Econometrics Lab 2

Estimation and Diagnostics of Linear Regression

1. Income Determination. The yearly income of a particular person may be quite random. Regardless of his education background and his experience, he may luckily win a fortune in a lottery and just likely he may lose his job and bring home nothing. But the income of an "average" person relates in a determined manner with his education, experience, gender, and other factors. The study of a particular person is fortune telling and the study of an "average" person is social science, economics in particular.

This exercise uses data from the Chinese Social Survey Open Database, China General Social Survey (CGSS) 2005 (cgss05.csv). There 5778 observations randomly taken on households across China.

At this stage, we are interested in the following three variables in the dataset:

income	yearly income
edu	years of schooling
expr	number of years after graduation

We study how much an "average" Chinese worker earns, given his education and work experience. The model can be constructed as follows,

$$\log(income) = \beta_0 + \beta_1 edu + \beta_2 expr + u. \tag{1}$$

(1) What would be the economic meanings of β_0, β_1, and β_2?

(2) Estimate the regression in (1). Does the signs of the estimates make sense? Discuss your results.

(3) In average, how much would an individual with 10 years of schooling and 5 years of work experience make in 2005 China? (Tip: If $\log(y) \sim N(\mu, \sigma^2)$, then $\mathbb{E}(y) = \exp(\mu + \sigma^2/2)$.)

(4) Calculate SSE, SST, SSR, and R^2 and $\bar{R}^2$. Add a quadratic term $expr^2$ to the regression, re-calculate the R^2 and $\bar{R}^2$.

2. Cost Function This exercise examines the production cost of 145 electricity generating companies in the United States. We use the dataset nerlove1.csv. It contains two variables *costs* and *kwh*, representing total costs and the quantity of electricity generated.

(1) Draw the scatter plot of *costs* v.s. *kwh*. What do you learn from the diagram?

(2) Define a new variable $avc = costs/kwh \times 1000$, where 1000 is a scaling factor. *avc* measures the unit cost of generating 1 kwh electricity. Draw the scatter plot of *avc* v.s. *kwh* and $\log(avc)$ v.s. *kwh*. What do you learn from these two diagrams.

(3) Estimate the following regression,

$$costs = \beta_0 + \beta_1 kwh + u. \tag{2}$$

(4) Calculate SSE, SST, SSR, and R^2.

(5) Plot residuals itself, residuals v.s. *kwh*, residuals v.s. fitted value.

(6) Draw partial residual plot. That is, $\hat{u} + \hat{\beta}_1 kwh$ v.s. *kwh*.

(7) From these diagrams, do you detect any violations of the CLR assumptions?

(8) To remedy the problems, we run

$$costs = \beta_0 + \beta_1 kwh + \beta_2 kwh^2 + u. \tag{3}$$

(9) Perform the above model diagnostics on the new model. Do you see any improvement?

3. Engel's Law. We use the dataset engel.dat. The data description can be found in the Lab 2 assignment.

(1) In addition to income per capita, we may conjecture that the consumption of food depends on the average age of the adult household members. To verify this conjecture, let's

estimate the following model,

$$food/totcper = \beta_0 + \beta_1 age + \beta_1 \log(totcinc) + u. \tag{4}$$

(2) What does the model tell us about the age effect on food consumption?

(3) Perform the usual model diagnostics. Report anything that catch your eyes.

(4) If there is anything wrong, try to improve your model in (4). And remember to perform diagnostics on your new model.