Chapter 2

Simple Linear Regression Model : Two Variable Case

Learning Objectives :

After learning this chapter you will understand :

- **Population Regression Function.**
- **Stochastic Error Term.**
- **Sample Regression Function.**
- **Method of Ordinary Least Squares.**
- **Assumptions of CLRM.**
- **Properties of OLS Estimators.**
- **Gauss-Markov Theorem.**
- **Hypothesis Testing of OLS Estimators.**
- **Coefficient of Determination.**
- **Normality Tests :**
  - ✓ **Normal Probability Plot,**
  - ✓ **Jarque-Bera Test.**
- **Forcasting.**
- **Scaling and units of measurement.**

Basic Concepts

1. **Regression :** Regression means returning or stepping back to average or normal. It was first used by Sir francis Galton. Regression analysis, in general sense, means the estimation or prediction of the unknown value of one variable from the known value of other variable. In the words of M. M. Blair, "Regression analysis is a mathematical measure of the average relationship between two or more variables in terms of the original units of the data". It is one of the very important statistical tools which is extensively used in business and economics to study the relationship between two or more variables that are related casually and for estimation of demand and supply curves, cost functions, production and consumption functions, etc.

2. **Linear Regression :** Since in this text we are only concerned with linear regression models, so it is essential to know the meaning of linearity. The term linearity can be interpreted in two different ways as under :

   (i) **Linearity in the Variables :** If all the variables involved in the regression equation have degree one then such regression equation is said to linear in variables.

   (ii) **Linearity in the Parameters :** If the conditional expectation of Y, $E(Y|X_i)$ is a linear function of parameters, the β's, then it is linear in parameters. Here it may or may not be linear in variables.

   In our analysis by linearity we mean linear in parameters only. The regression model may or may not be linear in variables, the X's, but it is essentially linear in parameters, the β's.

3. **Regression Equations :** Regression equations are used to estimate the value of one variable on the basis of the value of other variable. If we have two variables X and Y the two regression for them are X on Y and Y on X.

4. **Regression Equation of Y on X :** The line of regression of Y on X is the line which gives the best estimate of Y for any given value of X.

5. **Regression Equation of X on Y :** The line of regression of X on Y is the line which gives the best estimate of X for any given value of Y.

> *Note :* Generally we study only regression equation of Y on X, where Y is the dependent or explained variable and X is independent or explanatory variable.

6. **A Linear Probabilistic Model :** As a first approximation we may assume that the ***Population Regression Function*** (PRF) is a linear function, *i.e.*, it is of the type :

   $$E(Y/X_i) = \beta_1 + \beta_2 X_i$$

   Where,

   $E(Y/X_i)$ means the mean or expected value of Y corresponding to or conditional upon a given value of X, here by linearity we mean linear in parameters.

   β_1 and β_2 are unknown but fixed parameters known as regression coefficients. β_1 is called intercept term and β_2 is called slope term.

7. **Stochastic Specification of PRF :** $E(Y/X_i) = \beta_1 + \beta_2 X_i$ form of the regression function implies that the relationship between X and Y is exact, that is all the variations in Y are solely due to changes in X and there are no other factors affecting the dependent variable. If this were true then all the points of X and Y pairs, if plotted on a two dimensional plane, would fall on a straight line. However if we if we gather observations on the actual data and plot them on a diagram, we see they do not fall on a straight line (or any other smooth curve for that matter). The deviations of the observations from the line may be attributed to several factors.

   In Statistical analysis, however, one generally acknowledges the fact that the relationship is not exact by explicitly including a random factor, known as disturbance term in the linear regression model. So the linear regression model becomes :

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

   Where u_i is known as the stochastic or random or residual or error term.

8. **Sample Regression Function :** Suppose that we have taken a sample of few values of Y corresponding to given X_i. The line which is drawn to fit the data is called sample regression line. Mathematically, sample regression line can be expressed as :

   $$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \quad \text{or} \quad \hat{Y}_i = b_1 + b_2 X_i$$

   Where,

   $\hat{Y}_i$ is the estimator of $E(Y/X_i)$ or the estimator of population conditional mean,

   $\hat{\beta}_1$ or b_1 is the estimator of β_1, and

   $\hat{\beta}_2$ or b_2 is the estimator of β_2.

9. **Stochastic Specification of SRF :** The stochastic sample regression function can be expressed as :

   $$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \quad \text{or} \quad Y_i = \hat{Y}_i + e_i$$

   Where, e_i is the estimator of population error term u_i.

10. **Estimating Model Parameters :** Our basic objective in regression analysis is to estimate the stochastic PRF

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

    On the basis of the SRF

    $$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

    Because, generally our assumption is based on a single sample from some population. But due to sampling variations our estimate of PRF based on SRF is only approximate. To estimate the regression equation we use the method of least squares.

11. **Assumptions of Classical Linear Regression Model :** In order to use the method of ordinary least squares (OLS), the following basic assumptions must hold for a two variable regression model :

Assumption 1: Linear regression model. The regression model $Y_i = \beta_1 + \beta_2 X_i + u_i$ is **linear in the parameters.** This interpretation of linearity is that the conditional expectation of Y, $E(Y \mid X_i)$, is a linear function of the parameters, the β's; it may or may not be linear in the variable X. In this interpretation $E(Y \mid X_i) = \beta_1 + \beta_2 X^2_i$ is a linear (in the parameter) regression model. To see this, let us suppose X takes the value 3. Therefore, $E(Y \mid X = 3) = \beta_1 + 9\beta_2$, which is obviously linear in β_1 and β_2.

Note : A function is said to be linear in the parameter, say, β_1, if β_1 appears with a power of 1 only and is not multiplied or divided by any other parameter (for example, $\beta_1\beta_2$, β_2/β_1, and so on).

Assumption 2: *X* values are fixed in repeated sampling. Values taken by the regressor X are considered fixed in repeated samples. More technically, X is assumed to be *nonstochastic.*

Assumption 3: Zero mean value of disturbance u_i. Given the value of X, the mean, or expected, value of the random disturbance term u_i is zero. Technically, the conditional mean value of u_i is zero. Symbolically, we have $E(u_i \mid X_i) = 0$. A violation of this assumption introduces bias in the intercept of the regression equation.

Assumption 4: Homoscedasticity or equal variance of u_i. Given the value of X, the variance of u_i is the same for all observations. That is, the conditional variances of *ui* are identical. Symbolically, we have **var**$(u_i \mid X_i) = E[u_i - E(u_i \mid X_i)]^2 = E(u_i^2 \mid X_i)$ $= \sigma^2$

where **var** stands for variance.

Assumption 5: No autocorrelation between the disturbances. Given any two X values, X_i and X_j ($i \neq j$), the correlation between any two u_i and u_j ($i \neq j$) is zero. Symbolically,

cov $(u_i, u_j \mid X_i, X_j) = E\{[u_i - E(u_i)] \mid X_i\}\{[u_j - E(u_j)] \mid X_j\}$

$= E(u_i \mid X_i)(u_j \mid X_j)$

$= 0$

where i and j are two different observations and where cov means covariance.

Assumption 6: Zero covariance between ui and Xi, or E(uiXi) = 0. Formally,

cov $(u_i, X_i) = E[u_i - E(u_i)][X_i - E(X_i)]$

$= E[u_i (X_i - E(X_i))]$ since $E(u_i) = 0$

$= E(u_iX_i) - E(X_i)E(u_i)$ since $E(X_i)$ is nonstochastic

$= E(u_iX_i)$ since $E(u_i) = 0$

$= 0$ by assumption

Assumption 7: The number of observations n must be greater than the number of parameters to be estimated. Alternatively, the number of observations n must be greater than the number of explanatory variables.

Assumption 8: Variability in X values. The X values in a given sample must not all be the same. Technically, var (X) must be a finite positive number.

Assumption 9: The regression model is correctly specified. Alternatively, there is no specification bias or error in the model used in empirical analysis.

Assumption 10: There is no perfect multicollinearity. That is, there are no perfect linear relationships among the explanatory variables.

12. **Method of Least Squares :** The method of Ordinary Least Squares is attributed to a German Mathematician Carl Friedrich Gauss. This method is based on the assumption that the sum of the squares of differences between the estimated values and the actual observed values of the observations is minimum and the variables X and Y are related according to the simple linear regression model. The values of β_1, β_2 and σ^2 will almost never be known to an investigator. Instead, sample data consisting of *n* observed pairs (X_1, Y_1), (X_2, Y_2), …., (X_n, Y_n) will be available, from which the model parameters and the true regression line itself can be estimated. The least squares method is used to obtain the best fitting straight line to the given data and consists in selecting the best fitting straight line to the values of independent variable for dependent variable. Using this method the regression equation may be found as under :

Regression Equation of Y on X :

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

Normal Equations

$$\sum Y = n\hat{\beta}_1 + \hat{\beta}_2 \sum X \text{ and } \sum XY = \hat{\beta}_1 \sum X + \hat{\beta}_2 \sum X^2$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are constants and their values can be obtained by solving the normal equations.

The least square estimate of the slope coefficient $\hat{\beta}_2$ of true regression line is

$$\hat{\beta}_2 = \frac{\sum[(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum[X_i - \bar{X}]^2} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2}$$

The least square estimate of the intercept coefficient $\hat{\beta}_1$ of true regression line is :

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Note : **Regression Equation of *X* on *Y* :**

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

Normal Equations

$$\sum X = n\hat{\beta}_1 + \hat{\beta}_2 \sum Y \text{ and } \sum XY = \hat{\beta}_1 \sum Y + \hat{\beta}_2 \sum Y^2$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are constants and their values can be obtained by solving the normal equations.

13. **Variations in Y_i :** The variations in Yi may be classified as under :

(i) **Total Variation :** The total variation in actual Y values about their sample mean $\bar{Y}$ is called total variation in Y. it may also be called ***total sum of squares (TSS)***. $\sum y_i^2$ is a measure of total variation in Y, such that,

TSS = $\sum(Y_i - \bar{Y})^2$, *i.e.*, $\sum y_i^2$,

(ii) **Explained Variation :** $\sum \hat{y}_i^2 = \sum\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2 = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 = b_2^2 \sum\left(X_i - \bar{X}\right)^2$ is the variation of the estimated Y values $\left(\hat{Y}_i\right)$ about their mean $\left(\bar{\hat{Y}} = \bar{Y}\right)$, which appropriately may be called the sum of squares due to regression, (i.e., due to explanatory variables) or simply the ***explained sum of squares (ESS)***, such that ESS = $\sum(\hat{Y}_i - \bar{Y})^2 = \frac{\left(\sum (X-\bar{X})(Y-\bar{Y})\right)^2}{\sum (X-\bar{X})^2}$, *i.e.,* $\frac{\left(\sum xy\right)^2}{\sum x^2} = \beta_2^2 \sum x_i^2$

(iii) **Unexplained Variation :** $\sum e_i^2$ represents the unexplained variation of Y values about the regression line, or simply they are called ***residual sum of squares (RSS)***. Such that RSS = $\sum(Y_i - \hat{Y})^2$ or $\sum e_i^2$

Where,

Total Sum of Squares = Expected Sum of Squares + Residuals Sum of Squares

i.e., ***TSS = ESS + RSS***

Y_i = Actual Value, $\hat{Y}_i$ = Estimated Value, $\bar{Y}$ = Actual Mean

14. **Estimating σ² and σ :** The parameter σ² determines the amount of variability inherent in the regression model. A large value of σ² means that the observed (x_i, y_i) are quite spread out about the true regression line, whereas when σ² is small the observed points will tend to fall very close to the true regression line. The variance of the population error term σ² is usually unknown. We therefore need to replace it by an estimate using sample information. Since the population error term is unobservable, one can use the estimated residuals to find an estimate. We start by forming the residual term

$$e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

After estimating the residuals, we estimate the **residuals sum of squares** denoted by $\sum e_i^2$ as :

$$\sum e_i^2 = \sum(Y_i - \hat{Y}_i)^2 = \sum[Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i]^2$$

We observe that, first of all two parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ must be estimated, which implies a loss of two degrees of freedom. With this information we may use the following formula for estimating σ² as under :

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$$

In passing, note that the positive square root of $\hat{\sigma}^2$

$$\hat{\sigma} = \sqrt{\frac{\sum e_i^2}{n-2}}$$

is known as the standard error of estimate or the standard error of the regression (se). It is simply the standard deviation of the Y values about the estimated regression line and is often used as a summary measure of the "goodness of fit" of the estimated regression line.

Note : The term number of **degrees of freedom** means the total number of observations in the sample (= n) less the number of independent (linear) constraints or restrictions put on them. In other words, it is the number of independent observations out of a total of n observations. For example, before the $\sum e_i^2$ can be computed, $\hat{\beta}_1$ and $\hat{\beta}_2$ must first be obtained. These two estimates therefore put two restrictions on the $\sum e_i^2$. Therefore, there are n – 2, not n, independent observations to compute the $\sum e_i^2$.

15. **Variances and Standard Errors of OLS Estimators :** As we know that OLS estimators are random variables, because their values will change from sample to sample. So we would like to know something about the sampling variability of these estimators. These sampling variabilities are measured by the variances of these estimators. The variances and standard errors of OLS estimators are computed by the following formulae :

$$\text{Var}(b_1) = \frac{\hat{\sigma}^2.\sum X_i^2}{n\sum(X_i - \bar{X})^2} \qquad \Rightarrow \qquad \text{SE}(b_1) = \sqrt{Var(b_1)} = \hat{\sigma}\sqrt{\frac{\sum X_i^2}{n\sum(X_i - \bar{X})^2}}$$

$$\text{Var}(b_2) = \frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2} \qquad \Rightarrow \qquad \text{SE}(b_2) = \sqrt{Var(b_2)} = \frac{\hat{\sigma}}{\sqrt{\sum(X_i - \bar{X})^2}}$$

16. **Test of Significance of Regression Coefficients :** As per central limit theorem, the regression coefficients b_1 and b_2 follow normal distribution with their means equal to true β_1 and β_2 and variances as computed above. The following steps are taken to test the significance of slope term of regression equation :

(i) Define Null hypothesis(H_0) and Alternative hypothesis(H_1).

H_0 : $\beta_2 = 0$, *i.e.*, slope term is statistically insignificant.

H_1 : Slope term is statistically insignificant, *i.e.*,

$\beta_2 \neq 0$ (Two tailed test)

$\beta_2 > 0$ (Upper tailed test)

$\beta_2 < 0$ (Lower tailed test)

(ii) Find out the tail of the test, determine whether it is single tail or two tail test.

(iii) Calculate the standard error of b_2.

(iv) Calculate the test statistic ‘t’ as under :

$$t = \frac{b_2 - \beta_2}{S.E.(b_2)}$$

(v) Set the Level of Significance ‘α’.

(vi) Find t_α (for single tail test) or $t_{\alpha/2}$ (for two tail test) for n – 2 degrees of freedom from the table.

(vii) Compare |t| and $t_{\alpha/2}$ (or t_α).

(a) If $|t| < t_{\alpha/2}$ (or t_α), then do not Reject Null hypothesis.

(b) If $|t| > t_{\alpha/2}$ (or t_α), then Reject Null hypothesis.

Similarly we can test the statistical significance of intercept term β_1.

17. **Confidence Interval :** Let us assume that 'α' is the level of significance or the probability of committing type I error, then the confidence interval of regression coefficients is computed as under :

Confidence Interval of Intercept Term :

$$P\left(b_1 - t_{\alpha/2}.SE(b_1) \le \beta_1 \le b_1 + t_{\alpha/2}.SE(b_1)\right) = 1 - \alpha$$

Confidence Interval of Slope Term :

$$P\left(b_2 - t_{\alpha/2}.SE(b_2) \le \beta_2 \le b_2 + t_{\alpha/2}.SE(b_2)\right) = 1 - \alpha$$

18. **The Coefficient of Determination** : The coefficient of determination is a measure of how well a statistical model likely to predict future outcomes. The coefficient of determination r^2 is the square of sample correlation coefficient between outcomes and predicted values. The coefficient of determination denoted by r^2 is given by :

$$\text{Coefficient of determination} = r^2 = \frac{\text{Explained Variance}}{\text{Total Variance}}$$

$$r^2 = 1 - \frac{RSS}{TSS}$$

Properties of *r*^{2} : The following two properties of *r*^{2} may be noted :

(i) *r*^{2} is a non negative quantity.

(ii) Its limits are $0 \le r^2 \le 1$.

19. **The Goodness of Fit Test :** Once the regression line has been fitted, we would like to know how good the fit is; in other words, we would like to measure the discrepancy of actual observations from the fitted line. This is important since the closer the data to the line, the better the fit or, in other words, the better the explanation of variation of the dependent variable by the independent variables. A usual measure of the goodness of fit is the square of the correlation coefficient, *r*^{2}. This is the proportion of the total variation of the dependent variable caused by the independent variable. In other words,

$$r^2 = \frac{\text{ExpectedSum of Squares(ESS)}}{\text{TotalSum of Squares(TSS)}}$$

The closer the value of *r*^{2} to 1 the better fit is the regression model, because an *r*^{2} = 1 means regression is perfect fit, *i.e.*, $\hat{Y}_i = Y_i$.

20. **Gauss Markov Theorem :** The Gauss Markov theorem states that, provided that the assumptions of CLRM are satisfied, the OLS estimators are BLUE, *i.e.*, Best (most efficient) linear (combinations of Y$_i$) unbiased estimators of the regression parameters. Thus, the OLS estimators have the following properties :

(i) **Linearity :** b$_1$ and b$_2$ are *linear* estimators, *i.e.*, they are linear functions of random variable Y$_i$.

(ii) **Unbiasedness :** OLS estimators are unbiased.

(a) b$_1$ and b$_2$ are unbiased estimates of B$_1$ and B$_2$, *i.e.*, E(b$_1$) = B$_1$ and E(b$_2$) = B$_2$.

(b) The OLS estimator of the error variance is unbiased, *i.e.*, $E(\hat{\sigma}^2) = \sigma^2$.

(iii) **Minimum Variance :** b_1 and b_2 are *efficient* estimators, *i.e.*, var(b_1) is less than the variance of any other linear unbiased estimator of B_1, var(b_2) is less than the variance of any other linear unbiased estimator of B_2. Therefore, we will be able to estimate the true B1 and B2 more precisely if we use OLS rather than any other method.

21. **Forecasting :** Although econometric theory shows that under CLRM $\hat{Y}_{X = X_o}$ or more generally $\hat{Y}_0$ is an unbiased estimator of the true mean value, it is not likely to be equal to the latter in any given sample. The difference between them is called the forecasting or prediction error. To determine such error we need to find out the sampling distribution of $\hat{Y}_0$. Given the assumptions of CLRM, it can be shown that $\hat{Y}_0$ is normally distributed with the following mean and variance :

$$\text{Mean}(\hat{Y}_0) = \beta_1 + \beta_2 X_0$$

$$Var\left(\hat{Y}_0\right) = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum x_i^2}\right]$$

So, the ***confidence interval*** or ***confidence band*** for the true $E(Y_0|X_0)$ can be computed as :

$$P\left(\hat{Y}_0 - t_{\alpha/2}.SE(\hat{Y}_0) \le \beta_1 + \beta_2 X_0 \le \hat{Y}_0 + t_{\alpha/2}.SE(\hat{Y}_0)\right) = 1 - \alpha$$

22. **Jarque Bera Test :** In Statistics, the **Jarque–Bera test** is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The test is named after Carlos Jarque and Anil K. Bera. The JB test of normality is an asymptotic or large sample test. The test statistic *JB* is defined as :

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

where *n* is the number of observations (or degrees of freedom in general); *S* is the sample skewness, and *K* is the sample kurtosis.

The Jarque-Bera statistic has an asymptotic **chi-square distribution with two degrees of freedom and can be used to test the null hypothesis that the data are from a normal distribution**. The null hypothesis is a joint hypothesis of both the skewness and excess kurtosis being 0, since samples from a normal distribution have an expected skewness of 0 and an expected excess kurtosis of 0. As the definition of JB shows, any deviation from this increases the JB statistic.

The following steps are used in the JB test :

(I) Set the hypothesis

H_0 : Errors are normally distributed

H_1 : Errors are not normally distributed

(II) Compute the test statistic

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right)$$

(III) Set the level of significance ‘α’.

(IV) Compute χ^2_α for 2 degrees of freedom.

(V) Compare JB and χ^2_α

If JB < χ^2_α, then do not reject the H_0, *i.e.*, errors are normally distributed.

If JB ≥ χ^2_α, then reject the H_0, *i.e.*, errors are not normally distributed.

Exercise 1

Proofs

Q1. Derive the line of best fit $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + u_i$ using the method of ordinary least squares.

Or

Consider the two variable population regression function $Y_i = B_1 + B_2 X_i + u_i$. Explain the principal of least squares used to obtain estimators of B_1 and B_2. In this context derive the least squares normal equations and estimators of parameters of the population regression function. **[Eco. (H) 2012]**

Q2. Prove that $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$.

Q3. Prove that $\overline{\hat{u}_i} = 0$.

Q4. Prove that $\sum X_i \hat{u}_i = 0$.

Q5. Prove that $\sum \hat{u}_i x_i = 0$, where $\hat{u}_i$ is the residual term and x_i is the deviation of X_i from $\bar{X}$.

Q6. Prove that $\sum \hat{Y}_i \hat{u}_i = 0$, where $\hat{u}_i$ is the residual term and $\hat{Y}_i$ is the estimated value of Y_i.

Q7. Prove that residual term is uncorrelated with independent variable.

Q8. Prove that residual term is uncorrelated with the predicted value.

Q9. Prove that the mean of predicted value of Y_i is always equal to actual mean, *i.e.*, $\bar{Y} = \bar{\hat{Y}}$.

Q10. Prove that $\sum x_i y_i = \sum x_i Y_i = \sum X_i y_i$, where $x_i = X_i - \bar{X}$ & $y_i = Y_i - \bar{Y}$.

Q11. Show that the variance of ordinary least squares estimator of the slope coefficient in the regression model $Y_i = B_1 + B_2 X_i + u_i$, is given by :

$$Var(b_2) = \frac{\sigma_u^2}{\sum_i \left(X_i - \bar{X}\right)^2}$$

Where σ_u^2 is the variance of the population error term. **[Eco. (H) III Sem. 2013]**

Q12. Prove that the least square estimator $\hat{\beta}_2$ is linear, unbiased and consistent. **[BBE 2011]**

Q13. In CLRM, show that OLS estimator for the slope coefficient is linear and unbiased. **[Eco. (H) 2011]**

Q14. Show that the OLS estimators have the property of being linear and unbiased. **[BBE 2007]**

Q15. Prove that the least square estimators have the minimum variance amongst the class of estimators. **[BBE 2007, 2011]**

Q16. Prove that least square estimators are unbiased and efficient in the class of all estimators. **[BBE 2013]**

Q17. Prove that OLS estimators are Best Linear Unbiased Estimators (BLUE).

Q18. Prove BLUE properties of least square estimator of *b*. **[BBE 2008]**

Q19. State and Prove Gauss Markov Theorem.

Q20. State and prove the Gauss Markov theorem for the slope coefficient in the classical linear regression model. **[Eco. (H) III Sem. 2014]**

Q21. Prove that OLS estimators are linear estimates of population parameters.

Q22. Prove that OLS estimators are unbiased estimates of population parameters.

Q23. Derive the variances and standard errors of intercept and Slope terms of a two variable linear trend equation.

Q24. Prove that the OLS estimators are efficient estimators of population parameters.

Q25. Derive the numerical properties of OLS estimators and the regression line. **[BBE 2007]**

Q26. Prove that $Cov\left(\hat{\beta}_1, \hat{\beta}_2\right) = -\overline{X}\, Var\left(\hat{\beta}_2\right)$. **[BBE 2011]**

Q27. Derive the coefficient of determination (r^2).

Q28. Show that all the following formulas to compute r^2 are equivalent :

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum y_i^2} = \frac{\sum \hat{y}_i^2}{\sum y_i^2} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} = \frac{\left(\sum y_i \hat{y}_i\right)^2}{\left(\sum y_i^2\right)\left(\sum \hat{y}_i^2\right)}$$

Q29. Show that : $r^2 = \frac{(\Sigma y_i \hat{y}_i)^2}{(\Sigma y_i^2)(\Sigma \hat{y}_i^2)}$ **[Eco. (H) 2011]**

Q30. For the two variable regression model $Y_t = B_1 + B_2 X_t + u_t$, Show that :

(i) $r^2 = \frac{\left(\sum y_i \hat{y}_i\right)^2}{\left(\sum y_i^2\right)\left(\sum \hat{y}_i^2\right)}$ (ii) $\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$

[Eco. (H) III Sem. 2014]

Q31. Show that the coefficient of determination r^2, in a simple regression $Y_t = B_1 + B_2 X_t + u_t$, is obtained using the formula :**[Eco. (H) III Sem. 2017(ER)]**

$$r^2 = \frac{b_2^2 \sum x_i^2}{\sum y_i^2}$$

Where, b$_2$ is the ordinary least squares (OLS) estimator of B$_2$; and x$_i$ and y$_i$ denote the deviation of X and Y observations from their respective means.

Q32. Prove that : $\sum \hat{u}_i = n\bar{Y} - n\hat{\beta}_1 - n\hat{\beta}_2\bar{X}$.

Q33. You propose to study the relationship $PT_i = \alpha + \beta.RD_i + u_i$, where PT is the number of patents applications field during a given period and RD is the expenditures on research and development as ratio of gross domestic product. The error term u_i satisfies all the assumptions made by the classical linear regression model. What are the properties of the least squares estimators $\hat{\beta}$. Prove any two of these properties. **[Eco. (H) 2010]**

Q34. Consider two regression models :

Regression Model A : $Y_t = B_1 + B_2X_t + u_t$

Reverse Regression Model B : $Y_t = B^*_1 + B^*_2X_t + u^*_t$

If b_2 and b^*_2 are the OLS estimators of B_2 and B^*_2 respectively and r is the sample correlation between X and Y, show that $b_2.b^*_2 = r^2$. **[Eco. (H) 2013]**

Q35. For the model $Y_i = \beta_1 + u_i$, Given that all the CLRM assumptions are satisfied, use OLS to find the estimator of β_1. Show that this estimator can be decomposed into the true value plus a linear combination of the disturbance term in the sample. Also demonstrate that this estimator is an unbiased estimator of β_1.

[Eco. (H) IV Sem. 2015]

Theory Questions

Q1. Describe a population Regression Function. Explain the significance of the stochastic disturbance term in econometric analysis? **[BBE 2011]**

Q2. On what factors does the precision of OLS estimators depend? Find out the expression for the covariance between $\hat{\beta}_1$ and $\hat{\beta}_2$. **[BBE 2011]**

Q3. Describe the ten assumptions of Classical Linear Regression Model (CLRM) for the OLS estimators. **[BBE 2007, 2011]**

Q4. List down the assumptions of the CLRM. Bring out clearly the difference between BLUE and BUE. **[BBE 2011]**

Q5. Explain the Error term in regression analysis.

Or

Write explanatory short note on : Stochastic Error term [BBE III Sem. 2012]

Ans. A Stochastic error term is a term that is added to a regression equation to introduce all of the variation in Y that cannot be explained by the included Xs. It is, in effect, a symbol of the econometrician's ignorance or inability to model all the movements of the dependent variable.

A **statistical error** is the amount by which an observation differs from its expected value, the latter being based on the whole population from which the statistical unit was chosen randomly. For example, if the mean height in a population of 21-year-old men is 1.75 meters, and one randomly chosen man is 1.80 meters tall, then the "error" is 0.05 meters; if the randomly chosen man is 1.70 meters tall, then the "error" is −0.05 meters. The expected value, being the mean of the entire population, is typically unobservable, and hence the statistical error cannot be observed either.

In stochastic form the regression model can be represented as $Y_i = \beta_1 + \beta_2 X_i + u_i$, where the deviation u_i is an unobservable random variable taking positive or negative values. Technically, u_i is known as the **stochastic disturbance** or **stochastic error term.**

Q6. What is the role of the stochastic error term u_i in regression analysis? Why do we employ the normality assumption for u_i ? **[BBE 2011]**

Q7. Suppose for a regression $Y_i = \alpha + \beta X_i + \in_i$, you are told that the error term $\in_i$ are not normally distributed. What can you say about the sampling distribution of $\hat{\beta}$. Explain. **[Eco. (H) 2010]**

Q8. How do you test for normality of error terms in the population regression function using Jarque-Bera test. **[Eco. (H) IV Sem. 2015]**

Q9. Write a short note on Jarque-Bera Test of normality. **[BBE 2009, 2011]**

Q10. Consider the population regression function $Y_i = B_1 + B_2 X_i + u_i$. The stochastic error term u_i is assumed to follow the normal distribution. What is the rationale behind this assumption? What is the role of this assumption in determining the sampling distribution of the ordinary least squares estimators? **[Eco. (H) III Sem. 2013]**

Q11. Write short notes on the following : **[BBE 2014]**
(i) JB test.
(ii) SRF versus PRF.

Q12. Write Short note on the following :
(i) Sample Regression Function,
(ii) Population Regression Function. **[BBE 2011]**

Q13. Differentiate between Population Regression Function (PRF) and Sample Regression Function SRF. [BBE 2007, 2008, 2009, 2011]

Ans. The Following are the major Differences in population regression function and sample regression function :

S. No.	**Population Regression Function**	**Sample Regression Function**
1	The population regression function gives the average or mean value of the dependent variable corresponding to each value of the independent variable.	Each sample produces its own scatter plot. Through this sample scatter plot, we can plot a sample regression line. The SRF characterizes this line.
2	PRF is based on population data as a whole.	SRF is based on Sample data
3	PRF curve or line is the locus of the conditional mean/ expectation of the independent variable Y for the fixed variable X in a sample data.	SRF shows the estimated relation between dependent variable Y and explanatory variable X in a sample.
4	We can draw only one PRF line from a given population.	We can Draw one SRF for each sample from that population, and there can be many samples.
5	PRF is of the form :	SRF is of the form :

	$Y_i = \beta_1 + \beta_2 X_i + u_i$	$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$
6	PRF may exist only in our conception and imagination.	SRF exists in reality based upon the given sample.

Q14. Ordinary least square estimates of the slope coefficients in a simple regression model will be more precisely estimated with a smaller variance if the X values are close to their sample means. True/False. Explain. **[Eco. (H) 2012]**

Q15. Comment whether the following statements are true or false. Give reasons or explanation for the same in 2-3 sentences. **[BBE III Sem. 2012]**

(i) In ordinary least squares the attempt is to minimize the sum of residuals.

(ii) Regression line passes through the sample means of X and Y.

(iii) Sufficient variation in ‘X’ values is desired for greater precision of the estimators.

Q16. The confidence interval can be built up either using normal distribution or students ‘*t*’ distribution. Build a confidence interval for beta using both these distributions. When would a researcher use a ‘*t*’ and not a normal distribution to build a confidence interval. **[BBE III Sem. 2012]**

Q17. In a simple linear regression model, is it possible that all the observed (actual) values of the dependent variable Y_i lie above the estimated regression line obtained by the method of least squares? Justify your answer. **[Eco. (H) II Sem. 2013]**

Q18. You are told that monthly wages, W (in Rupees) earned by a person depends on his age A (in years). Write an appropriate model to study the effect of age on monthly wages. **[Eco. (H) 2009]**

Q19. A car rental company charges a fixed amount of Rs. 400 per day in addition to Rs. 20 per km for renting a car. Let Y be the total charge payable for a car for a day and X be the km.’s driven. **[Eco. (H) II Sem. 2013]**

(i) Is the relationship between X and Y exact or non exact? Explain.

(ii) What do you expect the coefficient of determination to be? Why?

Q20. Two individuals fit earnings function relating Earnings (E) to Years of Schooling (S). The first individual does it correctly and obtains the result as :

$$\hat{E} = -13.93 + 2.46S$$

The second individual makes a mistake and regresses S on E obtaining the following result :

$$\hat{S} = 12.29 + 0.07E$$

From this result the second individual derives :

$$\hat{E} = -175.57 + 14.29S$$

Explain why this equation is different from that fitted by the first individual.

Q21. A researcher has international cross-sectional data on aggregate wages (W), aggregate profits (P) and aggregate income (Y), for a sample of n countries. By, definition,

$$Y_i = W_i + P_i$$

The regressions

$$\hat{W}_i = a_1 + a_2 Y_i \qquad \text{and} \qquad \hat{P}_i = b_1 + b_2 Y_i$$

are fitted using OLS regression analysis. Show that the regression coefficients will automatically satisfy the following equations :

$a_2 + b_2 = 1$ and $a_1 + b_1 = 0$

Q22. What are the properties of a good estimator? **[BBE 2013]**

Q23. State whether the following statement is True or False. Give reasons for your answer : **[Eco. (H) III Sem. 2013]**

In the regression model $Y_i = B_1 + B_2X_i + u_i$, suppose we obtain a 95% confidence interval for B_2 as (0.1934, 1.8499). We can say that the probability is 95% that this interval includes the true B_2.

Q24. Comment on the following. Give reasons in support of your comment.**[BBE 2014]**

(a) Testing the significance of the slope coefficient in a 2 variable linear regression model is the same as testing the overall significance of the model.

(b) $Y_i = \beta_1 + \beta_2^3 X_i + u_i$ and $Y_i = \beta_1 + \beta_2 (1/X_i) + u_i$ are linear regression models and therefore meet all the assumptions of CLRM (Investigate Separately.

(c) The stochastic error term is irrelevant in the regression analysis as its mean value is always zero.

Q25. Is the following statement correct? Justify your answers carefully and provide proofs wherever necessary : **[Eco. (H) III Sem. 2012]**

If you choose a higher level of significance, a regression coefficient is more likely to be significant.

Q26. State whether the following statement is True or False. Give reasons for your answer : **[Eco. (H) III Sem. 2014]**

In a two variable PRF, if the slope coefficient β_2 is zero, the intercept β_1 is estimated by the sample mean.

Q27. Are the following statements correct? Justify your answers carefully and provide proofs wherever necessary : **[Eco. (H) IV Sem. 2015]**

(a) If we multiply both Y and X by 1000 and re-estimate the regression, the slope coefficient will get multiplied by 1000.

(b) In a simple regression model, at a given level of significance, it is not necessary to perform both t-test of estimated slope coefficient and F-test for the goodness of fit.

Numerical Questions

Meaning of Linearity

Q1. State whether the following models are linear regression models :

(a) $Y_i = \beta_1 + \beta_2 \left(\dfrac{1}{X_i} \right) + u_i$ (b) $Y_i = \beta_1 + \beta_2 \ln(X_i) + u_i$

(c) $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$ (d) $Y_i = e^{\beta_1 + \beta_2 X_i + u_i}$

(e) $Y_i = \beta_1 - \beta_2^3 X_i + u_i$ **[BBE 2007]**

Ans. [(a) LIP, (b) LIP, (c) LIP, (d) LIP, (e) LIV]

Q2. Determine whether the following models are linear in the parameters, or the variables, or both. Which of these models are linear regression models?

S. No.	**Model**	**S. No.**	**Model**
(i)	$\ln Y_i = \beta_1 + \beta_2 \ln(X_i) + u_i$,	(ii)	$Y_i = \beta_1 + \frac{1}{\beta_2} X_i + u_i$
(iii)	$Y_i = \beta_1 + \beta_2^2 X_i + u_i$,	(iv)	$\ln Y_i = \beta_1 - \beta_2 \left(\frac{1}{X_i}\right) + u_i$,
(v)	$Y_i = e^{\beta_1 - \beta_2 X_i + u_i}$,	(vi)	$Y_i = \beta_1 - \beta_2^3 X_i + u_i$

Ans. (i) LIP, (ii) LIV, (iii) LIV, (iv) LIP, (v) LIP, (vi) LIV

Q3. Determine whether the following models are linear in parameters or variables or both. Which of these models are linear regression models :

(a) $Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i}\right) + u_i$ (b) $Y_i = \beta_1 + \beta_2 \ln(X_i) + u_i$

(c) $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$ (d) $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i$

Calculate the elasticity for all the cases. **[BBE III Sem. 2012]**

Ans. (a) LIP, (b) LIP, (c) LIP, (d) None

Formulating Regression Equation

Q4. You are given the following data on X and Y.

X	1	2	3	4	5
Y	3	5	7	14	11

(i) Obtain the estimated regression equation using ordinary least squares when Y is regressed on X with an intercept term.

(ii) Prepare the ANOVA table for this data. **[Eco. (H) 2009]**

[**Ans. :** $\hat{Y} = 0.5 + 2.5X$]

Q5. The following sample is obtained for the random variable X and Y :

X	1	2	3	4	5
Y	2	6	10	16	16

Compute a regression line with Y being the explained variable. Find the fitted values of Y and the corresponding error/ residual terms. **[Eco. (H) III Sem. 2014]**

[**Ans. :** $\hat{Y} = 0.5 + 0.25X$]

Q6. From the following hypothetical data on Weekly Family Consumption Expenditure (Y) and Weekly Family Income (X) fit a two variable linear regression model $Y_i = \beta_1 + \beta_2 X_i + u_i$

Y_i	70	65	90	95	110	115	120	140	155	150
X_i	80	100	120	140	160	180	200	220	240	260

Also find standard errors of β_1, β_2 and coefficient of determination.

[**Ans. :** $Y_i = 24.4545 + 0.5091X_i + u_i$, $SE(\beta_1) = 6.4138$, $SE(\beta_2) = 0.0357$, $\hat{\sigma}^2 = 42.1591$, $r^2 = 0.9621$]

Q7. Fit the linear regression $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ for the following data :

X	-4	-3	-2	-1	0	1	2	3	4
Y	10	20	30	40	50	60	70	80	90

Also find the variance and standard errors of intercept and slope coefficients.

[**Ans. :** $\hat{Y} = 50 + 10X$, $SE(\hat{\beta}_1) = 0$, $SE(\hat{\beta}_2) = 0$, $r^2 = 1$]

Q8. Given $\sum Y = 2220$, $\sum X = 3400$, $\sum XY = 411000$, $\sum X^2 = 644000$, $\sum Y^2 = 264200$, n = 10. Obtain the OLS estimators, b_1 (intercept) and b_2 (slope) for the two variable model $Y_i = A + B X_i + u_i$. **[Eco (H) III Sem 2017(ER)]**

Q9. You are given the following data based on 10 pairs of observations on Y and X :

$\sum X_i = 1700$, $\sum Y_i = 1110$, $\sum X_i Y_i = 205,500$, $\sum X_i^2 = 322,000$, $\sum Y_i^2 = 132,100$

Suppose the assumptions of the simple two variable CLRM are fulfilled, obtain

(i) OLS estimators, b_1 and b_2.

(ii) Standard errors of these estimators.

(iii) What is the value of r^2. **[Eco. (H) III Sem. 2014]**

[**Ans. :** $\hat{\beta}_1 = 24.453, \ \hat{\beta}_2 = 0.5091$]

Q10. The true linear relation explaining the height of the oldest son (Y) and the height of the father (X), both in inches, is given by Y = 35.82 + 0.478X with variance of the population disturbance term as 2.25.

(i) A sample of 5 fathers has heights of 64, 66, 68, 70 and 72 inches. Calculate the standard deviation of the estimated slope coefficient, $\hat{\beta}_1$.

(ii) A sample of 7 fathers is drawn up with heights of 65, 66, 67, 68, 69, 70 and 71, would this larger sample be preferable to the one chosen in (i) above? Why? **[Eco. (H) III Sem. 2014]**

Interpretation of Regression Equations

Q11. Suppose someone has presented the following regression results for your consideration :

$$\hat{Y}_t = 2.6911 - 0.4795X_t$$

Where,

Y : Coffee consumption in cups per person per day

X : retail price of coffee (` per cup)

t : time period

(i) Is this a time series or cross sectional regression.

(ii) Sketch the regression line.

(iii) What is the interpretation of intercept term in this example? Does it make economic sense?

(iv) Interpret the slope coefficient.

Q12. From the data on GNP and four definitions of the money stock for the United states for 1970-1983. Regressing GNP on the various definitions of money we obtain the following models :

(I) $\hat{GNP} = -787.4723 + 8.0863 M_{1t}$ $r^2 = 0.9912$

$(77.9664) \quad (0.2197)$

(II) $\hat{GNP} = -44.0626 + 1.5875 M_{2t}$ $r^2 = 0.9905$

$(61.0134) \quad (0.0448)$

(III) $\hat{GNP} = 159.1366 + 1.2034 M_{3t}$ $r^2 = 0.9943$

$(42.9882) \quad (0.0262)$

(IV) $\hat{GNP} = 164.2071 + 1.0290 L_t$ $r^2 = 0.9938$

$(44.7658) \quad (0.0234)$

Note : The figures in parentheses are the estimated standard errors.

Definitions :

M_1 = currency + demand deposits + travelers cheques & other chequable deposits
$M_2 = M_1$ + savings and small deposits
$M_3 = M_2$ + large time deposits
$L = M_3$ + other liquid assets

Determine which of these definitions of money seems closely related to nominal GNP in USA? Also determine which of these money measures is a better target if *Fed* wants to control the money supply in USA?

Hypothesis Testing and Confidence Intervals of OLS Estimators

Q13. Given the following summary results for 6 pairs of observations on the dependent variable Y and the independent variable X, calculate the 95% confidence interval for the true regression slope coefficient β₁. **[Eco. (H) II Sem. 2013]**

$$\sum_{i=1}^{6} X_i = 90; \quad \sum_{i=1}^{6} Y_i = 10.5; \quad \sum_{i=1}^{6} X_i^2 = 1694; \quad \sum_{i=1}^{6} Y_i^2 = 20.29; \quad \sum_{i=1}^{6} X_i Y_i = 181.1$$

Q14. Suppose that a simple linear regression of the fuel consumption rate in miles per gallon (Y) on car weight (X) has been performed using 32 observations. Suppose that the least-squares estimates for the intercept and slope coefficients are $\hat{\beta}_0 = 68.17$ and $\hat{\beta}_1 = -1.112$ respectively, with sample standard deviation of the residuals being, $\hat{\sigma} = 4.281$. Other useful statistics are $\bar{x} = 30.91$, and $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 2054.8$

(i) What should be the predicted fuel consumption rate for a new observation with car weight of 24?

(ii) Find a 95% confidence interval for $\hat{\beta}_1$. **[Eco. (H) III Sem. 2014]**

Q15. Let the following data based on the 10 pairs of observations of X and Y, which is $\sum Y_i = 112$, $\sum X_i = 168$, $\sum X_i Y_i = 211$, $\sum X_i^2 = 3240$, $\sum Y_i^2 = 1330$

Assume that it follows all the usual classical assumptions of CLRM, then obtain

(i) $\hat{\beta}_1$ and $\hat{\beta}_2$, (ii) Standard errors of these estimators,

(iii) r^2, (iv) Establish 95% confidence intervals for $\hat{\beta}_1$ and $\hat{\beta}_2$.

Q16. A linear regression model was estimated using the OLS method for a sample of 15 observations. In order to test the hypothesis that there is no relationship between random variables X and Y the following statistics were calculated :

$\hat{\beta}_0 = 0.9645$, $\hat{\beta}_1 = 1.6699$, $S_{\hat{\beta}_0} = 0.5262$, $S_{\hat{\beta}_1} = 0.1569$, where $\hat{\beta}_0$ is the intercept coefficient and $\hat{\beta}_1$ is the slope coefficient.

(i) State null and alternative hypotheses.

(ii) What statistic would you use to test the hypotheses, and derive its value?

(iii) Do you reject the null hypothesis at the 0.05 level of significance?

(iv) Interpret your result in (iii) above. **[Eco. (H) III Sem. 2014]**

Q17. Using the following data :

$n = 10$, $\sum Y_i = 5070$, $\sum X_i = 5{,}60{,}000$, $\sum Y_i X_i = 30{,}55{,}50{,}000$, $\sum X_i^2 = 47{,}60{,}00{,}00{,}000$, $\sum Y_i^2 = 26{,}07{,}100$

(i) Fit the linear regression $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$,

(ii) Find $S.E.(\hat{\beta}_1)$ and $S.E.(\hat{\beta}_2)$,

(iii) Find 95% confidence intervals of slope and intercept coefficients,

(iv) Test the significance of slope coefficient at 5%.

[**Ans. :** $\hat{\beta}_1 = 432.4138, \hat{\beta} = 0.0013, S.E.(\hat{\beta}_1) =, S.E.(\hat{\beta}_1)$]

Q18. To reduce crime, the minister has budgeted more money to put more police force in the city. A regression to study the effect of police deployment (Y in ‘000) on the number of reported crimes (X) was studied. From the data pertaining to 8 weeks, following results were obtained. **[BBE 2014]**

$\sum X_i = 76$, $\sum Y_i = 130$, $\sum x_i^2 = 116$, $\sum y_i^2 = 241.5$, $\sum x_i y_i = 155$, RSS = 34.38

Where x_i and y_i are deviations from their respective mean.

(i) Estimate both the regression coefficients interpret the regression equation.

(ii) Test the significance of the slope coefficient at 5% level of significance.

(iii) Calculate the value of coefficient of determination and interpret it.

Q19. For the given regression model answer the questions that follow :

$$\hat{Y}_i = 432.4138 + 0.0013 X_i$$

se (16.9061) (0.000245)

$n = 10$, $r^2 = 0.7849$

Y_i → Marks Obtained

X_i → Family Income

(i) Interpret the regression model.

(ii) Find 95% confidence intervals of slope and intercept coefficients,

(iii) Test the significance of slope coefficient at 5%.

Q20. For a sample of 10 observations the following results are obtained :

$\sum Y_i = 1110$, $\sum X_i = 1700$, $\sum x_i y_i = 205500$, $\sum x_i^2 = 322000$, $\sum y_i^2 = 132100$

Where x_i and y_i are deviations from their respective means.

(i) Find the regression coefficients and regression line.

(ii) Test whether regression coefficients are statistically significant at 5% level of significance.

(iii) Calculate and interpret coefficient of determination. **[BBE III Sem. 2012]**

Q21. Consider the following regression

$$\hat{Y}_i = -66.1058 + 0.0650 X_i$$

se (10.7509) ( )

t ( ) (18.73)

n = 20, $r^2 = 0.946$

Fill in the missing numbers. Would you reject the hypothesis that true B_2 is zero at $\alpha = 0.05$? Tell whether you are using a one tailed or two tailed test and why?

Q22. From the data on earnings and education, we obtained the following regression output :

Mean wage earnings	=	0.7437 +	0.6416 Education
SE	=	(0.8355)	( )
t		( )	(9.6536)

n = 528, $r^2 = 0.8944$ **[BBE 2013]**

(i) Fill the missing numbers.

(ii) How would you interpret the coefficient 0.6416?

(iii) Would you reject the hypothesis that education has no effect on wages? Which test would you use?

Q23. Based on the data for a country for the period 1965 to 2006, the following regression results were obtained :

$$\widehat{GNP} = -995.5183 + 8.7503 M_{1t} \qquad r^2 = 0.9488$$

se = ( ) (0.3214)

t = (-3.8258) ( )

where GNP is the gross national product (Rs., in billions) and M_1 is the money supply (Rs., in billions).

(a) Fill in the missing figures.

(b) The monetarists maintain that money supply has a significant positive impact on GNP. How would you test this hypothesis?

(c) What is the meaning of negative intercept?

(d) Suppose M_1 for 2007 is Rs. 750 billions, what is the mean forecast value of GNP for this year.

Q24. A regression was run between personal consumption expenditure (Y) and gross domestic product (X) all measured in billions of dollars for the years 1982 to 1996 and the following results were obtained :

$\hat{Y}_i = -184.0780 + 0.7064X_i$

Se (46.2619) (0.007827)

$r^2 = 0.9984$

(i) What is the economic interpretation of regression coefficients.

(ii) What is MPC.

(iii) Interpret r^2.

(iv) Prepare 95% confidence intervals of regression coefficients.

(v) Test the significance of $\hat{\beta}_1$ and $\hat{\beta}_2$ writing the hypothesis.

Q25. ***The relationship between nominal exchange rate and relative prices***. From the annual observations from 1980 to 1994, the following regression results were obtained, where Y = exchange rate of the German mark to the U.S. dollar (GM/$) and X = ratio of the U.S. consumer price index to the German consumer price index; that is, X represents the relative prices in the two countries:

$\hat{Y}_t = 6.682 - 4.318X_t$ $r^2 = 0.528$

se = (1.22) (1.333)

(a) Interpret this regression. How would you interpret r^2?

(b) Does the negative value of X_t make economic sense? What is the underlying economic theory?

(c) Suppose we were to redefine X as the ratio of German CPI to the U.S. CPI. Would that change the sign of X? And why?

Q26. Based on 48 years annual data on inflation and unemployment, both measured in percentage, the following model was estimated : **[Eco. (H) III Sem. 2013]**

$\Delta\text{inf} = B_1 + B_2\Delta\text{unemp} + u$

Where Δinf and Δunemp are, respectively, inflation and unemployment measured in first difference.

The following table gives partial results of the regression :

	Coefficient	Standard Error
Constant	-0.0781	0.348
Δunemp	-0.8429	0.314

$r^2 = 0.832$

(i) Interpret the slope coefficient.

(ii) Test at 5% level of significance, the claim that there is a one to one tradeoff between inflation and unemployment. Set up the null hypothesis carefully.

Q27. Given the following regression between retail sales of passenger cars (S_i) and real disposable income (X_i) for the years 1962 tom1977 :

$S_i = 5807 + 3.24X_i$

SE (1.634)

$r^2 = 0.22$

(i) Interpret the regression coefficient of X_i.

(ii) Establish a 95% confidence interval for coefficient of X_i.

(iii) Compute t-value under zero null hypothesis and test at 5% level of significance. Which t-test would you use one tail or two tail and why?

Q28. In an economic housing model price (in thousand rupees) of houses was studied to be a function of area (measured in square feet) *SQFT*, that is

$$Price = \beta_1 + \beta_2 SQFT + u_i$$

The results on the basis of a sample of 14 observations were as follows

$$Price = 52.351 + 0.13875 SQFT$$

$$(37.29) \quad (0.019) \qquad\qquad R^2 = 0.82$$

(i) Interpret the result.

(ii) Test the significance of $\hat{\beta}_1$ and $\hat{\beta}_2$ writing the hypothesis.

(iii) Prepare confidence intervals of both the coefficients at 5% level of significance.

(iv) Comment upon the overall significance of the model. **[BBE 2011]**

Q29. Following are the regression results of exports of services (EXPORTS, in lakhs of rupees) over time (t) for the years 1991-2015 (standard errors are mentioned in paranthesis) : **[Eco (H) III Sem 2017(ER)]**

$$\hat{EXPORTS_i} = 3.3890 + 0.00843\, t$$

$$(se) \qquad (0.002) \quad (0.0001)$$

(a) Test the statistical significance of slope coefficients at 5% level of significance.

(b) Compute the instantaneous rate of growth in exports of services.

(c) Can compound rate of growth be computed? If yes, how?

Q30. A regression was run between per capita savings (S) and per capita income (Y) (in rupees) of 10,000 households in a city and the following results were obtained :

$$S_i = 450.03 + 0.67Y_i$$

$$SE \qquad (151.105) \qquad (0.011)$$

$$r^2 = 0.8214$$

(i) What is the economic interpretation of regression coefficients.

(ii) What do you think about the sign of constant term? What can be the possible reason behind it.

(iii) Say something about goodness of fit. Also carry out 't' test for slope coefficients at 1%.

(iv) Reform the above model by stating this is per 100 rupees (instead of per Rs.). what do you think would be impact on slope and intercept.

(v) Prepare 99% confidence intervals. **[BBE 2010]**

Q31. The rational expectation hypothesis claims that expectations are unbiased, *i.e.*, the average predicted value is equal to the actual values of the variable under investigation. A researcher wished to see the validity of this claim with reference to the interest rates on 3 months US treasury bills for 30 quarterly observations. The results of the regression of actual interest rates (r_i) on the predicted interest rates (r_i*) were as follows :

$$\hat{r}_i = 0.0240 + 0.9400\, r_i\, *$$

Se (0.86) (0.14)

Carry out the test(s) to see the validity of the rational expectation hypothesis. (choose α = 5%). Assume all basic assumptions of the classical linear regression model are satisfied. **[Eco. (H) 2009]**

Q32. From the cross sectional data of 55 rural households in India, the following results were obtained :

$$\hat{Y}_i = 94.2087 + 0.4368 X_i$$

variances = (2560.9401) (0.0061)

p = (0.0695) (0.0000)*

$r^2 = 0.3698$

where, * denotes extremely small

Y : Food Expenditure

X : Total Expenditure

(i) Interpret the intercept and slope term.

(ii) Test the statistical significance of intercept and slope term, using p values, clearly specifying the null and alternative hypothesis.

(iii) Interpret r^2.

Q33. Using cross-sectional data on total sales and profits for 27 German companies in 1995, the following model is estimated : **[Eco. (H) III Sem. 2012]**

$$\text{Profits}_i = B_1 + B_2\, \text{Sales}_i + u_i$$

Where

Profits : Total profits in millions of dollars

Sales : Total sales in billions of dollars

The regression results are given below :

	Estimates of Coefficients	**Standard errors**
Constant	83.5753	118.131
Sales	18.4338	4.4463

$r^2 = 0.4074$

(a) Construct a 95% confidence interval for the slope coefficient. What can you say about its statistical significance?

(b) Prove that in a simple regression model with an intercept, the F statistic for goodness of fit of the model is equal to the square of the t statistic for a two sided t test on the slope coefficient. Verify this statement for the regression results given in this question.

(c) Find the forecasted mean profits if annual sales are 25 billion dollars. Explain the concept of a confidence band for true mean profits.

Q34. What is known as the characteristic line of modern investment analysis is simply the regression line obtained from the following model :

$$r_{it} = \alpha_i + \beta_i\, r_{mt} + u_t$$

Where,

r_{it} = the rate of return on the i^{th} security in time t

r_{mt} = the rate of return on the market portfolio in time t

u_t = stochastic disturbance term

In this model β_i is known as the beta coefficient of the i^{th} security, a measure of market (or systematic) risk of a security.

On the basis of 240 monthly rates of return for the period 1956–1976, Fogler and Ganapathy obtained the following characteristic line for IBM stock in relation to the market portfolio index developed at the University of Chicago†:

$$\hat{r}_{it} = 0.7264 + 1.0598 r_{mt} \qquad r^2 = 0.4710$$

$$\text{se} = (0.3001) \quad (0.0728) \qquad df = 238 \qquad F_{1,238} = 211.896$$

(a) A security whose beta coefficient is greater than one is said to be a volatile or aggressive security. Was IBM a volatile security in the time period under study?

(b) Is the intercept coefficient significantly different from zero? If it is, what is its practical meaning?

ANOVA Table

Q35. For a simple linear regression model, $Y_i = B_1 + B_2X_i + u_i$, the following data are given for 22 observations :

$$\bar{X} = 10 \qquad \bar{Y} = 20 \qquad \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = 60 \qquad \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = 100$$

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) = 30$$

(i) Compute the least squares estimates of the slope and intercept parameters.

(ii) Prepare an ANOVA table for the above results.

(iii) Test the hypotheses that $B_2 = 1$ at 5% level of significance. How would your testing procedure change if you were given the true value of the error variance? **[Eco. (H) 2013]**

Q36. From the data on earnings and education, we obtained the following regression output :

Mean wage earnings	=	0.7437 +	0.6416 Education
SE	=	(0.8355)	( )
t		( )	(9.6536)

n = 528, $r^2 = 0.8944$

(i) Fill the missing numbers.

(ii) How would you interpret the coefficient 0.6416?

(iii) Would you reject the hypothesis that education has no effect on wages? Which test would you use?

(iv) Set up the ANOVA table for this example and test the hypothesis that the slope coefficient is zero. Which test would you use?

Q37. You have the following data based on 50 observations : **[BBE 2009]**

$\bar{X} = 4$, $\bar{Y} = 6.2$, $\sum x_i^2 = 800$, $\sum x_i y_i = 490$

Where *x* and *y* are in deviations from respective means

(i) Estimate the linear regression of Y on X,

(ii) Interpret the slope coefficient,

(iii) If $\sum y_i^2 = 396.2$ construct the ANOVA table and calculate R^2.

Q38. Given below is the data for 10 years from the Economic Survey of India :

Year	**Private Final Consumption Expenditure (PFCE) (in Rs. '0000 cr.)**	**GDP (in Rs. '0000 cr.)**
1985-86	43	54
1986-87	43	55
1987-88	45	56
1988-89	48	62
1989-90	51	67
1990-91	53	69
1991-92	54	70
1992-93	55	74
1993-94	57	78
1994-95	61	85

We take PFCE as dependent variable Y and GDP as independent variable X. We are given the following results in terms of deviations : **[BBE 2007]**

$\sum xy = 568$, $\sum x^2 = 966$, $\sum y^2 = 338$

Find :

(i) Marginal Propensity to consume, (ii) ESS, (iii) RSS

(iv) Coefficient of Determination,

(v) Given the standard error of estimated MPC = 0.022. Test the null hypothesis Ho : MPC > 0.5 at 5% level of significance.

(vi) For the above question where does the rejection region lie? Show graphically.

(vii) Construct the ANOVA table for the above data and find the F- statistic.

Jarque Bera Test for Normality of Residuals

Q39. Test the normality of residuals using the following data :

Skewness	1.50555
Kurtosis	6.432967
No. of observations	379

[**Ans. :** JB = 328.9, Reject H$_o$]

Q40. Information was collected on daily changes in rupee (Distribution A) and daily return on nifty (Distribution B) for a six month period (150 days) and following are the summarized results :

	Distribution A	**Distribution B**
Mean	39.29	10.53
Standard Deviation	8.17	8.24
Skewness	0.38	1.78

Kurtosis	2.61	6.24

Determine which of the above distributions is normally distributed clearly specifying the test employed. **[BBE III Sem. 2012]**

[For A : JB = 4.5606, For B : JB = 144.82]

Q41. Explain the steps involved in the Jarque-Bera test for testing the validity of the normality assumption in an empirical exercise. Perform the test for a JB test statistic value equal to 0.8153 at 5% level of significance. **[Eco. (H) III Sem. 2013]**

Q42. State whether the following statement is True or False. Give reasons for or proofs : If Jarque-Bera statistic is computed for a large sample as 7.378, it provides evidence in favour of normality of the error term at 5% level of significance.

[Eco (H) III Sem 2017 (ER)]

Change of Scale

Q43. A regression was run between gross private domestic investment (Y) and gross domestic product (X) all measured in billions of dollars for the years 1988 to 1997 and the following results were obtained :

$Y_t = -1026.498 \quad + \quad 0.3016X_t$

Se (257.5874) (0.0399) $r^2 = 0.8772$

Fit the new regression model by finding both the parameters, their standard errors and coefficient of determination in the following cases :

(i) Both GPDI and GDP are measured in millions of dollors,

(ii) GPDI in billions of dollors and GDP in millions of dollors,

(iii) GPDI in millions of dollors and GDP in billions of dollors.

Ans. : (i) $Y_t = -1026{,}498 \quad + \quad 0.3016X_t$

Se (257,587.4) (0.0399) $r^2 = 0.8772$

(ii) $Y_t = -1026.498 \quad + \quad 0.0003016X_t$

Se (257.5874) (0.0000399) $r^2 = 0.8772$

(iii) $Y_t = -1026{,}498 \quad + \quad 301.6X_t$

Se (257,587.4) (39.9) $r^2 = 0.8772$

Q44. A regression was run between gross private domestic investment (Y) and gross domestic product (X) all measured in billions of dollars for the years 1997 to 2006 and the following results were obtained :

$Y_t = 461.511 \quad + \quad 5.8046X_t$

Se (1331.451) (0.762) $r^2 = 0.8787$

Fit the new regression model by finding both the parameters, their standard errors and coefficient of determination in the following cases :

(i) Both GPDI and GDP are measured in millions of dollors,

(ii) GPDI in billions of dollors and GDP in millions of dollors,

(iii) GPDI in millions of dollors and GDP in billions of dollors.

Ans. (i) $Y_t = 461511 \quad + \quad 5.8046X_t$

Se (1331451) (0.762) $r^2 = 0.8787$

(ii) $Y_t = 461.511 \quad + \quad 0.0058046X_t$

Se (1331.451) (0.000762) $r^2 = 0.8787$

(iii) Y_t = 461511 + 5804.6X_t
Se (1331451) (762) $r^2 = 0.8787$

Q45. Given the estimated regression of savings (Y, in Rs. '000) on income (X, in Rs. '000) : $\hat{Y}_i = -0.0386 + 0.863X_i$, derive what would be the new regression equation if savings and income were both expressed in rupees. Interpret the slope coefficient for the given equation. **[Eco. (H) II Sem. 2013]**

Q46. A researcher has data on the aggregate expenditure on services, Y, and aggregate disposable personal income, X, both measured in $ billion at constant prices, for each of the US states and fits the equation

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

The researcher initially fits the equation using OLS regression analysis. However suspecting that tax evasion causes both Y and X to be substantially underestimated, the researcher adopts two alternative methods of compensating for the under reporting :

(i) The researcher adds $90 billion to the data for Y in each state and $200 billion to the data for X.

(ii) The researcher increases the figures for both Y and X in each state by 10 percent.

Evaluate the impact of the adjustments on the regression results.

Q47. Suppose that you are considering opening a restaurant at a location where traffic volume is 1000 cars per day. To help you decide whether to open the restaurant or not, you collect data on daily sales (in thousands of rupees) and average traffic volume (in hundreds of cars per day) for a random sample of 22 restaurants. You set up your model as : **[Eco. (H) III Sem. 2013]**

$$\text{Sales}_i = B_1 + B_2\text{Atraffic}_i + u_i$$

You know that $\sum X_i Y_i = 17170$, $\sum X_i^2 = 13055$, $\bar{Y} = 32$, $\bar{X} = 22.5$

(i) Obtain the ordinary least square estimator of the slope coefficient and interpret it.

(ii) Estimate the average sales for your potential restaurant location.

(iii) Will the value of coefficient of determination change if you want to change the unit of measurement of sales from thousands rupees to rupees, leaving units of traffic volume unchanged? Explain your answer.

Relation Between F and *t*

Q48. Given the following regression results (t statistics are reported in parentheses)

$$\hat{Y}_i = 16{,}899 - 2978.5X_{2t}$$ **[Eco. (H) 2013]**

(8.51) (-4.72) $R^2 = 0.6149$

Use the relationship between R^2, F and t to find out the underlying sample size.

Q49. You are given the following regression result

Sales(Y_t) = 4.3863 + 1.08132ADV$_t$

t (4.42) (13.99) $r^2 = 0.938$

Find the sample size underlying the result. **[BBE 2014]**

Forcasting

Q50. The following regression equation was estimated for 10 observations on X and Y.

$$\hat{Y}_i = 24 - 0.5X_i\,, \qquad \bar{X} = 170\,, \qquad \sum x_i^2 = 33{,}000\,, \qquad \hat{\sigma}^2 = 42$$

Establish a 95% confidence interval for E(Y/X = 100).

Q51. Based on the data collected on a particular Monday for 13 B. A. (H) Economics second year students we want to estimate the following population regression equation : **[Eco. (H) IV Sem. 2015]**

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

Where :

Y_i : Travelling time (in hrs) for the *i*th student from her home to college.

X_i : Distance from home to college for *i*th student in km.

The sample gave the following values :

$$\sum X_i = 195,\ \sum Y_i = 26,\ \sum X_i^2 = 3050,\ \sum Y_i^2 = 53,\ \sum X_i Y_i = 400$$

Using the above data and assuming that all the CLRM assumptions are satisfied, establish a 95% confidence interval for the predicted mean travelling time when the distance between college and a student's house is 11 km.

[$\hat{\beta}_1 = 0.8, \hat{\beta}_2 = 0.08, \hat{\sigma}^2 = 0.0182$, CI = [1.5462, 1.8138]]

MA Entrance

Q52. In a linear regression of Y on X, changing the units of measurement of the Y variable will affect all of the following except : **[DSE MA Ent. Eco. 2009]**

(a) the estimated intercept parameter

(b) the estimated slope parameter

(c) the Total Sum of Squares for the regression

(d) R squared for the regression