# Lecture 10: Nonlinear Regression Functions

Zheng Tian

# Outline

# Overview

### Linear population regression function

$E(Y_i \mid \mathbf{X}_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})'$.

### Nonlinear population regression function

$E(Y_i \mid \mathbf{X}_i) = f(X_{i1}, X_{i2}, \ldots, X_{ik}; \beta_1, \beta_2, \ldots, \beta_m)$, where $f(\cdot)$ is a nonlinear function.

### Study questions

- Why do we need to use nonlinear regression models?
- What types of nonlinear regression models can we estimate by OLS?
- How can we interpret the coefficients in nonlinear regression models?

# Test Scores and district income

- Test scores can be determined by average district income
- We estimate a simple linear regression model

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

- What's the problem with the simple linear regression model?

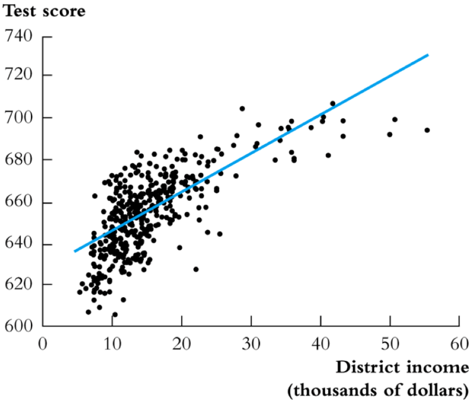


Figure: Scatterplot of test score vs district income and a linear regression line

# Why does a simple linear regression model not fit the data well?

- Data points are below the OLS 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

$$\Delta TestScore / \Delta Income = \beta_1$$

  where $\beta_1$ is constant.

# Estimate a quadratic regression model

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

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

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

- It is linear with respect to $\beta$. So we can still use the OLS estimation and carry out hypothesis testing as we do with a linear regression model.

# Estimate a quadratic regression model (cont'd)

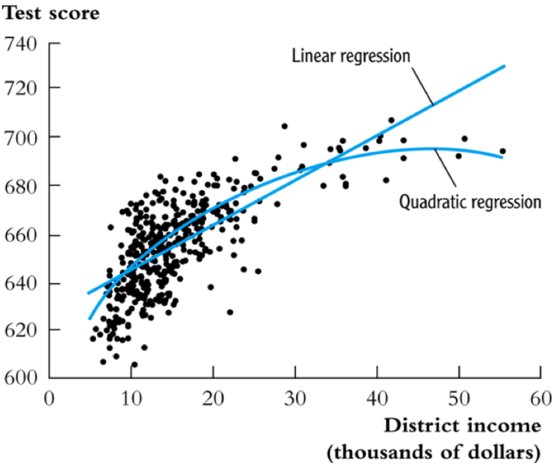


Figure: Scatterplot of test score vs district income and a quadratic regression line

# A general formula for a nonlinear population regression function

A general nonlinear regression model is

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

- The population nonlinear regression function:

$$E(Y_i | X_{i1}, \ldots, X_{ik}) = f(X_{i1}, X_{2i}, \ldots, X_{ik}; \beta_1, \beta_2, \ldots, \beta_m)$$

- The number of regressors and the number of parameters are not necessarily equal in the nonlinear regression model.
- In vector notation

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

- We focus on the nonlinear regression models such that $f(\cdot)$ is nonlinear with $\mathbf{X}_i$ but linear with $\boldsymbol{\beta}$.

# 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

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

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

$$\Delta Y = \frac{\partial f(X_1, \ldots, X_k; \beta)}{\partial X_1} \Delta X_1$$

# Application to test scores and income \ Estimation and inference

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.

# Application to test scores and income \ The effect of change in 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

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

- 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.
- 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.

# 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 $\beta$.
3. Determine whether the nonlinear model improves upon 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 $X$.