UNIT 6 EXTENSION OF TWO VARIABLE REGRESSION MODELS*

Structure

6.0 OBJECTIVES

After going through this Unit, you should be in a position to

- interpret regression models passing through the origin;
- explain the impact of changes in the unit of measurement of dependent and independent variables on the estimates;
- interpret parameters in semi-log and log-linear regression models; and
- identify the correct functional form of a regression model.

6.1 INTRODUCTION

In the previous two Units we have discussed how a two variable regression model can be estimated and how inferences can be drawn on the basis of the estimated regression equation. In this context we discussed about the ordinary least squares (OLS) method of estimation. Recall that the OLS estimators are the best linear unbiased estimators (BLUE) in the sense that they are the best in the class of linear regression models.

The two variable regression model has the function as follows:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \ldots (6.1)$$

*Prof. Kaustuva Barik, Indira Gandhi National Open University, New Delhi

where Y is the dependent variable and X is the independent variable. We added a stochastic error term (u_i) to the regression model. We cited three reasons for inclusion of the error term in the regression model: (i) it takes care of the excluded variables in the model, (ii) it incorporates the unpredictable human nature into the model, and (iii) it absorbs the effects measurement error, incorrect function form, etc.

We assumed that the regression model is correctly specified. All relevant variables are included in the model. No irrelevant variable is included in the regression model. In this Unit we will continue with the two variables case as in the previous two units. We also continue with the same assumptions, as mentioned in Unit 4.

Let us look into the regression model given at equation (6.1). We observe that the regression model is linear in parameters. We do not have complex forms of the parameters such as β_2^2 or $\beta_1\beta_2$ as parameters. Further, the regression model is linear variables. We do not have X^2 or log X as explanatory variable. Can we have these sorts of variables in a regression model? How do we interpret the regression model if such variables are there? We will extend the simple regression model given in equation (6.1) and explain how the interpretation of the model changes with the modifications.

6.2 REGRESSION THROUGH THE ORIGIN

Let us look into the simple regression model given at equation (6.1). There are two parameters in the regression model: β_1 and β_2. The intercept parameter is β_1 and the slope parameter is β_2. The intercept β_1 indicates the value of the dependent variable when the explanatory variable takes the value zero, i.e., $E(Y_0|X_0) = \beta_1$.

Suppose regression model takes the following form:

$$Y_i = \beta_2 X_i + u_i \qquad \ldots (6.2)$$

In equation (6.2) there is only one slope parameter, β_2. There is no intercept. The implication is that the regression line passes through the origin. The population regression function is $Y = \beta_2 X_i + u_i$ and the sample regression function is $Y_i = b_2 X_i + e_i$.

Now let us apply OLS method and find out the OLS estimator b_2. As you know from Unit 4, in OLS method we minimise the error sum of squares (ESS). Thus we minimise

$$ESS = \sum e_i^2 = \sum (Y_i - b_2 X_i)^2 \qquad \ldots (6.3)$$

We take derivative of the ESS and equate it to zero.

$$\frac{d\sum e_i^2}{db_2} = 0 \qquad \ldots (6.4)$$

$$\frac{d\sum e_i^2}{db_2} = 2\sum (Y_i - b_2 X_i)(-X_i) = 0 \qquad \ldots (6.5)$$

This implies

$-2\sum e_i(X_iY_i - b_2X_i^2) = 0$

$\sum X_iY_i - b_2\sum X_i^2 = 0$

$b_2 = \frac{\sum X_iY_i}{\sum X_i^2}$ … (6.6)

The estimator given at (6.6) is unbiased. The variance of the estimator is given by

$var(b_2) = \frac{\sigma^2}{\Sigma X_i^2}$ … (6.7)

Let us compare the above estimator with the estimator for the regression model $Y_i = \beta_1 + \beta_2 X_i + u_i$ (see equation (4.18) in Unit 4)

$b_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$ … (6.8)

and

$var(b_2) = \frac{\sigma^2}{\Sigma x_i^2}$ … (6.9)

Note that in equation (6.6) the variables are not in deviation form. Thus when we do not have an intercept in the regression model, the estimator of the slope parameter is different from that of a regression model with intercept. Both the estimators will be the same if and only if $\bar{X} = 0$.

We present a comparison between the regression model with intercept and without intercept in Table 6.1.

Table 6.1: Features of Regression Model without Intercept

Regression Model with Intercept	Regression Model without Intercept
$b_2 = \frac{\sum x_i y_i}{\sum x_i^2}$	$b_2 = \frac{\sum X_i Y_i}{\sum X_i^2}$
$\text{var}(b_2) = \frac{\sigma^2}{\sum x_i^2}$	$\text{var}(b_2) = \frac{\sigma^2}{\sum X_i^2}$
$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$	$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-1}$
R^2 is non-negative	R^2 can be negative

The estimated regression model is given as

$\hat{Y}_i = b_2 X_i$ … (6.10)

Note that the coefficient of determination R^2 is not appropriate for regression models without the intercept. If the intercept in a regression model is not statistically significant, then we can have regression through the origin. Otherwise, it leads to specification error. There is omission of a relevant variable.

6.3 CHANGES IN MEASUREMENT UNITS

Suppose you are given time series data on GDP and total consumption expenditure of India for 30 years. You are asked to run a regression model with consumption expenditure as dependent variable and income as the independent variable. The objective is to estimate the aggregate consumption function of India. Suppose you took GDP and Consumption Expenditure in Rs. Crore. The estimated regression equation you found is

$$Y_i = 237 + 0.65X_i \quad \ldots (6.11)$$

When you presented the results before your seniors, they pointed out that the measure of GDP and consumption expenditure should have been in Rs. Million, so that it is comprehensible outside India also. If you re-estimate the results by converting the variables, will estimates be the same? Or, do you expect some changes in the estimates? Let us discuss the issue in details.

Suppose we transform both the dependent and independent variables as follows:

$$Y_i^* = w_1 Y_i \text{ and } X_i^* = w_1 X_i \quad \ldots (6.12)$$

The regression model (6.1) can be transformed as follows:

$$Y_i^* = \beta_1 + \beta_2 X_i^* + u_i \quad \ldots (6.13)$$

Estimation of equation (6.13) by OLS method gives us the following estimators

$$b_1^* = \bar{Y}^* - b_2^* \bar{X}^* \quad \ldots (6.14)$$

$$b_2^* = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}} \quad \ldots$$
(6.15)

In a similar manner you can find out the variance of b_1^* and b_2^*, and the estimator of the error variance.

From equation (6.15) we can find out that

$$b_2^* = \frac{w_1}{w_2} b_2 \quad \ldots (6.16)$$

and

$$b_1^* = w_1 b_1 \quad \ldots (6.17)$$

Now let us look into the implications of the above.

(i) Let us begin with the dependent variable, Y_i . Suppose Y_i is doubled ($w_1 = 2$) and X_i is unchanged ($w_2 = 1$). What will happen to b_1 and b_2? Substitute the values of w_1 and w_2 in equations (6.16) and (6.17). We find that both the estimates are doubled. Thus, if the dependent variable is multiplied by a constant *c*, then all OLS coefficients will be multipled by *c*.

(ii) Now let us take the case of the independent variable. Suppose X_i is doubled ($w_2 = 2$) and Y_i is unchanged ($w_1 = 1$). On substitution of the values of w_1 and w_2 in equations (6.16) and (6.17) we find that the slope coefficient (b_2) is halved, but the intercept (b_1) remains unchanged.

(iii) If we double both the variables X_i and Y_i, then the slope coefficient (b_2)will remain unchanged, but the intercept will change. Remember that the intercept is changed by a change in the scale of measurement of the dependent variable.

Now the question arises: Will there be a change in the t-ratio and F-value of the model? No, the *t* and *F* statistics are not affected by a change in the scale of measurement of any variable.

Check Your Progress 1

1) Under what condition should we run a regression through the origin?

....
....
....
....
....
....

2) What are the implications of a regression model through origin?

....
....
....
....
....

3) What are the implications on the estimates if there is a change in the measurement scale of the explanatory variable?

....
....
....
....
....

4) What are the implications on the estimates if there is a change in the measurement scale of the dependent variable?

....
....
....
....
....

6.4 SEMI-LOG MODELS

In some of the cases the regression model is non-linear, but by taking logarithm on both sides of the regression equation, we get a linear model. If a model is non-linear , but becomes linear after transformation of its variables, then the model is said to be intrinsically linear . Thus, semi-log and log-linear models are intrinsically linear models. We discuss about the semi-log model in this section. We will discuss about the log-linear model in the next section.

Let us begin with a functional form as follows:

$$Y_t = e^{\beta_1 + \beta_2 X_t + u_t} \quad \ldots (6.18)$$

This regression model, in its present form, is non-linear . Therefore, it cannot be estimated by OLS method. However, if we take natural logs of both the sides, we obtain

$$\ln Y_t = \beta_1 + \beta_2 t + u_t \quad \ldots (6.19)$$

It transforms into a semi-log equation. It is called a semi-log model as one of the variables is in log form.

If we take $\ln Y_t = Y_t^*$, then equation (6.19) can be written as

$$Y_t^* = \beta_1 + \beta_2 t + u_t \quad \ldots (6.20)$$

Estimation of equation (6.20) is simple. The equation is linear in parameters and in variables. Thus, we can apply OLS method to estimate the parameters. The implication of the regression model (6.20), however, is much different from the regression model (6.1).

If we take the differentiation of the regression model $Y_t = \beta_1 + \beta_2 X_t + u_t$, we obtain

$$\frac{dY}{dt} = \beta_2 \quad \ldots (6.21)$$

Equation (6.21) shows that the slope of the regression equation is constant. An implication of the above is that the absolute change in the dependent variable for unit increase in the independent variable is constant throughout the sample. If there is an increase in X by one unit, Y increases by β_2 unit.

Now let us consider the regression model $\ln Y_t = \beta_1 + \beta_2 t + u_t$. If we take differentiation of equation (6.19) we find that

$$\frac{d \ln Y_t}{dt} = \beta_2$$

which means

$$\frac{1}{Y_t}\frac{dY_t}{dt} = \beta_2 \quad \ldots (6.22)$$

An implication of equation (6.22) is that the slope of the regression model is variable. Thus its interpretation is different from that of the regression model $Y_t = \beta_1 + \beta_2 X_t + u_t$.

For equation (6.19), we interpret the slope coefficient (β_2) as follows: For every unit increase in *X*, there is β_2 per cent increase *Y*. Thus, for a semi-log model the change in the dependent variable in terms of percentages. The semi-log model is useful is estimating growth rates.

6.5 LOG-LINEAR MODELS

Let us consider the following regression equation:

Let us take the case of the following non-linear model

$$Y = \beta_1 X^{\beta_2} \quad \ldots (6.23)$$

This model will be intrinsically linear if it can be transformed into

$$Y^* = \beta_1 + \beta_2 X^* + u \quad \ldots (6.24)$$

Using the logarithm of each of the variable in equation (6.23), we get the following transformed equation:

$$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i \quad \ldots (6.25)$$

The regression model given at (6.25) is called log-linear model (because it is linear in logs of the variables) or double-log model (because both variables are in log form).

Let us take differentiation of equation (6.25) with respect to X_i

$$\frac{d(\ln Y_i)}{dX_i} = \frac{1}{Y_i} \cdot \frac{dY_i}{dX_i} \quad \ldots (6.26)$$

$$\frac{dY_i}{dX_i} = \frac{\beta_2}{X_i} \quad \ldots (6.27)$$

By combining equations (6.26) and (6.27) we find that

$$\frac{dY_i}{dX_i} = \frac{Y_i}{X_i}\beta_2$$

Or,

$$\frac{dY_i}{dX_i}\frac{X_i}{Y_i} = \beta_2 \quad \ldots (6.28)$$

A closer look at equation (6.28) shows that the slope parameter represents the elasticity between Y and X.

This attractive feature of the log-linear model has made it popular in applied work. The slope coefficient β_2 measures the elasticity of *Y* with respect to *X*, that is, the percentage change in *Y* for one per cent change in *X*. Thus, if *Y* represents the quantity of a commodity demanded and *X* its unit price, then β_2 measures the price elasticity of demand.

6.6 CHOICE OF FUNCTIONAL FORM

By you would have observed that the two variable regression model could have three functional forms as given below.

(I) $Y_i = \beta_1 + \beta_2 X_i + u_i$

(II) $lnY_i = \beta_1 + \beta_2 X_i + u_i$

(III) $\ln Y_i = \beta_1 + \beta_2 \ln x_i + u_i$

A question arises: which one is the best model? The choice of functional form depends on our objective. We should choose the model that gives us relevant answer to our queries. Suppose our objective is to estimate the impact of change in the independent variable on the dependent variable. In this case we can use model-I. On the other hand, if our objective is to estimate growth rate in the dependent variable as a result of the change in the independent variable, we should opt for semi-log model (model II). If our objective is to estimate elasticity between two variables, we choose the log-linear model.

The three regression models (Models –I, II, III) will give different estimates of the parameters. The standard error of the estimators will also be different. Further, the coefficient of determination, R^2, will be different for all three models. Can we compare the R^2 of the models and say that the model with the highest R^2 is the best fit? We cannot compare the value of R^2 obtained from regression models with different dependent variables. However, we can compare R^2 of regression models with the same dependent variable and the same estimation method. Thus the R^2 value of Model-I and Model-II cannot be compared. We can compare Model-II and Model-III in terms of their best fit.

If two regression models are almost similar in terms of their coefficient of determination, statistical significance of estimators and diagnostic checking (to be discussed in Units 13 and 14), we prefer the simpler model. The simpler model is easier to comprehend and usually accepted by others.

The log-linear regression model has certain advantages: (i) the parameters are invariant to change of scale since they measure percentage changes, (ii) the model gives elasticity figures directly, and (iii) the model moderates the problem of heteroscedasticity to some extent (see Unit 11 for the problem of heteroscedasticity).

Check Your Progress 2

1) In a semi-log model how do you interpret the slope coefficient?

....
....
....

2) Describe how the slope parameter of a log-linear regression model is estimated.

....
....

....
....
....

2) What are the advantages of the log-linear model?

....
....
....
....
....

3) What is meant by intrinsically linear model? Can you compare the results of an intrinsically linear model with that of a linear model? Why or why not?

....
....
....
....
....

6.7 LET US SUM UP

In this Unit we discussed about the functional forms that can be accommodated in a two variable regression model. We began with the regression model passing through the origin (there is no intercept). We pointed out the impact of changes in the scale of measurement of variables. Subsequently we considered three functional forms: the original model, the semi-log model and the log-linear model. The interpretations of the parameters in all three functional forms have been discussed in the Unit.

6.8 ANSWERS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) The exclusion of intercept term from a regression model has serious implication. It should be omitted only when the intercept term in the unrestricted model is statistically not significant.

2) We have listed the implications of the omission of the intercept term in table 6.1. Go through it and answer.

3) When there is a change in the measurement scale of the explanatory variable the concerned estimate is affected. If X is multiplied by *c*, the parameter is divided by *c*.

4) If Y is multiplied by c, all parameters in the model are multiplied by c.

Check Your Progress 2

1) In a semi-log model the slope parameter indicates growth rate. If there is 1 unit increase in the value of X, the expected value of Y increases by β per cent.

2) The estimation of the log-linear model is the same as the simple regression model, except that the variables are transformed. Write down the steps followed in estimation of a regression model.

3) We have mentioned three advantages in the text: (i) the parameters are invariant to change of scale since they measure percentage changes, (ii) the model gives elasticity figures directly, and (iii) the model lessens the problem of heteroscedasticity to some extent.

4) You cannot compare the results of two regression models unless the dependent variable is the same.