Lecture 4

Multiple Regression Analysis: Further Issues

Nurgul Tilenbaeva

American University - Central Asia

17.10.2022

Contents

More on Functional Form

Logarithmic Functional Forms

$log(price) = \beta_0 + \beta_1 log(nox) + \beta_2 rooms + u,$

- β_1 is the elasticity of *price* with respect to *nox*(pollution).
- β_2 is the change in $log(price)$, when $\Delta rooms = 1$. When multiplied by 100, this is the approximate percentage change in *price*. $100\beta_2$ is the semi-elasticity of *price* with respect to *rooms*.

More on Functional Form

Logarithmic Functional Forms

$$\widehat{log(price)} = \underset{(0.19)}{9.23} - \underset{(0.066)}{0.718}\, log(nox) + \underset{(0.019)}{0.306}\, rooms$$

$n = 506, R^2 = 0.514$

- When *nox* increases by 1%, *price* falls by 0.718%, holding *rooms* fixed.
- When *rooms* increases by one, *price* increases by approximately $100(0.306) = 30.6\%$. This estimate turns out to be somewhat inaccurate. The approximation error occurs because, as the change in $log(y)$ becomes larger and larger, the approximation $\%\Delta y \approx 100\Delta log(y)$ becomes more and more inaccurate.

More on Functional Form

Logarithmic Functional Forms

$\widehat{log}(y) = \hat{\beta}_0 + \hat{\beta}_1 log(x_1) + \hat{\beta}_2 x_2$

Using simple algebraic properties of the exponential and logarithmic functions gives the **exact** percentage change in the predicted y as

$$\boxed{\%\Delta\hat{y} = 100[exp(\hat{\beta}_2 \Delta x_2) - 1]}$$

where the multiplication by 100 turns the proportionate change into a percentage change.

When $\Delta x_2 = 1$,

$$\boxed{\%\Delta\hat{y} = 100[exp(\hat{\beta}_2) - 1]}$$

More on Functional Form

Logarithmic Functional Forms

Why use logarithms?

- Coefficients with appealing interpretations.
- We can be ignorant about the units of measurement of variables appearing in logarithmic form because the slope coefficients are invariant to rescalings.
- When $y > 0$, models using $log(y)$ as the dependent variable often satisfy the CLM assumptions more closely than models using the level of y. Strictly positive variables often have conditional distributions that are heteroskedastic or skewed; taking the *log* can mitigate, if not eliminate, both problems.
- Taking the *log* of a variable narrows its range. This can make OLS estimates less sensitive to outlying (or extreme values).

More on Functional Form

Logarithmic Functional Forms

When not to use logarithms?

- When a variable y is between zero and one (such as proportion) and takes on values close to zero. In this case, $log(y)$ can be very large in magnitude whereas the original variable, y, is bounded between zero and one.
- If a variable takes on zero or negative values.

More on Functional Form

Logarithmic Functional Forms

Rules of thumb for taking logs

- When a variable is a positive dollar amount, the *log* is often taken.
  Examples: wages, salaries, firm sales, and firm market value.
- Variables with large integer values often appear in *log* form.
  Examples: population, number of employees, and school enrollment.
- Variables that are measured in years usually appear in their original form.
  Examples: education, experience, tenure, age.
- A variable that is a proportion or a percent can appear in either original or logarithmic form, although there is a tendency to use it in level form.
  Examples: unemployment rate, participation rate in a pension plan, percentage of students passing an exam, arrest rate on reported crimes.

More on Functional Form

Models with Quadratics

Quadratic functions are used quite often in applied economics to capture decreasing or increasing marginal effects.

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

We write the estimated equation as

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$

Then,

$\Delta\hat{y} \approx (\hat{\beta}_1 + 2\hat{\beta}_2 x)\Delta x$, so $\Delta\hat{y}/\Delta x \approx \hat{\beta}_1 + 2\hat{\beta}_2 x$

More on Functional Form

Models with Quadratics

$$\widehat{wage} = \underset{(0.35)}{3.73} + \underset{(0.041)}{0.298}\, exper - \underset{(0.0009)}{0.0061}\, exper^2$$

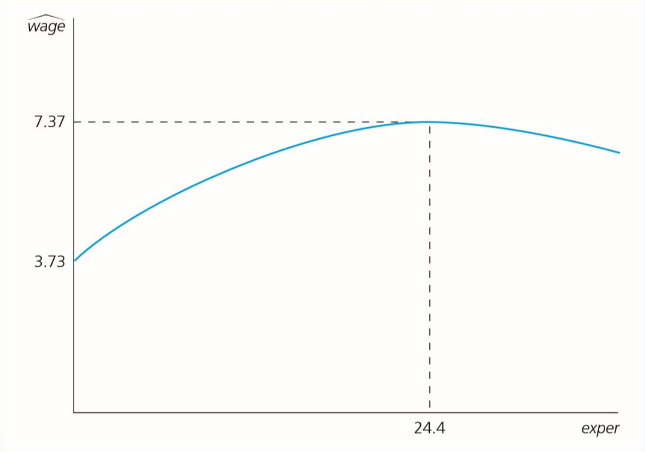
$n = 526, R^2 = 0.093$

exper has a **diminishing** effect on *wage*. The first year of experience is worth 29.8 cents per hour. The second year of experience is worth less [about $0.298 - 2(0.0061)(1) \approx 0.286$, or 28.6 cents per hour. In going from 10 to 11 years of experience, *wage* is predicted to increase by about $0.298 - 2(0.0061)(10) \approx 0.176$, or 17.6 cents per hour.

More on Functional Form

Models with Quadratics

FIGURE 6.1 Quadratic relationship between $\widehat{wage}$ and *exper*.



More on Functional Form

Models with Quadratics

- A U-shape arises when $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive; this captures and **increasing** effect of x on y.
- If the coefficients on the level and squared terms have the *same* sign (either both positive or both negative) and the explanatory variable is necessarily nonnegative, there is no turning point for $x > 0$.
  - If β_1 and β_2 are both positive, the smallest expected value of y is at $x = 0$, and increases in x always have a positive and increasing effect on y.
  - If β_1 and β_2 are both negative, the largest expected value of y is at $x = 0$, and increases in x have a negative effect on y, with the magnitude of the effect increasing as x gets larger.

More on Functional Form

Models with Interaction Terms

$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrooms + \beta_3 sqrft * bdrooms + \beta_4 bthrooms + u$

the partial effect of *bdrooms* on *price* (holding all other variables fixed) is

$\frac{\Delta price}{\Delta bdrooms} = \beta_2 + \beta_3 sqrft$

If $\beta_3 > 0$, then an additional bedroom yields a higher increase in housing price for larger houses. In other words, there is an **interaction effect** between square footage and number of bedrooms.

More on Goodness-of-Fit and Selection of Regressors

Adjusted R-Squared

The R-squared can be written as:

$R^2 = 1 - (SSR/n)/(SST/n)$

The adjusted R-squared is:

$$\boxed{\bar{R}^2 = 1 - [SSR/(n-k-1)]/[SST/(n-1)]}$$

$$\boxed{\bar{R}^2 = 1 - (1 - R^2)(n-1)/(n-k-1)}$$

The primary attractiveness of $\bar{R}^2$ is that it imposes a penalty for adding additional independent variables to a model.

More on Goodness-of-Fit and Selection of Regressors

Using Adjusted R-Squared to Choose between Nonnested Models

We want to choose between the models

$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + u$
and
$log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 rbisyr + u$

These two equations are **nonnested models** because neither equation is a special case of the other.

$\bar{R}^2$ for the regression containing *hrunsyr* is 0.6211.
$\bar{R}^2$ for the regression containing *rbisyr* is 0.6226.

Thus, based on the adjusted R-squared, there is a very slight preference for the model with *rbisyr*.

More on Goodness-of-Fit and Selection of Regressors

Using Adjusted R-Squared to Choose between Nonnested Models

Comparing $\bar{R}^2$ to choose among different nonnested sets of independent variables can be valuable when these variables represent different functional forms. Consider two models relating R and D intensity to firm sales:

$rdintens = \beta_0 + \beta_1 log(sales) + u$
$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u$

R^2 is 0.061 for the first model.
R^2 is 0.148 for the second model.

$\bar{R}^2$ is 0.030 for the first model.
$\bar{R}^2$ is 0.090 for the second model.

Thus, even after adjusting for the difference in degrees of freedom, the quadratic model wins out.
BUT! We cannot use $\bar{R}^2$ to choose between different functional forms for the dependent variable.