_1,\ldots,X_k; b_1,\ldots,b_m)]^2$$

The solution to this minimization problem is the nonlinear least squares estimators.

6 Nonlinear effects on test scores of the student-teacher ratio

We apply the nonlinear regression models to examine the effect of the student-teacher ratios on test scores in California elementary school districts.

Let's read replicate_ch8.pdf.

The goal is to reproduce Table 8.3 and Figures 8.10 and 8.11. in the textbook.