Econometrics

COURSEWORK 1

Module Code: ECON22100

Level: 2

Deadline for submitting: 16th April 2010

1. (12.5%) Explain in detail the difference between the following concepts.

a) Population Regression Function vs Sample Regression F unction.

The Population Regression Function (PRF) is a description of the model that is thought to be generating the actual data and it represents the true relationship between the variables. The PRF embodies the true values of β_1 and β_2, and is expressed as:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where Y_i is the actual value obtained by adding the error term u_i; or as

$$E(Y) = \beta_1 + \beta_2 X$$

where E(Y) may be regarded as the average or expected value of Y for a given value of X. The PRF tells us how the mean of Y varies for different values of X.

The population is the total collection of all objects to be studied. The population may be either finite or infinite, while a sample is a selection of just some items from the population. In general either all of the observations for the entire population will not be available, or they may be so many in number that it is infeasible to work with them, in which case a sample of data is taken for analysis. (Brooks, 2002, p g.112)

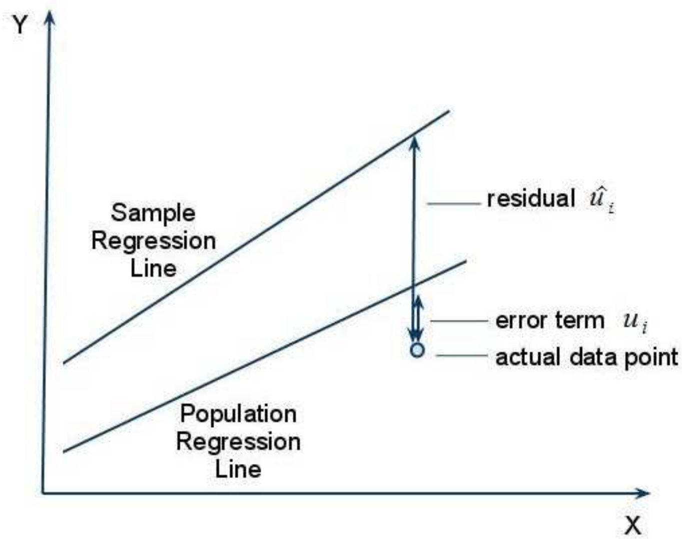
The Sample Regression Function (SRF)

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$$

It allows us to calculate the estimated value of Y for a given value of X.

b) Error terms vs residuals.

The error term – u_i is the disturbance term of the PRF. It represents factors other than X that affect Y and is calculated as the difference between the actual data point and the



estimated value of Y from the PRF.

The residual is the difference between the actual value of Y and it s fitted value:

$$\hat{u}_i = Y_i - \hat{Y}_i$$

In most cases we cannot calculate the parameters for PRF, therefore we cannot calculate the error term, which is why the residual is used more often in empirical studies.

c) Regression coefficients vs estimators.

An estimator is also known as a statistic, is simply a rule or formula or method that tells how to estimate the population parameter from the information provided by the sample at hand for example $\hat{\beta}_2$ is an estimator of β_2 , $\hat{Y}_i$ is an estimator of $E(Y \mid X_i)$ (Gujarati, 2003, pg.49)

β_1 , β_2 , $\hat{\beta}_1$ and $\hat{\beta}_2$ are regression coefficients; also know as intercept and slope coefficients.

Estimators are like proxies of the real values, as opposed to coefficients which are the parameters of an equation.

2. (12.5%) Consider the following model:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i} X_{3i} + \beta_5 X_i + u_i$$

where Y = the annual salary of a college teacher; X = years of teaching experience;

D_2 = 1 if male and 0 otherwise; D_3 = 1 if white and 0 otherwise.

a) What does the term ($D_{2i}D_{3i}$) mean?

Using a new variable $D_{2i}D_{3i}$ the interaction effect is taken into account, it shows the multiplicative effect of the variables D_{2i} and D_{3i} on mean Y.

b) What is the meaning of β_4?

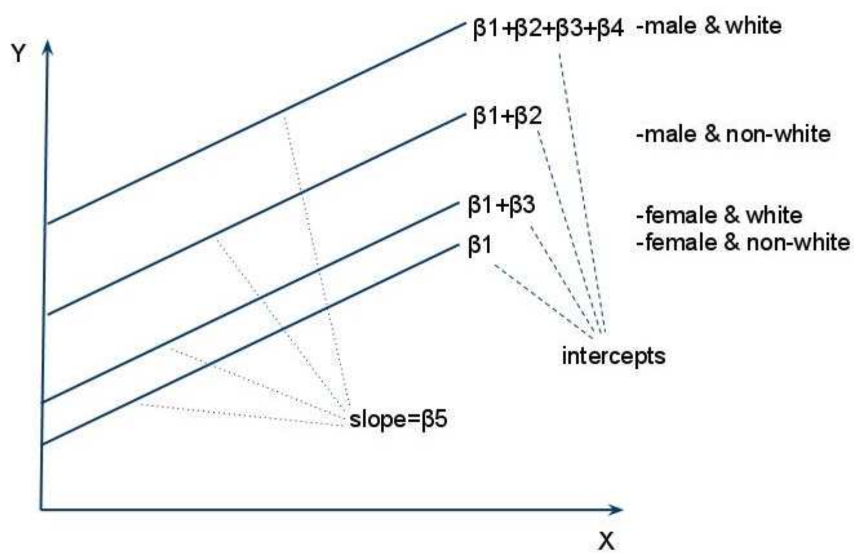
β_4 is the parameter that shows the extent of the expected change of the annual salary of college teachers if they are white males.

c) Find $E(Y_{ij}D_2 = 1; D_3 = 1; X_i)$ and interpret it.

$$E(Y_i|D_2=1, D_3=1,X_i)=\beta_1+\beta_2(1)+\beta_3(1)+\beta_4(1*1)+\beta_5X_i=\beta_1+\beta_2+\beta_3+\beta_4+\beta_5X_i$$

($\beta_1+\beta_2+\beta_3+\beta_4+\beta_5X_i$) is the expected annual salary of white male college teacher (i) with X_i years of teaching experience, which is different by ($\beta_2+\beta_3+\beta_4$) on average than the annual salary of non-white female college teacher with the same number of years of teaching experience, representing the reference salary.

The following diagram illustrates a possible combination of parameters, for simplicity we assume that all parameters are positive:



3. (20%) Consider the following regression

$\hat{Y}_i = \quad 50 \quad -0.1\, X_i$

se= (10.7509)

$|t| = \qquad 18.73$

$r^2=0.935 \qquad n=17$

Fill in the missing numbers and establish a 95% confidence interval for β_2. Give a proper economic interpretation. Would you reject the null hypothesis that the true β_2 is zero at α= 5%? Tell whether you are using a one-tailed or two-tailed test and why. (Critical values $t_{15}^{0.025}=2.131$ and $t_{15}^{0.05}=1.753$)

Let the intercept (50) be β_1 and the slope (-0.1) be β_2.

To find the t-value for β_1 the following formula will be used:

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$$

Assuming β_2 to be 0 we obtain:

$$t = \frac{50 - 0}{10.7509} = 4.6508$$

For finding the standard error of β_2 we rearrange the formula; to keep the standard error non-negative the assumption that t-value for β_2 is negative is made.

$$se(\hat{\beta}_2) = \frac{\hat{\beta}_2 - \beta_2}{t} = \frac{-0.1 - 0}{-18.73} = 0.0053$$

In order to establish a 95% confidence interval we need to use the following formula:

$$\hat{\beta}_2 - t_{n-2}^{\alpha/2}\, se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{n-2}^{\alpha/2}\, se(\hat{\beta}_2)$$

where α=5%; n-2=15; $\hat{\beta}_2 = -0.1$; $se(\hat{\beta}_2) = 0.0053$; $t_{15}^{0.025} = 2.131$

replacing these values in the above formula we obtain

$$-0.1 - 2.131 \cdot 0.0053 \le \beta_2 \le -0.1 + 2.131 \cdot 0.0053$$

$$-0.1114 \le \beta_2 \le -0.0886$$

We can state that there is 95% chance that the value of β_2 is in this interval [-0.1114; -0.0886]. In other words in 95 cases out of 100 the true β_2 value will fall within the interval.

Null Hypothesis $\qquad H_0$: $\beta_2=0$

Alternative hypothesis H_1: $\beta_2 \neq 0$

The two-tailed test is going to be used in this situation because the alternative hypothesis (H_1) has two possibilities $\beta_2<0$ or $\beta_2>0$.

Now we need to compare the absolute value of the t -statistic |t|=18.73 with the critical value $t_{15}^{0.025} = 2.131$.

$|t| \quad t_{15}^{0.025}$

18.73>2.131 => We reject the null hypothesis, which means that β_2 is statistically significant.

4. (25%) Based on 20 annual observations, the following regressions were obtained

Model A: $\hat{Y}_t = 4.00 \quad -0.5\, X_t \qquad r^2=0.70$

$se = (0.1216) \quad (0.1140)$

Model B: $\ln \hat{Y}_t = 1.78 \quad -0.37 \ln X_t \quad r^2 = 0.7448$

$se = (0.0152) \quad (0.0494)$

where Y= the cups of tea consumed per person per day and X= the price of tea in pounds per kilo.

a) Interpret the slope coefficients in the two models.

The slope in model A is -0.5, this means that when the price of tea goes up by £1 per kilo the consumption of tea will go down by 0.5 cups per person per day. In model B, because it's logarithmic the meaning of slope is different: the slope is -0.37, meaning that an increase in price of tea by 1% will result in a decrease by 0.37% in the number of cups of tea consumed per person per day.

b) You are told that $\bar{Y} = 5$ *and* $\bar{X} = 3.6$*. At these mean values, estimate the price elasticity for model A.*

To calculate the price elasticity we need to observe the relative change in Y over the relative change in X, as the slope represents the change in Y $\Delta Y = -0.5$ when X changes by one unit $\Delta X = 1$ we use this formula for price elasticity:

$$\varepsilon = \frac{\Delta Y / Y}{\Delta X / X} = \frac{\Delta Y \cdot X}{\Delta X \cdot Y} = \frac{-0.5 \cdot 3.6}{1 \cdot 5} = -0.36$$

c) What is the price elasticity for model B?

For model B price elasticity is equal to the slope because the logarithms are used for both X and Y. Price elasticity is -0.37.

d) From the estimated elasticities, can you say that the demand for tea is price inelastic?

Yes, the demand for tea is price inelastic because the price elasticities in both cases are less than 1.

e) Since the r^2 of Model B is larger than that of Model A, Model B is preferable to Model A. Comment on this statement.

Despite the fact that r^2, the measure of goodness of fit of a regression, is different in the two models we cannot state that one is preferable to the other because the controlled variables are different in models A and B, therefore we cannot compare the r^2 values directly.

5. (30%) You are given the following data based on 20 pairs of observations on Y and X.

$se(\hat{\beta}_2) = 0.05$; $\sum Y_i = 1100$; $\sum X_i = 1500$; $\sum X_i Y_i = 200{,}000$;

$\sum X_i^2 = 300{,}000$; $\sum Y_i^2 = 145{,}000$; $n = 20$.

Assuming all the assumptions of CLRM are fulfilled, obtain

a) $\hat{\beta}$

Based on the assumptions of the Classical Linear Regression Model we can write the general form of the function:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

For calculating the regression coefficients we use:

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{20 \cdot 200{,}000 - 1500 \cdot 1100}{20 \cdot 300{,}000 - 1500^2} = 0.6267$$

Mean of Y is:

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{1100}{20} = 55$$

Mean of X is:

$$\bar{X} = \frac{\sum X_i}{n} = \frac{1500}{20} = 75$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 55 - 0.6266667 \cdot 75 = 8$$

b) standard error of β_1

$se(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n\sum x_i^2}}\,\sigma$ (5.b.1)(Gujarati, pg. 58)

$$se(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$$

$$\Rightarrow \sigma = se(\hat{\beta}_2) \cdot \sqrt{\sum x_i^2} \quad (5.b.2)$$

Now we combine (5.b.2) with (5.b.1):

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}} \left(se(\hat{\beta}_2) \cdot \sqrt{\sum x_i^2} \right)$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n}} \, se(\hat{\beta}_2)$$

$$se(\hat{\beta}_1) = \sqrt{\frac{300000}{20}} \, 0.05 = 6.1237$$

c) Establish 95% confidence interval for the parameters

n-2=20-2=18

α=5%

$$t_{n-2}^{\alpha/2} = t_{18}^{0.025} = 2.101$$

$$\hat{\beta}_1 - t_{n-2}^{\alpha/2} se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{n-2}^{\alpha/2} se(\hat{\beta}_1)$$

$$8 - 2.101 \cdot 6.1237 \le \beta_1 \le 8 + 2.101 \cdot 6.1237$$

$$-4.8659 \le \beta_1 \le 20.8629$$

$$\hat{\beta}_2 - t_{n-2}^{\alpha/2} se(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{n-2}^{\alpha/2} se(\hat{\beta}_2)$$

$$0.6267 - 2.101 \cdot 0.05 \le \beta_2 \le 0.6267 + 2.101 \cdot 0.05$$

$$0.5216 \le \beta_2 \le 0.7317$$

d) On the basis of the confidence intervals established in (c), can y ou reject the hypothesis that β_2 is statistically significant?

Confidence interval established for β_2 :

$$0.5216 \le \beta_2 \le 0.7317$$

This confidence interval [0.5216; 0.7317] doesn't contain zero, therefore β_2 **is** statistically significant or is statistically different from zero. That is why we cannot reject the hypothesis of β_2 being statistically significant.

e) Test whether $\beta_2 > 1$.

The appropriate type of hypothesis in this case is the left -tail test:

Null hypothesis

$H_0: \beta_2 > 1$

Alternative hypothesis

$H_1: \beta_2 \leq 1$

$$t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} \qquad t_{\beta_2} = \frac{0.6267 - 1}{0.05} = -7.4667$$

$$t^{\alpha}_{n-2} = t^{0.05}_{18} = 1.734$$

Decision rule

$$t_{\beta_2} < -t^{\alpha}_{n-2}$$

-7.4667<-1.734 -True => we reject the null hypothesis that $\beta_2 > 1$.

Bibliography

BROOKS, Chris, 2002. **Introductory econometrics for finance**. New York: Cambridge University Press.

GUJARATI, Damodar N., 2003. **Basic Econometrics**. 4th ed. New York: McGraw-Hill.

KOOP, Gary, . **Analysis of Economic Data. Chichester**, England: John Wiley & Sons.

WOOLDRIDGE, Jeffrey M., 2006. **Introductory Econometrics**. 3rd ed. Manson, USA: Thomson South-Western.