Economics Honours (Semester VI)

Basic Econometrics -I

Testing of Regression Coefficients

A. Important Statistical Concepts

1. Population and Sample

In collecting data concerning the characteristics of a group of individuals or objects, such as the heights and weights of students in a university or the numbers of defective and non-defective bolts produced in a factory on a given day, it is often impossible or impractical to observe the entire group, especially if it is large. Instead of examining the entire group, called the *population*, or universe, one examines a small part of the group, called a *sample*.

A population can be *finite* or *infinite*. For example, the population consisting of all bolts produced in a factory on a given day is finite, whereas the population consisting of all possible outcomes (heads, tails) in successive tosses of a coin is infinite.

If a sample is representative of a population, important conclusions about the population can often be inferred from the analysis of the sample. The phase of statistics dealing with conditions under which such inference is valid is called *statistical inference*. Because such inference cannot be absolutely certain, the language of *probability* is often used in stating conclusions.

2. Sampling Theory

Sampling theory is a study of relationships existing between a population and samples drawn from the population. It is of great value in many connections. For example, it is useful in estimating unknown population quantities (such as mean and variance), often called *population parameters* or briefly *parameters*, from the knowledge of corresponding sample quantities (such as sample mean and variance), often called *sample statistics* or briefly *statistics*.

Sampling theory is also useful in determining whether the observed differences between two samples are due to chance variation or whether they are really significant. Such questions arise, for example, in testing new serum for use in treatment of a disease or in deciding whether one production process is better than another. Their answers involve the use of *tests of significance and hypotheses*.

In general, a study of the inferences made concerning a population by using samples drawn from it, together with indications of the accuracy of such inferences by using probability

theory, is called *statistical inference*. In order that the conclusions of sampling theory and statistical inference be valid, samples must be chosen so as to be *representative* of a population.

3. PRF and SRF

Given *population* data, the classical linear regression model in two variables is expressed as $\mathbf{Y_i = f(X_i) = \beta_1 + \beta_2 X_i + u_i, (i=1,2,\ldots,N),}$ is the *Population Regression Function* (PRF).

Here X_i is an explanatory variable, Y_i is a dependent variable and function of X_i, β_1 and β_2 are population regression coefficients and u_i = error term.

In most practical situations we do not have access to the population data, but deal with a sample of Y values corresponding to some fixed Xs from the population. The regression model estimated on the basis of the sample is called the *Sample Regression Function* (SRF). This can be expressed as $\mathbf{Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i, \text{ for } i=1,2,\ldots,n}$

where $\hat{\beta}_1$ & $\hat{\beta}_2$ are estimated sample regression coefficients and $\hat{u}_i$ = estimated residual term.

This shows that each new sample data from the population of size N, will help us to determine a new set of sample regression coefficients. If the estimated regression coefficients are dependent on sample data, as the sample data are likely to change from sample to sample, the estimates will change as a result. Hence we need some measure of accuracy of the estimators $\mathbf{\hat{\beta}_1}$ **&** $\mathbf{\hat{\beta}_2}$ in predicting the behaviour of the population. In statistics the accuracy of an estimate is measured by its standard error (se).

4. Standard Error

The standard errors of the OLS estimates can be obtained as follows:

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

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

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2 \sum X_i^2}{n \sum x_i^2}$$

$$\text{se}(\hat{\beta}_1) = \frac{\sigma \sqrt{\sum X_i^2}}{\sqrt{n \sum x_i^2}}$$

where var = variance, se = standard error and σ^2 is estimated from the sample data.

The formula for estimated σ^2 is given by

$$\hat{\sigma}^2 = (\sum \hat{u}_i^2) / (n-2)$$

a. where $\hat{\sigma}^2$ is the OLS estimator of true but unknown σ^2,

b. the expression n-2 is known as the **number of degrees of freedom** (df) i.e. [number of observations in the sample – number of parameters estimated which in our case is 2] ,

c. $\Sigma \hat{u}_i^2$ = sum of the residuals squared or the residual sum of squares.

5. Normal distribution of the estimated regression coefficients

$\hat{\beta}_1$(being a linear function of u_i) is *normally distributed* with

Mean: $E(\hat{\beta}_1) = \beta_1$

var $(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \dfrac{\sigma^2 \Sigma X_i^2}{n \Sigma x_i^2}$

which can also be expressed as

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

Then by the properties of the normal distribution, the variable Z is defined as

$$Z = (\hat{\beta}_1 - \beta_1) / \sigma^2_{\hat{\beta}_1}$$

follows the standard normal distribution, that is, a normal distribution with zero mean and unit (=1) variance or

$$Z \sim N(0,1)$$

Similarly we can state, $\hat{\beta}_2$(being a linear function of u_i) is *normally distributed* with

Mean: $E(\hat{\beta}_2) = \beta_2$

var $(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \dfrac{\sigma^2}{\Sigma x_i^2}$

which can also be expressed as

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$$

Then by the properties of the normal distribution, the variable Z is defined as

$$Z = (\hat{\beta}_2 - \beta_2) / \sigma^2_{\hat{\beta}_2}$$

follows the standard normal distribution, that is, a normal distribution with zero mean and unit (=1) variance or

$$Z \sim N(0,1)$$

6. Testing the statistical significance of $\beta\hat{}_1$ & $\beta\hat{}_2$: The *t* Test

Let us take the help of the following **example** on **Corn production with Fertilizer used** to understand the concept of 'Test of Significance of Parameter Estimates'

Table 1: Statistical workings

Obs	(X_i) (Fertilizer)	(Y_i) (Corn)	x_i	x_i^2	y_i	y_i^2	x_iy_i
1	6	40	-12	144	-17	289	204
2	10	44	-8	64	-13	169	104
3	12	46	-6	36	-11	121	66
4	14	48	-4	16	-9	81	36
5	16	52	-2	4	-5	25	10
6	18	58	0	0	1	1	0
7	22	60	4	16	3	9	12
8	24	68	6	36	11	121	66
9	26	74	8	64	17	289	136
10	32	80	14	196	23	529	322
n =10	$\Sigma X_i = 180$ $\bar{X} = 18$	$\Sigma Y_i = 570$ $\bar{Y} = 57$	$\Sigma x_i = 0$	$\Sigma x_i^2 = 576$	$\Sigma y_i = 0$	$\Sigma y_i^2 = 1634$	$\Sigma x_iy_i = 956$

Source: Schaum's Outlines on Statistics and Econometrics by Dominick Salvatore and Derrick Reagle

Note:

a. $\hat{\beta}_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = 956/576 = 1.66$ [slope coefficient i.e. slope of the estimated regression line]

b. $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 57 - 1.66 \times 18 = 27.12$ [the Y intercept i.e. the intercept coefficient]

c. $\hat{Y}_i = 27.12 + 1.66 X_i$ [the estimated regression equation]

Table 1: Statistical workings contd…

Obs	(X_i) (Fertilizer)	(Y_i) (Corn)	$\hat{Y}_i$	u_i	u_i^2	X_i^2
1	6	40	37.08	2.92	8.5264	36
2	10	44	43.72	0.28	0.0784	100
3	12	46	47.04	-1.04	1.0816	144
4	14	48	50.36	-2.36	5.5696	196
5	16	52	53.68	-1.68	2.8224	256
6	18	58	57.00	1.00	1.0000	324
7	22	60	63.64	-3.64	13.2496	484
8	24	68	66.96	1.04	1.0816	576
9	26	74	70.28	3.72	13.8384	676
10	32	80	80.24	-0.24	0.0576	1024
n =10	$\Sigma X_i = 180$ $\bar{X} = 18$	$\Sigma Y_i = 570$ $\bar{Y} = 57$		$\Sigma u_i = 0$	$\Sigma u_i^2 = 47.3056$	$\Sigma X_i^2 = 3816$

Source: Schaum's Outlines on Statistics and Econometrics by Dominick Salvatore and Derrick Reagle

Calculation of se of $\hat{\beta}_1$ & $\hat{\beta}_2$

$\hat{\sigma}^2 = (\Sigma \hat{u}_i^2) / (n-2) = 47.3056/(10-2) = 47.3056/8 = 5.9132$

$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\Sigma x_i^2} = 5.9132/576 = 0.01027$

d. Therefore se($\hat{\beta}_2$) = $\sqrt{0.01027}$ = ± 0.101

$\text{var}(\hat{\beta}_1) = \frac{\sigma^2 \Sigma X_i^2}{n \Sigma x_i^2} = \frac{5.9132 \times 3816}{10 \times 576} = 22564.7712/5760 = 3.92$

e. Therefore se($\hat{\beta}_1$) = $\sqrt{3.92}$ = ± 1.98

As u_i is normally distributed, Y_i and therefore $\hat{\beta}_1$ and $\hat{\beta}_2$ are also normally distributed, so we use the *t* distribution with (n-2) [i.e. (10-2) =8] degrees of freedom at α = 5% level of significance to test statistical significance of $\hat{\beta}_1$ & $\hat{\beta}_2$.

Under the normality assumption,

***t_1* = ($\hat{\beta}_1 - \beta_1$) / se($\hat{\beta}_1$) = ± (27.12 - 0) / 1.98 = ± 13.7**

and

***t_2* = ($\hat{\beta}_2 - \beta_2$) / se($\hat{\beta}_2$) = ± (1.66 – 0) / 0.101 = ± 16.43**

Since both t_1 and t_2 exceed *t* = 2.306 with 8 degrees of freedom at the 5% level of significance (see *t* table), we conclude that both **$\hat{\beta}_1$ & $\hat{\beta}_2$ are statistically significant at the 5% level.**

[Note: Under normality assumption, by property of Least square estimators, $E(\hat{\beta}) = \beta_1$ and $E(\hat{\beta}_2) = \beta_2$ and according to standard normal distribution mean of the parameters is zero.]

*N.B. Explanations of all the theories have been taken from the following references:

References:

Gujarati, D. N. Porter, D.C., Gunasekar, S. (2009), *Basic econometrics*. (Fifth ed.) McGraw-Hill Education (India).

Salvatore, D., & Reagle, D. (2011). *Schaum's outline of statistics and econometrics* (2nd ed.). McGraw-Hill Education.