Lecture 10: Regression for Nonlinear Relationships

Xiaozhou Ding

University of Kentucky

April 16, 2019

- The models we leant before all assume there is a linear relationship between x and y.
- e.g. wage and education; wage and experience; Keenland attendance and temperature; food consumption and income, etc.
- But really? Do you really believe their relationship can be represented by a straight line?

Why Do We Need Nonlinear Model?

- Theory predicts nonlinear relationship
  - Optimal solution.
    For example, the “golden rate” saving rate; the optimal hours of study time every week; the optimal tax rate; etc.
  - Changing marginal effect.
    For example, the return to education may increase with year of schooling; productivity and working experience; utility you get from the apple and the number of apple you eat; etc.

Polynomial Regression Models

- A simple linear regression model,

$$y = \beta_0 + \beta_1 x + \epsilon$$

  is easy to interpret: if x increases by one unit, we expect y to change by β_1, holding other variables constant.
- However, sometimes the relationship cannot be represented by a straight line and, rather, must be captured by an appropriate curve.
- Since one of the assumptions in Chapter 15 replaces the restriction of linearity on the parameters, not the x values, we can capture many interesting nonlinear relationships within this framework.

The Quadratic Regression Model

- For example, a firm's average cost curve tends to be "U-shaped".
- Due to economies of scale, average cost initially falls as output increases, before rising once output reaches a certain threshold.
- Such a relationship can be estimated by a quadratic regression model:

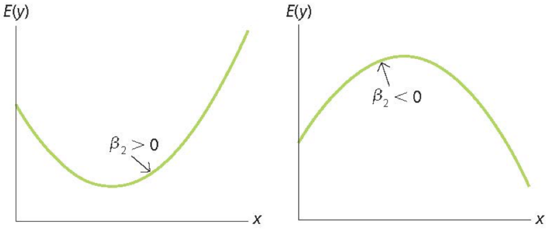
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

The “Flexible” Quadratic Model

- For a quadratic regression, we estimate:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

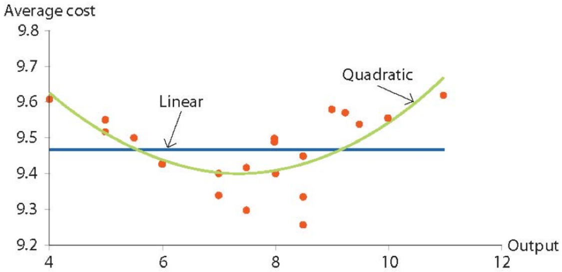
- The sign of β_2 determines the shape:



- With quadratic regression model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$,
- The marginal effect of x on y is $\beta_1 + 2\beta_2 x$. The marginal effect is NOT a constant, but a function of x.
- Predictions with this model are made by $\hat{y} = b_0 + b_1 x + b_2 x^2$.
- When $x = -\frac{b_1}{2b_2}$, $\hat{y} = \{max, min\}$ values. $\hat{y}$ reaches its maximum ($b_2 < 0$) or minimum ($b_2 > 0$) when the marginal effect =0.

Example

- Suppose we want to estimate the relationship between average cost and output. We gather data for 20 manufacturing firms on output and average cost.
- When using a scatterplot to display the relationship, notice that a quadratic curve seems to better fit the data.



The model is

$$\text{average cost} = \beta_0 + \beta_1 \text{output} + \beta_2 \text{output}^2 + \epsilon$$

Results

$$\widehat{\text{average cost}} = 10.5225 - .3073\text{output} + 0.210\text{output}^2$$

- Is the average cost curve concave or convex? Explain how you know.
- Find the output that maximizes/minimizes the average cost. (Hint: first order condition).

$$-.3073 + 2 \times .0210\text{output} = 0$$

$$\text{output} = 7.32$$

Prediction

- What is the change in average cost going from an output level of 4 million units to 5 million units?

$$\widehat{AC} = 10.5225 - 0.3073 \times 4 + 0.0210 \times 4^2 = 9.63$$

$$\widehat{AC} = 10.5225 - 0.3073 \times 5 + 0.0210 \times 5^2 = 9.51$$

  An increase in output from 4 to 5 million units(one unit increase in x) results in a \$0.12 decrease in predicted average cost.

- What is the change in average cost going from an output level of 8 million units to 9 million units? Compare this result to the result found in part 1.

$$\widehat{AC} = 10.5225 - 0.3073 \times 8 + 0.0210 \times 8^2 = 9.41$$

$$\widehat{AC} = 10.5225 - 0.3073 \times 9 + 0.0210 \times 9^2 = 9.46$$

  An increase in output from 8 to 9 million units(one unit increase in x) results in a \$0.05 increase in predicted average cost.

Depending on the value at which x is evaluated, a one-unit change in x may have positive or negative influence on y, and the magnitude of this effect is not constant.

Higher Order Models

- The quadratic regression model allows one sign change of the slope capturing the influence of x on y.
- Polynomial regression models, more generally, are able to describe various numbers of sign changes.
- For example, the cubic regression model allows for two changes to the slope:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

The n-th order polynomial regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \ldots + \beta_n x_1^n + \epsilon$$

It allows $n - 1$ signs changes of the slope.

Regression Models with Logarithms

- Another commonly used transformation to capture nonlinearities between the response and the explanatory variables is based on the natural logarithm.
- Linearity assumes that an increase of one unit in the explanatory variable has the same impact on the response variable regardless of whether x is increasing from 100 to 101 or 1000 to 1001.
- That may not be true if, for example, we want to predict how food expenditure responds to changes in income.

The Log-Log Model

- In a log-log model both the response and the explanatory variables are transformed into natural logs. We can write this model as:

$$\ln(y) = \beta_0 + \beta_1 \ln(X) + \epsilon$$

- The relationship between y and x is captured by a curve whose shape depends on β_1.

Notice, β_1 **is the marginal effect. It denote the percentage change of** y **if** x **increases by one percentage.**

The Slope as an Elasticity

- In the model $\ln y = \beta_0 + \beta_1 x + \epsilon$, we would interpret the slope as the percent change in y given a 1% increase in x. In other words, β_1 is a measure of elasticity.
- Suppose y represents quantity demanded and x is price. If $\beta_1 = -1.2$, it would imply that a 1% increase in price is expected to lead to a 1.2% decrease in its quantity demanded.

Prediction

- Even though we estimate the equation with transformed data, it is relatively easy to predict in the original units.
- After the logarithm are computed, the equation is estimated as:

$$\widehat{\ln y} = b_0 + b_1 \ln x$$

- But $\hat{y} = \exp(b_0 + b_1 \ln x)$ is known to systematically underestimate the expected value of y, so we correct for that by making predictions using:

$$\hat{y} = \exp(b_0 + b_1 \ln x + se^2/2)$$

  where se is the standard error of the estimate.

Example

Refer back to the expenditure example where y is expenditure on food and x represents income. Let the sample regression be

$$\widehat{\ln y} = 3.64 + 0.5 \ln x$$

with the standard error of the estimate $se = 0.18$.

1. What is the predicted food expenditure for an individual whose income is \$20,000?
2. What is the predicted value if income increases to \$21,000?
3. Interpret the slope coefficient, $b_1 = 0.5$.

1. For the log-log model, $\hat{y} = \exp(b_0 + b_1 \ln x + se^2/2)$. If income equals 20,000, $\hat{y} = \exp(3.64 + 0.5 \ln 20000 + \frac{0.18^2}{2}) = 5475$.
2. If income equals 21,000, $\hat{y} = \exp(3.64 + 0.5 \ln 21000 + \frac{0.18^2}{2}) = 5610$.
3. The slope means as x increases by 1%, y increases by 0.5%. As shown in the first two parts, income increases 5% from 20,000 to 21,000, and expenditure on food increases by 2.47% from 5475 to 5610, roughly by 2.5%.

Semi-Log Model

- Another common application is a “semi-log” model where only one of the variable is transformed.
- In a logarithmic model only the x is expressed as a natural log:

$$y = \beta_0 + \beta_1 \ln x + \epsilon$$

y is the original units of measurement. x measurement unit now is the percentage.

β_1 is still the marginal effect. $\beta_1/100$ measures the unit change of y when x increases by 1 **percent**.
Prediction model is: $\hat{y} = b_0 + b_1 \ln x$.

Example

Continuing with the earlier example of food expenditure. Let the estimated logarithmic regression be:

$$\widehat{Food} = 12 + 566\ln(Income)$$

- For an income of $20,000, predicted food expenditure is:

$$\widehat{Food} = 12 + 566\ln(20,000) = 5617$$

- The slope $b_1 = 566$ implies that a 1 percent increase in income leads to an increase in food expenditures of $566/100 = 5.66$.

The Exponential Model

- When the y variable is transformed, but not the x, we have the exponential model:

$$\ln y = \beta_0 + \beta_1 x + \epsilon$$

- This model allows us to estimate the percent change in y when x increases by one **unit**.
- The sign of β_1 again determines the shape.
- The prediction model is

$$\widehat{\ln y} = b_0 + b_1 x$$

  $\hat{y} = \exp(b_0 + b_1 x)$ systematically underestimate the expected value of y. We use

$$\hat{y} = \exp(b_0 + b_1 x + se^2/2)$$

  as the prediction model, where *se* is the standard error of the estimate.

Example

Suppose we estimate the food expenditure-income relationship using an exponential model and find that the estimated exponential model is:

$$\ln \widehat{Food} = 7.6 + 0.00005\text{Income}$$

where the standard error of the estimate is $se = 0.2$.

- An individual with an income of \$20,000 is predicted to have food expenditure of:

$$\widehat{Food} = \exp(7.6 + 0.00005 \times 20000 + 0.2^2/2) = 5541$$

- The slope coefficient of 0.00005 implies that if income increases by \$1, food expenditure would increase by $0.00005 \times 100 = 0.005$ percent.

- The following table summarizes the simple linear and the logarithmic regression models:

Model	Predicted Value	Estimated Slope Coefficient
$y = \beta_0 + \beta_1 x + \epsilon$	$\hat{y} = b_0 + b_1 x$	change in $\hat{y}$ when $x \uparrow$ by 1 unit
$\ln y = \beta_0 + \beta_1 \ln x + \epsilon$	$\hat{y} = \exp(b_0 + b_1 \ln x + se^2/2)$	percentage change in $\hat{y}$ when $x \uparrow$ by 1%.
$y = \beta_0 + \beta_1 \ln x + \epsilon$	$\hat{y} = b_0 + b_1 \ln x$	$\frac{b_1}{100}$ change in $\hat{y}$ when $x \uparrow$ by 1%.
$\ln y = \beta_0 + \beta_1 x + \epsilon$	$\hat{y} = \exp(b_0 + b_1 x + se^2/2)$	$100b_1$ percentage change in $\hat{y}$ when $x \uparrow$ by 1

Although these models involve nonlinear functions of the two variables y and x, they are linear in β parameters so can be estimated by OLS.

Compare Linear Models with Models with Logarithms

- For logarithmic model, we can still compare them with linear models using R^2.
- For log-log model, and exponential model, we cannot compare them with linear models using R^2 directly. Because they have different dependent variables.
- For a valid comparison, we need to compute the percentage of explained variations of y even through the estimated model use $\ln(y)$ as the response variable.
- The coefficient of determination R^2 can be computed as $R^2 = r^2_{\hat{y},y}$, where $r_{\hat{y},y}$ is the sample correlation coefficient between y and $\hat{y}$.

Practice Examples

- Consider the following models:
  1. $\hat{y} = 200 - 12x$
  2. $\hat{y} = 19 - 350 \ln x$
  3. $\widehat{\ln y} = 3 + .1x, se = .5$
  4. $\widehat{\ln y} = 9 - .4 \ln x, se = .1$

  Answer the following questions for each:
  1. Interpret the slope coefficient for each of the estimated models
  2. For each model, what is the predicted unit change in y when x increases by 100 to 101, or 1%.

Summary

- Polinomial regression models
  - Functional form
  - Graphic representation
  - Model estimation and prediction
- Logarithms regression models
  - Log-log model
  - log-linear model
  - Linear-log model
  - Interpretation of coefficients and model predictions.