COURSE TITLE: ECONOMETRICS

CODE : ECO 2102

HOURS TAUGHT : HOURS PER WEEK

INSTRUCTOR : NANSAMBA BITIYALI

PURPOSE OF THE COURSE

The course contains the models and methods used to estimate relationships and test hypotheses concerning economic variables. This helps students in their research work.

LEARNING OUTCOMES

By the end of the course unit, students should be able to ;

- Explain the different terms used in econometrics
- State and explain the methodology of econometrics
- Test hypotheses concerning different variables
- Formulate models
- Use different analytical tools

COURSE CONTENT

INTRODUCTION

- Definition of econometrics and other concepts
- Application
- Objectives of econometrics
- Types of econometrics(theoretical and applied econometrics)
- Methodology of econometrics
- Data used in econometrics (time series data, cross sectional data, panel data and pooled data)

CORRELATION ANALYSIS

- Correlation and linearity
- Pearson correlation coefficient
- Spearman's rank correlation coefficient
- Partial correlation

REGRESSION ANALYSIS

- Simple linear regression model
    - Basic assumptions of the regression model
    - Reasons for an error term

- Estimation of parameters
- Hypotheses testing

- Multiple regression
- Non linear regression

DATA ANALYSIS

- Types of tests
- Decision rule
- Errors and estimates
- Autocorrelation
- Multicollinearity
- Solving simultaneous equations

MODE OF DELIVERY

- Lecture method
- Group work
- Reading assignments

INSTRUCTIONAL MATERIALS

- White board and markers
- Computers

COURSE ASSESSMENT

- Continuous assessment test 20%
- Assignments/ group work 20%
- Final examination 60%

REFERENCES

1. **A**.K. Sharma: Elementary Statistics
2. Dominick S. and Derrick R: Statistics and Econometrics 2nd Edition
3. Freund and Williams: Modern Business Statistics
4. Robert D. Mason et al: Statistical Techniques in Business and Economics, Tenth edition.
5. Dimitrios Asteriou, Stephen hall (2007): applied econometrics; a model approach using views and Microfit.
6. Gujarat (2008), basic econometrics

ECONOMETRICS (ECO 2102) NOTES

CHAPTER ONE

1.0 INTRODUCTION:

1.1WHAT IS ECONOMETRICS?

Is concerned with the testing the theoretical propositions embodied in relations and with estimating the parameters involved. Econometrics is the science that combines economic theory with economic statistics and tries by mathematical and statistical methods to investigate the empirical support of the general law established by economic theory.

It is a composition of economics, mathematics and statistics. Where economics is for developing a hypothesis, mathematics is for model building in a mathematical form and statistics deals with using statistical techniques to analyse the economic model, to estimate the unknown parameters of the model and using the estimates for statistical inference.

Literally interpreted, *econometrics* means "economic measurement." Although measurement is an important part of econometrics, the scope of econometrics is much broader, as can be seen from the following quotations: Econometrics, the result of a certain outlook on the role of economics, consists of
the application of mathematical statistics to economic data to lend empirical support to the models constructed by mathematical economics and to obtain numerical results.

Econometrics may be defined as the quantitative analysis of actual economic phenomena based on the concurrent development of theory and observation, related by appropriate methods of inference.

Econometrics may be defined as the social science in which the tools of economic theory, mathematics, and statistical inference are applied to the analysis of economic phenomena.
Econometrics is concerned with the empirical determination of economic laws.

I.2 WHY A SEPARATE DISCIPLINE?

As the preceding definitions suggest, econometrics is an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. Yet the subject deserves to be studied in its own right for the following reasons.

- Economic theory makes statements or hypotheses that are mostly qualitative in nature. For example, microeconomic theory states that, other things remaining the same, a reduction in the price of a commodity is expected to increase the quantity demanded of that commodity. Thus, economic theory postulates a negative or inverse relationship between the price and quantity demanded of a commodity. But the theory itself does not provide any numerical measure of the relationship between the two; that is, it does not tell by how much the quantity will go up or down as a result of a certain change in the price of the commodity. It is the job of the econometrician to provide such numerical estimates. Stated differently, econometrics gives empirical content to most economic theory.
- The main concern of mathematical economics is to express economic theory in mathematical form (equations) without regard to measurability or empirical verification of the theory.
- Econometrics, mainly interested in the empirical verification of economic theory. The econometrician often uses the mathematical equations proposed by the mathematical economist but puts these equations in such a form that they lend themselves to empirical testing. And this conversion of

mathematical into econometric equations requires a great deal of ingenuity and practical skill.

- Economic statistics is mainly concerned with collecting, processing, and presenting economic data in the form of charts and tables. These are the jobs of the economic statistician. It is he or she who is primarily responsible for collecting data on gross national product (GNP), employment, unemployment, prices, etc. The data thus collected constitute the raw data for econometric work. But the economic statistician does not go any further,not being concerned with using the collected data to test economic theories.

1.3 OBJECTIVES/ GOALS OF ECONOMETRICS

a) To judge the validity of economic theory.
b) To supply the numerical estimates of the coefficients of the economic relationships that may be used for sound economic policies.
c) To forecast the future values of the economic magnitude with a certain degree of probability.

1.4 CATEGORIES OF ECONOMETRICS

It is distinguished into two categories;

i. Theoretical econometrics: deals with the development of the appropriate methods for measuring economic relationships described by econometric models. These methods may be classified into two groups;
   - Single equation techniques (simple regression analysis) which are applied to one relation at a time.
   - Simultaneous equation techniques (multiple regression) which are applied to all relationships of the model simultaneously.

Theoretical econometrics is concerned with spelling out the assumptions of the above methods, their properties and what happens when one or more of the assumptions of the methods are not fulfilled.

ii. Applied econometrics: describes the practical value of econometric research. It deals with the application of econometric techniques developed in theoretical econometrics to different fields of economic theory for its verification and forecasting. Applied econometrics makes it possible to obtain numerical results from studies that are of great importance to planners.

I.5 METHODOLOGY OF ECONOMETRICS

How do econometricians proceed in their analysis of an economic problem? That is, what is their methodology? Although there are several schools of thought on econometric methodology, we present here the **traditional** or **classical** methodology, which still dominates empirical research in economics
and other social and behavioral sciences.

The traditional econometric methodology involves the following steps:

1. Statement of theory or hypothesis.
2. Specification of the mathematical model of the theory
3. Specification of the statistical, or econometric, model
4. Obtaining the data
5. Estimation of the parameters of the econometric model
6. Hypothesis testing
7. Forecasting or prediction
8. Using the model for control or policy purposes.

To illustrate the preceding steps, consider the well-known Keynesian theory of consumption.

1. Statement of Theory or Hypothesis

Keynes stated:The fundamental psychological law . . . is that men [women] are disposed, as a

rule and on average, to increase their consumption as their income increases, but not as much as the increase in their income. In short, Keynes postulated that the **marginal propensity to consume (MPC),** the rate of change of consumption for a unit (say, a dollar) change in income, is greater than zero but less than 1.

2. Specification of the Mathematical form of the theory

Although Keynes postulated a positive relationship between consumption and income, he did not specify the precise form of the functional relationship between the two. For simplicity, a mathematical economist might suggest the following form of the Keynesian consumption function:

$Y = \beta 1 + \beta 2X$, $0 < \beta 2 < 1$ **,**where Y = consumption expenditure and X = income, and where $\beta 1$ and $\beta 2$,

known as the **parameters** of the model, are, respectively, the **intercept** and **slope** coefficients.

The slope coefficient $\beta 2$ measures the MPC. A model is simply a set of mathematical equations.

If the model has only one equation, as in the preceding example, it is called a **single-equation model,** whereas if it has more than one equation, it is known as a **multiple-equation model .T**he variable appearing on the left side of the equality sign is called the **dependent variable** and the variable(s) on the right side are called the **independent,** or **explanatory,** variable(s). Thus, in the Keynesian consumption function above ; consumption (expenditure) is the dependent variable and income is the explanatory variable.

3. Specification of the Econometric Model

This involves identifying the variables to be included and specifying the variable form.

The purely mathematical model of the consumption function given in the function described earlier is of limited interest to the econometrician, for it assumes that there is an *exact* or *deterministic* relationship between consumption and income. But relationships between economic variables are generally inexact.

Thus, if we were to obtain data on consumption expenditure and disposable income of a sample and plot these data on a graph paper with consumption expenditure on the vertical axis and disposable income on the horizontal axis, we would not expect all observations to lie exactly on the straight line of because, in addition to income, other variables affect consumption expenditure. For example, size of family, ages of the members in the family, family religion, etc., are likely to exert some influence on consumption. To allow for the inexact relationships between economic variables, the econometrician would modify the deterministic consumption function as follows:

$Y = \beta 1 + \beta 2X + u.$

Where *u*, known as the **disturbance,** or **error, term,** is a **random (stochastic) variable** that has well-defined probabilistic properties. The disturbance term *u* may well represent all those factors that affect consumption but are not taken into account explicitly.

4. Obtaining Data

Obtain data on the variables by identifying the right data sources

5. Estimation of the parameters of the Econometric Model

Involves selection of the estimation techniques and estimating the parameters.

6. Hypothesis Testing

Assuming that the fitted model is a reasonably good approximation of reality, we have to develop suitable criteria to find out whether the estimates obtained are in accordance with the expectations of the theory that is being tested.

Such confirmation or refutation of economic theories on the basis of sample evidence is based on a branch of statistical theory known as **statistical inference (hypothesis testing).**

7. Forecasting or Prediction

If the chosen model does not refute the hypothesis or theory under consideration, we may use it to predict the future value(s) of the

dependent, or **forecast, variable** *Y* on the basis of known or expected future value(s) of the explanatory, or **predictor, variable** *X*.

8. Use of the Model for Control or Policy Purposes

Estimated model may be used for policy purposes by appropriate fiscal and monetary policy needs.

1.6 DATA USED IN ECONOMETRICS

The success of any economic analysis depends on the availability of the appropriate data used.

a) Time series data: a time series is a set of observations on the values that a variable takes at different times.
b) Cross sectional data: data on one or more variables collected at the same point in time e.g census , surveys on consumer expenditure.
c) Pooled data (combined data): this comprises of both time series and cross sectional data.
d) Panel/ longitudinal/ panel data: special type of pooled data in which the same cross sectional unit (e,g family or firm) is surveyed over time.

1.7 RELEVANT TERMS

- Dependent variable: is the explained variable or predictand or reggressand or response variable or outcome or endogenous or controlled variable. Is the variable on which data is collected.
- Explanatory variable. Is the one used to explain the dependent variable. It is referred to as the independent variable or predictor or regressor, stimulus, exogenous, covariate or control variable.
- **Population**- All subjects or objects possessing some common specified characteristic. The population in a statistical investigation is arbitrarily defined by naming its unique properties
- ***Sample*** - A smaller group of subjects or objects selected from a large group (population)

- **Parameters**- A measurable characteristic of a population. A measurable quantity derived from a population, such as population mean or standard deviation
- Statistics (singular –a statistic),**Statistic** - A measure obtained from a sample. It is a measurable quantity derived from a sample, such as the sample mean or standard deviation
- **Variable** - A measurable characteristic. Individual measurements of a variable are called varieties, observations, or cases.
- ***Primary data*** is the data published or used by an organization which originally collects them. The data in the Population Census reports are primary because they are collected, compiled and published by the Population Census Commission. In the natural and social sciences, primary sources are often empirical studies -- research where an experiment was done or a direct observation was made.
- ***Secondary Sources*** is the data published or used by an organization other than the one which originally collected them. You can think of secondary sources as second-hand information.

- In **nominal** measurement the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number 30 is not more of anything than a player with number 15, and is certainly not twice whatever number 15 is.
- In **ordinal** measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0=less than H.S.; 1=some H.S.; 2=H.S. certificate; 3=some college; 4=college transcript; 5 = post collegeIn this measure, higher numbers mean *more* education. But is distance from 0 to 1 same as 3 to 4? Of course the interval between values is not interpretable in an ordinal measure.

- In **interval** measurement the distance between attributes *does* have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).
- **ratio** measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count" variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "...we had twice as many clients in the past six months as we did in the previous six months."
- Random or stochastic variable can take on a set of values positive or negative with a given probability.
- A model is a set of mathematical equations.

CHAPTER TWO

2.0 CORRELATION ANALYSIS

Correlation analysis is a group of techniques used to measure the strength of the relationship between two variables. This relationship can be positive or negative (linear) or non linear.

The scatter diagram is a chart that portrays the relationship between two variables. The values of the independent variable are portrayed on the horizontal axis (X – axis) and the dependent variable along the vertical axis (Y-Axis).

A***dependent variable*** is a variable that is being predicted or estimated *while* an ***independent variable*** is a variable that provides the basis for estimation. It is the predictor variable.

Linear Correlation Coefficient is a measure of the strength of the linear relationship between two sets of variables. It is a measure of the extent to which the points cluster about a straight-line. A measure of correlation between two variables is the Pearson's product moment correlation coefficient or Pearson's r after its founder Karl Pearson. The correlation coefficient is usually designated by the lower case r and may range from $^{-}1.00$ to $^{+}1.00$ inclusive ($^{-}1 \leq r \leq ^{+}1$). A value of $^{-}1.00$ indicates perfect negative correlation. A value of $^{+}1.00$ indicates perfect positive correlation. A correlation coefficient of 0.0 indicates that there is no linear relationship between the two variables under consideration.

The coefficient of correlation requires that both variables be at least of interval scale. The degree of strength of the relationship is not related to the sign (direction – or +) of the coefficient of correlation.

For example, an r value of -0.60 represents the same degree of correlation as +0.60. An r of -0.70 represents a stronger degree of correlation than 0.40. An r of -0.90 represents a strong negative correlation and +0.15 a weak positive correlation.

The Pearson's product moment correlation coefficient or sample correlation coefficient for variables x and y (r) is computed as below;

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)2]\,[n(\sum y^2) - (\sum y)2]}} = b\frac{s_x}{s_y}$$

Where;

n	is the number of paired observations.
ΣXY	is the sum of the products of X and Y.
ΣX	is the X variable summed.
ΣY	is the Y variable summed.
ΣY^2	is the X variable squared and the squares summed.

$(\Sigma X)^2$ is the X variable summed and the sum squared.
ΣY^2 is the Y variable squared and the squares summed.
$(\Sigma Y)^2$ is the Y variable summed and the sum squared.
b is the estimated value from the regression equation
or $r_{xy} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}$
where; $x_i = X_i - \bar{X}$ and $y_i = Y_i - \bar{Y}$

Testing the Significance of the Correlation Coefficient
A test of significance for the coefficient of correlation may be used to determine if the computed r could have occurred in a population in which the two variables are not related. To put it in the form of a question: Is the correlation in the population zero?
For a two-tailed test the null hypothesis and the alternate hypothesis are written as follows:
H_0: p = 0 (The correlation in the population is zero)
H_1: p ≠ 0 (The correlation in the population is different from zero)

The Greek lower case rho, p, represents the correlation in the population. The null hypothesis is that there is no correlation in the population, and the alternate that there is a correlation.

From the way H_1 is stated, we know that the test is two tailed. The alternate hypothesis can also be set as a one-tailed test. It could read "the correlation coefficient is greater than zero." The test statistic follows the t distribution with n - 2 degrees of freedom. And is denoted as below;

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

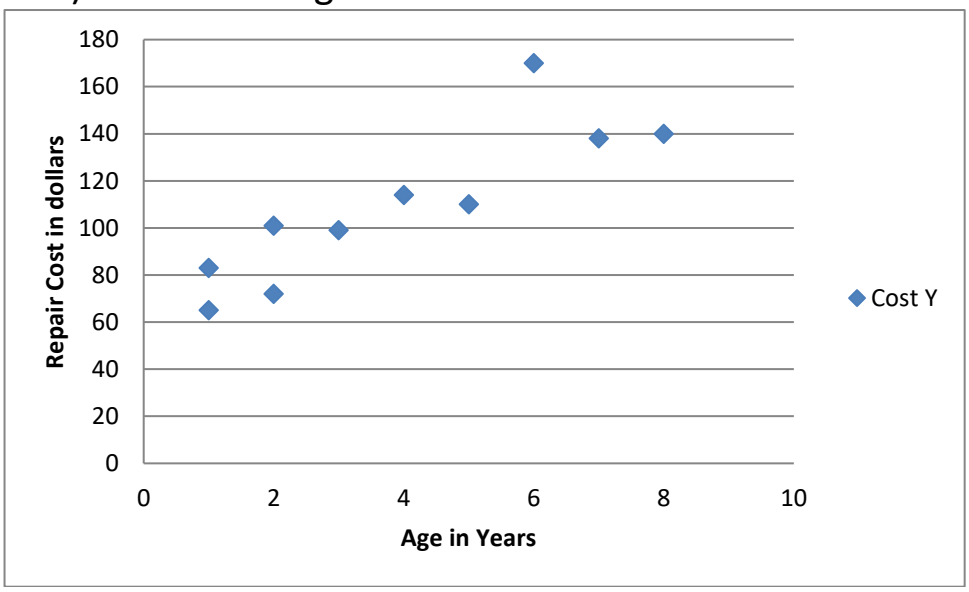
Example It is believed that the annual repair cost for a vehicle is related to its age. A sample of 10 vehicles revealed the results in the table below.

Repair cost in $ (y)	72	99	65	138	170	140	114	83	101	110
Age in years (x)	2	3	1	7	6	8	4	1	2	5

a) Plot these data in a scatter diagram. Does it appear there is a relationship between repair cost and age?
b) Compute the coefficient of correlation
c) Determine at the 0.05 significance level whether the correlation in the population is greater than zero.

Solution

a) A scatter diagram



b) The degree of association between age and repair cost is measured by the coefficient of correlation is computed as,

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)2]\ [n(\sum y^2) - (\sum y)2]}}$$

Y	**X**	**XY**	**X^2**	**Y^2**
72	2	114	4	5,184
99	3	297	9	9,801
65	1	65	1	4,225
138	7	966	49	19,044
170	6	1,020	36	28,900
140	8	1,120	64	19,600
114	4	456	16	12,996
83	1	83	1	6,889
101	2	202	4	10,201
110	5	550	25	12,100
ΣY = 1,092	**ΣX = 39**	**ΣXY = 4,903**	**ΣX^2 = 209**	**ΣY^2 = 128,940**

The totals are inserted into the formula and the value of r computed:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)2]\ [n(\sum y^2) - (\sum y)2]}}$$

$$= \frac{10*4903 - (39*1092)}{\sqrt{[10*209 - 39^2][10*128{,}940 - 1{,}092^2]}} = \frac{6{,}442}{\sqrt{569*96{,}936}} = \frac{6{,}442}{7{,}426.7479} = \mathbf{0.8674}$$

The r of 0.867 suggests a strong positive correlation between the age of this annual repair costs. Implying that as the age of the car increases so does the annual repair cost

The coefficient of determination (r^2) is the square of the coefficient of correlation thus; 0.867^2 = 0.7524 indicating that 75.2 percent of the variation in repair costs can be explained by the variation in the age of the car.

A test of hypothesis is used to determine if the correlation in the population could be zero. In this instance, suppose we want to show that there is a positive association between the variables. H_0: $p \leq 0$ and H_1: $p > 0$

If the null hypothesis is not rejected, it indicates that the correlation in the population could be zero. If the null hypothesis is rejected, the alternate is accepted, and this indicates there is correlation in the population between the two variables and it is positive.

The test statistic follows the Student's (distribution with (n - 2) degrees of freedom. The alternate hypothesis given above specifies a one-tailed test in the positive direction. There are 8 degrees of freedom, (n - 2) = (10 - 2). The critical value for a one-tailed test using the 0.05 significance level is 1.860. The decision rule is to reject the null hypothesis if the computed value of exceeds 1.860. The computed value of t is 4.92, found by using formula

$$\boldsymbol{t} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.8674\sqrt{10-2}}{\sqrt{1-(0.8674)^2}} = \frac{0.8674\sqrt{8}}{\sqrt{0.2476}} = \frac{2.4534}{0.4976} = \mathbf{4.930}$$

Since the computed *t* value (4.930) exceeds the critical value (1.860), the null hypothesis is rejected and the alternate accepted. Concluding that there is a positive association between the age of the vehicle and the annual repair cost. That is the p-value is less than 0.05.

The hypotheses can be tested using the Pearson statistical tables with n-2 degrees of freedom.

Example two

Using the following data results, test the hypotheses that there is no relationship between the variables at 5% level of significance.

$$n = 75, \sum x = 2535, \sum x^2 = 115748, \sum xy = 60168, \sum y = 1650, \sum y^2 = 303646$$

2.1 Spearman's Rank Correlation Coefficient

Is another type of correlation coefficient that is used to measure the strength of linear relationship between two variables x and y. The strength can be measured by converting the two sets of data into ranks and calculating the ordinary correlation coefficient using the rank. The Spearman's Rank Correlation Coefficient is denoted as;

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

Where; d_i is the difference in the corresponding pairs of ranks, n is the number of observations.

The coefficient lies between -1 and +1. The interpretation is the same as that for Pearson's correlation coefficient.

$0.7 \leq c \leq 1$ implies a strong positive correlation

$0.4 \leq c \leq 0.7$ implies a fairly positive correlation

$0 \leq c \leq 0.4$ implies a weak positive correlation

0 implies no correlation

$0 \geq c \geq -0.4$ implies a weak negative correlation

$-0.4 \geq c \geq -0.7$ implies a fairly negative correlation

$-0.7 \geq c \geq -1$implies a strong negative correlation

-1 implies a perfect negative correlation

Example: using the data below for the rankings of countries by two students, determine spearman's rank correlation coefficient and interpret your results.

Country	**$Rank_x$**	**$Rank_y$**	**$d_i = (R_x - R_y)$**	**d_i^2**
A	1	10	-9	81
B	10	4	6	36
C	4	8	-4	16
D	8	5	3	9
E	7	3	4	16

F	2	11	-9	81
G	3	9	-6	36
H	5	6	-1	1
I	6	7	-1	1
J	11	1	10	100
K	9	2	7	49
Total				**$\Sigma d_i^2 = 426$**

So, $r_S = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \times 426}{11(11^2-1)} = 1 - \frac{2556}{1320} = -0.936$

Interpretation: there is a high negative correlation between the ranks of the students.

To test the hypothesis about the relationship, the t-statistic with n-2 degrees of freedom is used as used for Pearson's coefficient.

2.2 PARTIAL CORRELATION

Refers to correlation between the dependent variable y and one of the explanatory variables that influence y given by x_2 while variable x_1 is considered but held constant. If the ordinary correlation coefficients for y and x_1 , y and x_2 and x_1 and x_2 are given as; r_{y1}, r_{y2} and r_{12} respectively, the sample partial correlation coefficient for y and x_2 with x_1 held fixed is given by the formula below;

$$r_{y2.1} = \frac{r_{y2} - r_{y1}r_{12}}{\sqrt{(1 - r_{y1}^2)(1 - r_{12}^2}}$$

The square of the sample correlation coefficient is the sample coefficient of partial determination, which represents the ratio of the unexplained variation to the previously unexplained variation. That is; r_{y2}^2 gives the proportion of the variation in the value of y that was unexplained by a regression line involving only x_1 that can now be explained by including x_2 in the model along with x_2.

EXAMPLE

Given the number of lecture hours missed by 12 students taking an econometrics course , their final examination marks and test marks as below;

student	Final examination (y)	Test marks (x_1)	Hours missed (x_2)
A	85	65	1
B	74	50	7
C	76	55	5
D	90	65	2
E	85	55	6
F	87	70	3
G	94	65	2
H	98	70	5
I	81	55	4
J	91	70	3
K	76	50	1
L	74	55	4

Find and interpret the partial correlation coefficient for y and x_2 when x_1 is held constant.

SOLUTION

From the data ; $\sum x_{1i} = 725, \sum x_{1i}^2 = 44475, \sum x_{1i}y_i = 61685, \sum x_{2i} = 43, \sum x_{2i}^2 = 195, \sum x_{2i}y_i = 3581, \sum x_{1i}x_{2i} = 2540, \sum y_i = 1011.$

$$r_{y1} = \frac{(12)(61685) - (1011)(725)}{\sqrt{[(12)(85905) - 1011^2][(12)(44475) - 725^2]}} = 0.862$$

And r_{y2}= -0.242, r_{12}= -0.349

Therefore; $r_{y2.1} = \frac{-0.242-(0.862)(-0.349)}{\sqrt{[1-(0.862)^2][1-(-0.349^2]}} = 0.124$

Interpretation: the value $r_{y2.1}^2$=0.015 indicates that the addition of $x_2$to the regression equation results in only a 1.5% reduction in the variation of y that is unexplained by a regression line using only x_1. Hence the number of lecture hours missed contributes very little in predicting a student's grade in econometrics.

CHAPTER THREE

3.0 REGRESSION ANALYSIS

To study the relationship between two variables we use two techniques namely; regression and correlation analysis.

3.1 DIFFERENCES BETWEEN CORRELATION AND REGRESSION ANALYSIS

- In regression analysis there is an asymmetry in the way the dependent and explanatory variables are treated.
- In regression the dependent variable is assumed to be statistical, random or stochastic, that is have a probability distribution.
- In regression the explanatory variable is assumed to have fixed values in repeated sampling.
- In correlation analysis, the two variables are treated symmetrically.
- In correlation there is no difference between the dependent and independent variables.
- In correlation both variables are assumed to be random.

3.11 OBJECTIVES OF REGRESSION ANALYSIS

- Prediction of future observations. To estimate the mean value of the dependent variable given the value of the independent variable.
- Assessment of the effect or relationship between explanatory on the response.
- A general description of the data structure.

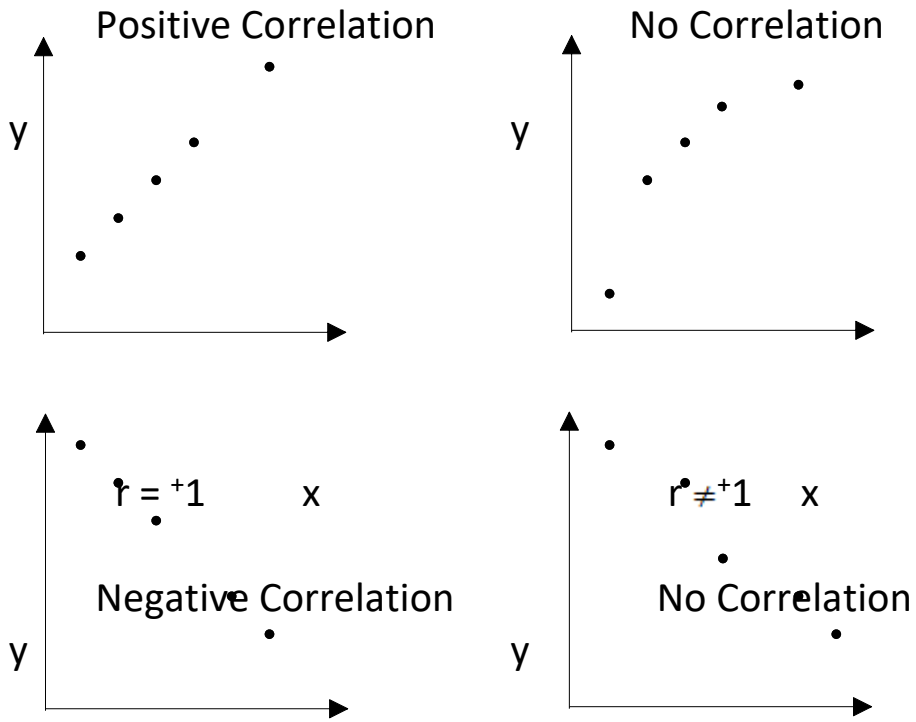
Regression analysis this refers to fitting of a mathematical relationship between two variables say *x* and *y* where one is known “Independent or explanatory or exogenous variable”. And the unknown variable is referred to as the “Dependent or explained or endogenous variable”.

Using the fitted relationship we can make prediction of the dependent variable for any given value of the independent variable. For example; if the relationship between advertising expenses and sales for a given company is given, we can determine the value of sales that may result from spending a given amount of money in advertising. Such information is very useful to a company manager in determining how much the company can spend on advertising. In this case; advertising cost or expenses is the independent variable while the volume of sales is the dependent variable.

Regression analysis involves two steps;

- **Specification;** this relates to determining the nature of the relationship between the explained (dependent) and the explanatory (independent) variables. The possible relationship that may result include;
  - **Linear thus;** $Y = a + bx$. Where; ***a*** is the intercept, ***b*** is the slope of the curve, **Y** is the dependent (explained) variable and ***x*** is the independent (explanatory) variable.
  - **Quadratic thus;** $Y = a + bx + Cx^2$
  - **Exponential thus;** $Y = ab^x$

Specification is facilitated by a scatter diagram; this is the plot of the dependent variable against the independent variable. Although the points may not fall on the same line; it generally shows the overall pattern of the relationship between the two variables thus a careful examination of the scatter diagram will indicate the possible relationship between these variables.



☞ Estimation; this relates to how best you can determine the coefficient of the regression equation specified above

Linear regression line; the linear regression line is the representation of the data plotted. The drawn line that fixes the data very well is called the line of best fit.

Method of least squares; is the method used to find the line of best of fit which is called the regression line. y = *a* + *bx*. Where; *a* is the y intercept, *b* is the slope, *x* is the dependent variable and y is the independent variable. For every point on a scattered diagram there is a

unique value of (y-$\bar{y}$). This is the error term committed in estimating y by $\bar{y}$. By the least square method, we scam to minimize the sum of squares of observation values of the dependent variable from those estimated by the regression line. Thus; Σ(y-$\bar{y}$)^2 should be minimal

Interpretation of regression coefficient

Given the regression equation or line y = *a* + *bx*, the regression coefficients of *a* and *b* can be interpreted as follows; a is the y intercept and it gives an estimate of a dependent variable when the independent variable is zero(*x* = 0). b is the slope of the regression line. It shows the change in y (the dependent variable) that the results from a unit change in *x* (the independent variable). The constant *a* and *b* are denoted by;

y = *a* + *bx*

$\bar{y}$ = *a* + *b*$\bar{x}$

a = $\bar{y}$ -b$\bar{x}$

$$a = \frac{\sum Y}{n} - \frac{b\sum x}{n} = \frac{\sum Y - b\sum x}{n}$$

$$b = \frac{n\sum XY - \sum X\sum Y}{n\sum X^2 - (\sum X)^2} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Example1:Given the bivariate data below fit a regression line of *y* and *x* and hence predict *y* if *x* = 0

x: 1 5 3 2 1 1 7 3

y: 6 1 0 0 1 2 1 5

Solution:

X	***Y***	x^2	***xy***	***y = a + bx***
1	6	1	6	2.5708
5	1	25	5	1.354
3	0	9	0	1.9624
2	0	4	0	2.2666
1	1	1	1	2.5708
1	2	1	2	2.5708
7	1	49	7	0.7456
3	5	9	15	1.9624
$\Sigma x = 23$	$\Sigma x = 16$	$\Sigma x^2 = 99$	$\Sigma xy = 36$	
$\overline{x} = 2.875$	$\overline{y} = 2$			

From $b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} = \frac{(8*36)-(23*16)}{(8*99)-(23)^2} = \frac{-80}{263}$ **- 0.3042**

$a = \overline{y} - b\overline{x}$ = 2 -(-0.3042*2.875) = 2 + 0.874575 = 2.8746

y = *a* + *bx* = 2.875 + -0.3042x

Example 2:Suppose the least squares principle was used to develop an equation expressing the relationship between annual salary and years of work experience. The equation is:

$$y' = \alpha + \beta x$$
$$= 20{,}000 + 500x \text{ (in dollars)}$$

In this example, annual income is the dependent variable,y' and is being predicted on the basis of the employee's years of work experience, *x*, the independent variable. The value of 500, which is *b*, means that for each additional year of work experience the employee's salary increases by $500. Thus, we would expect an employee with 40

years of work experience to earn $5,000 more than one with 30 years of work experience.

What does the 20,000 dollars represent? It is the value for y' when *x* = 0. Recall that this is the point where the line intersects the Y-axis. The values of α and *b* in the regression equation are usually referred to as the **'regression coefficients'**.

3.2 ECONOMETRIC MODELS

In econometrics, all relations between variables can be classified as either deterministic or stochastic. The variables are deterministic if one of the variables can explain the other with certainity, that is y=f(x) is a deterministic relationship between x and y if for each value of x there is only one corresponding value of y.

A relationship is stochastic or non deterministic if for each value of x there is a whole probability distribution of values of y. thus for any value of x, the variable y may assume some specific value or fall within some specific interval with a probability smaller than one and greater than zero.

A stochastic relationship is random in nature and can be derived from the model $y_i = \alpha + \beta x_i + \varepsilon_i$ where y is the dependent variable, x is the independent variable and ε_i is the error/random/disturbance term, α and β are parameters to be determined.

The model $y_i = \alpha + \beta x_i$ is deterministic while adding the error term produces a stochastic relationship. In econometrics we deal with stochastic relations which can be represented using a simple linear regression model of the form $y_i = \alpha + \beta x_i + \varepsilon_i$.

3.21 **BASIC ASSUMPTIONS OF THE MODEL**

These are referred to as the basic classical assumptions and include;

- Normality of the error term
- Error term has zero mean, i.e E(ε_i)=0
- Constant variance (homoscedasticity) i.e E(ε^2)=δ^2
- Non-auto-regression i.e $E(\varepsilon_i \varepsilon_j) = 0 \; for \; i \neq j$.
- The regression model is linear in parameters

- Zero covariance between the error term and the explanatory variable i,e $E(E_i x_i)=0$
- Non-stochastic explanatory variable. The values of x are fixed in repeated samples.
- 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.
- Variability in x values. The x values in a given sample must not all be the same.
- The regression model is correctly specified. Alternatively, there is no specification bias or error in the model used in empirical analysis.
- There is no perfect multicollinearity, that is there is no perfect linear relationship among the explanatory variables.

3.22 **THE SIGNIFICANCE OF THE STOCHASTIC DISTURBANCE TERM**

The disturbance term is a surrogate for all those variables that are omitted from the model but that collectively affect *Y*. The obvious question is: Why not introduce these variables into the model explicitly? Stated otherwise, why not develop a multiple regression model with as many variables as possible? The reasons are many.

1. *Vagueness of theory:* The theory, if any, determining the behavior of *Y* may be, and often is, incomplete. We might know for certain that weekly income *X* influences weekly consumption expenditure *Y,* but we might be ignorant or unsure about the other variables affecting *Y.* Therefore, *the term* may be used as a substitute for all the excluded or omitted variables from the model.

2. *Unavailability of data:* Even if we know what some of the excluded variables are and therefore consider a multiple regression rather than a simple regression, we may not have quantitative information about these. A further difficulty is that variables such as sex, education, and religion are difficult to quantify.

3. *Core variables versus peripheral variables:* Assume in a consumption income example that besides income *X*1, the number of children per family *X*2, sex *X*3, religion *X*4, education *X*5, and geographical region *X*6 also affect consumption expenditure. But it is quite possible that the joint influence of all or some of these variables may be so small and at best nonsystematic or
random that as a practical matter and for cost considerations it does not pay to introduce them into the model explicitly. One hopes that their combined effect can be treated as a random variable.

4. *Intrinsic randomness in human behavior:* Even if we succeed in introducing all the relevant variables into the model, there is bound to be some"intrinsic" randomness in individual *Y*'s that cannot be explained no matter how hard we try. The disturbances, may very well reflect this intrinsic randomness.

5. *Poor proxy variables:* Although the classical regression model assumes that the variables *Y* and *X* are measured accurately, in practice the data may be plagued by errors of measurement.
Consider, for example, Milton Friedman's well-known theory of the consumption He regards *permanent consumption* (*Yp*) as a function of *permanent income* (*Xp*). But since data on these variables are not directly observable, in practice we use proxy variables, such as current consumption (*Y*) and current income (*X*), which can be observable. Since the observed *Y*

and *X* may not equal *Yp* and *Xp*, there is the problem of errors of measurement. The disturbance term may in this case then also represent the errors of measurement.

6. *Principle of parsimony:* we would like to keep our regression model as simple as possible. If we can explain the behavior of *Y* "substantially" with two or three explanatory variables and if
our theory is not strong enough to suggest what other variables might be included, why introduce more variables? Let *the error term* represent all other variables. Of course, we should not exclude relevant and important variables just to keep the regression model simple.

7. *Wrong functional form:* Even if we have theoretically correct variables explaining a phenomenon and even if we can obtain data on these variables, very often we do not know the form of the functional relationship between the regressand and the regressors. Is consumption expenditure a linear (invariable) function of income or a nonlinear (invariable) function? If it is the former, *Yi* = *β*1 + *B*2*Xi* + *Ei* is the proper functional relationship between *Y* and *X,* but if it is the latter, *Yi* = *β*1 + *β*2*Xi* + *β*3*X*2*i*+ E*i* may be the correct functional form. In two-variable models the functional form of the relationship can often be judged from the scatter gram. But in
a multiple regression model, it is not easy to determine the appropriate functional form, for graphically we cannot visualize scattergrams in multiple dimensions.

For all these reasons, the stochastic disturbances assume an extremely critical role in regression analysis.

3.3 **DESIRABLE PROPERTIES OF AN ECONOMETRIC MODEL**

An econometric model is a model whose parameters have been estimated with some appropriate econometric technique. The goodness of an econometric model is judged according to the following desirable properties;

- Theoretical plausibility; the model should be compatible with the postulates of economic theory. It must describe adequately the economic phenomenon which it relates.
- Explanatory ability; it should be able to explain the observations of the actual world.
- Accuracy of the estimates of the parameters; it should approximate as best as possible the true parameters of the structural model. It should possess the desirable properties of unbiasedness, consistency and efficiency.
- Forecasting ability; it should produce satisfactory predictions of future values of the dependent variables.
- Simplicity; it should represent economic relationships with maximum simplicity.

3.4 **ESTIMATION OF PARAMETERS**

There are different methods of estimating the parameters in the regression model which include; Least squares estimation method, maximum likelihood estimation methods, moments etc.

3.31 **LEAST SQUARES ESTIMATION METHOD (LSE)**

The principle of LSE involves minimizing the sum of squared deviations of the observed values from their mean. Given the model $Y_i = \alpha + \beta x_i + \varepsilon_i$ and making the error term the subject leads to $\varepsilon_i = Y_i - \alpha - \beta x_i$

- Take sum of squares: $\sum \varepsilon_i^2 = \sum (y_i - \alpha - \beta x_i)^2$
- Differentiate with respect to $\alpha \; and \; \beta$ and equate the result to zero

- $\frac{d\sum \varepsilon_i^2}{d\alpha} = -2\sum (y_i - \alpha - \beta x_i) = 0$ and $\frac{d\sum \varepsilon_i^2}{d\beta} = -2\sum x_i (y_i - \alpha - \beta x_i) = 0.$
- $taking \; the \; two \; equations \; and \; making \; y \; the \; subject \; to \; form \; least \; squa$
- $\sum y_i = n\alpha + \beta \sum x_i \;\; and \;\; \sum x_i y_i = \alpha \sum x_i + \beta \sum x_i^2$
- $solving \; the \; two \; simultaneously \; for \; \alpha \; and \; \beta$
- $\hat{\beta} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x)^2} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$
- $\hat{\alpha} = \frac{\sum x^2 \sum y - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} \; or \; \alpha = \bar{y} - \hat{\beta}\bar{x}$

3.4 The Standard Error of Estimate

Rarely does the predicted value of Y' agree exactly with the actual Y value. That is, we expect some prediction error. One measure of this error is called the ***'standard error of estimate'.*** This is written as S_{yx}.

The 'Standard error of estimate' is a measure of the scatter, or dispersion, of the observed values around the line of regression. A small standard error of estimate indicates that the independent variable is a good predictor of the dependent variable.

The standard error, as it is often called, is similar to the standard deviation described in earlier. Recall that the standard deviation was computed by squaring the difference between the actual value and the mean. This squaring was performed for all n observations. For the standard error of estimate, the difference between the predicted value Y' and the actual value of Y is obtained and that difference squared and summed over all n observations. The formula is:

$$S_{y.x} = \sqrt{\frac{\Sigma (Y - Y')^2}{n-2}}$$

A more convenient computational form is formula:

$$S_{yx} = \sqrt{\frac{\sum y^2 - \alpha \sum y - \beta \sum xy}{n-2}}$$

Where;

α and β are the regression coefficients

Σy^2 is the sum of the squares of the dependent variable

ΣY is the sum of the values of the dependent variables

ΣXY is the sum of the products of the dependent and independent variable

n is the sample size.

Using example 1 above, calculate its standard error of estimate.

Solution:

X	***Y***	***x^2***	***Xy***	***$y = \alpha + \beta x$***	***y^2***
1	6	1	6	2.5708	36
5	1	25	5	1.354	1
3	0	9	0	1.9624	0
2	0	4	0	2.2666	0
1	1	1	1	2.5708	1
1	2	1	2	2.5708	4
7	1	49	7	0.7456	1
3	5	9	15	1.9624	25
$\Sigma x = 23$	***$\Sigma x = 16$***	***$\Sigma x^2 = 99$***	***$\Sigma xy = 36$***		***$\Sigma y^2 = 68$***

From; S_{yx} = $\sqrt{\frac{\sum y^2-\alpha\sum y-\beta\sum xy}{n-2}}$ = $\sqrt{\frac{\mathbf{68-(2.875*16)-(-0.3042*36)}}{\mathbf{6}}}$= $\sqrt{\frac{\mathbf{68-35.0488}}{\mathbf{6}}}$

= $\sqrt{\frac{\mathbf{32.9512}}{\mathbf{6}}}$ = $\sqrt{\mathbf{5.4919}}$ = **2.34**

Regression assumptions the linear regression is based on these four assumptions;

- For each value of *x*, there is a group of *y* values, and these *y* values are normally distributed
- The means of these normal distributions of *y* values all lie on the straight line of regression
- The standard deviations of these normal distributions are equal
- The *y* values are statistically independent. This means that in the selection of a sample, the *y* values chosen for a particular *x* value do not depend on the *y* values for any other *x* value

3.5 DISTRIBUTION OF THE DEPENDENT VARIABLE Y AND THE PARAMETER ESTIMATES OF α AND β

The dependent variable Y is normally distributed with mean ($\alpha+\beta x_i$) and variance δ^2 which is estimated by;

$S^2_{yx}= \frac{\sum y^2-\alpha\sum y-\beta\sum xy}{n-2}$ or $\frac{\sum(y-\hat{y})^2}{n-2}$ or $\frac{n-1}{n-2}[s_y^2-\hat{\beta}^2 s_x^2]$

The estimator $\hat{\alpha}$ is normally distributed with mean α and variance $V(\hat{\alpha})=\delta^2[\frac{1}{n}+\frac{\bar{x}^2}{\sum(x-\bar{x})^2}]$ and the estimator $\hat{\beta}$ is normally distributed with mean β and variance $V(\hat{\beta})=\frac{\delta^2}{\sum(x-\bar{x})^2}$

3.51 **PROPERTIES OF LSE**

Using finite sample properties, LSE are

- Unbiased because they are Best Linear Unbiased estimators. Unbiased estimator is one whose expected value is equal to the unknown population parameter.
- They are efficient. That is, have minimum variance among all unbiased estimators.
- Have all the desirable asymptotic properties since they are the same as the maximum likelihood estimators (consistent)

EXAMPLE

Using the data below, estimate the regression equation, the variance δ^2 and the variances of the estimators.

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	98

3.52 C**OVARIANCE ($\widehat{\alpha}$, $\widehat{\beta}$)**

By using the estimators instead of the parameters α and β , sampling errors are committed. The sign of this error $E(\widehat{\alpha} - \alpha)(\hat{\beta} - \beta)$ is the covariance of $\hat{\alpha}\ and\ \widehat{\beta}$. This covariance is given as $cov(\hat{\alpha}, \hat{\beta}) = \frac{-\bar{x}\delta^2}{\sum(x-\bar{x})^2}$.

Example: using the above example, compute the covariance of the estimators.

CHAPTER FOUR

4.0 ESTABLISHING A CONFIDENCE INTERVAL FOR α AND β.

Confidence intervals provide an alternative way of expressing the uncertainity in the estimates. For a (1-α)100% confidence region, any point that lies within the region represents a null hypothesis that would not be rejected at the 100α% level while every point outside represents a null hypothesis that would be rejected. Confidence region provides a lot more information than a single hypothesis test in that it tells us the outcome of a whole range of hypothesis about the parameter values.

The formula for the confidence interval for the parameter estimates:

- $\hat{\beta} \pm t_{\propto/2}$, n-2*$s_\beta$, $s_\beta = \sqrt{S^2_{\hat{\beta}}}$, Variance $S^2_{\hat{\beta}} = \frac{MSE}{\sum(x-\bar{x})^2}$, and $MSE = \frac{\sum(y-\hat{y})^2}{n-2}$
- $\hat{\alpha} \pm t_{\propto/2}$, n-2*$s_\alpha$, $s_\alpha = \sqrt{s^2_\alpha}$ where $s^2_\alpha = MSE[\frac{1}{n} + \bar{x}^2 / \sum(x - \overline{x})^2]$

EXAMPLE: Use the previous example 1, to set the confidence interval for the data; given that $\propto$ = 0.05

Solution

x	*y*	x^2	*xy*	*y = a+bx*	y^2	$(y-\hat{y})$	$(y-\hat{y})^2$	$(x-\bar{x})$	$(x-\bar{x})^2$
1	6	1	6	2.5708	36	-30	900	-1.875	3.51563
5	1	25	5	1.354	1	0	0	2.125	4.51563
3	0	9	0	1.9624	0	0	0	0.125	0.01563
2	0	4	0	2.2666	0	0	0	-0.875	0.76563
1	1	1	1	2.5708	1	0	0	-1.875	3.51563
1	2	1	2	2.5708	4	-2	4	-1.875	3.51563
7	1	49	7	0.7456	1	0	0	4.125	17.01563
3	5	9	15	1.9624	25	-20	400	0.125	0.01563
23	**16**	**99**	**36**	**16.0034**	**68**	**-52**	**1304**	**0**	**32.875**

From;

$\widehat{\boldsymbol{\beta}} \pm \boldsymbol{\varepsilon}$ but; $\boldsymbol{\varepsilon} = t_{\propto/2}$, n-2*$s_\beta$, $S = \sqrt{S^2_{\hat{\beta}}}$, Variance $S^2_{\hat{\beta}} = \frac{MSE}{\sum(x-\bar{x})^2}$,

and *MSE* = $\frac{\sum(y-\hat{y})^2}{n-2}$

MSE = $\frac{\sum(y-\hat{y})^2}{n-2} = \frac{1304}{6}$ = **217.33, such** that

Variance$S^2_{\hat{\beta}} = \frac{MSE}{\sum(x-\bar{x})^2} = \frac{217.33}{32.875}$ = **6.611**

$S = \sqrt{S^2_{\hat{\beta}}} = \sqrt{6.611}$ = **2.571;** $t_{0.025}$**, 6 (2.447)**

Therefore;
$\hat{\beta} \pm \varepsilon$ = **-0.3042 ± (2.447*2.571) = [0.3042-6.2912, 0.3042+6.2912] = [-5.987, 6.5954].**

4.2 CONFIDENCE INTERVAL FOR Y

To determine the confidence interval for any given point (x_i) on the population/regression line within the given x domain, we need to get the mean and variance of $\hat{y}$. the confidence interval is given as

$$\hat{y} \pm t_{\frac{\gamma}{2},n-2}\sqrt{s^2_{xy}[\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x-\overline{x})^2}]}$$

EXAMPLE: given the data below on absenteeism and age of a worker at an institution as below;

X (age)	19	22	25	27	30	33	36	39
Y (absent)	8	10	9	7	5	6	5	4

Determine a 95% confidence interval on the regression line when x is 35.

4.3 COVARIANCE OF X AND Y

Covariance of x and y [Cov (*x, y*)] describes the strength and the direction of the linear relationship for random variables *x* and *y*. it is a measure of the relationship or association between the value of *x* and *y*. it is denoted as;

Cov (*x, y*) = $\frac{\sum[(x_i-\bar{x})(y_i-\bar{y})]}{n-1}$

It can be looked at as a measure of the way in which values of *x* and *y* vary together. Where, large values of *x* tend to go with large values of *y* and small values of *x* with small values of *y*.

The covariance will be ***positive*** for *x* larger and *y* smaller. For *x* smaller and *y* larger, covariance will be ***negative***. We can express the correlation coefficient r as a function of standard deviation of random variables x and y and covariance thus;

For a Population; **r =** $\frac{Cov\,(x,y)}{\sigma_x\sigma_y}$ and a Sample; **r =** $\frac{Cov\,(x,y)}{S_xS_y}$

Where;$\sigma_x S_x$ is the standard deviation of variable *x* and
$\sigma_y S_y$ Is the standard deviation of variable *y*

CHAPTER FIVE

5.0 GOODNESS OF FIT

To establish how good the fitted line $\hat{y}$ is to the sample observations of y, we can use the coefficient of determination R^2. For no variations, all the points will lie on a horizontal line equal to the mean of y but in reality when values of y are plotted against x, they scatter around the line $\hat{y}$ so that the variation of y can be measured by the difference between the observed values of y and $\bar{y}$ on the right hand side and the residual term on the left.

The total variation is equal to the variation due to the residual and due to regression;

- $\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$, that is
  SST=SSR+SSE(residual)

- Which is the same as; $\sum y^2 = \sum \hat{y}^2 + \sum e^2$

The measure of goodness of fit known as the coefficient of determination R^2 which is computed using the following;

$$R^2 = \frac{SSR}{SST} = \frac{\sum \hat{y}^2}{\sum y^2} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}$$

Or $R^2 = 1 - \frac{\sum e^2}{\sum y^2}$ $Or\ R^2 = \frac{\beta^2 \sum (x - \bar{x})^2}{\sum (y - \bar{y})^2}$

This coefficient lies between 0 and 1. A zero or near zero value of the coefficient signifies a poorest fit and a unit value (1) or near unit signifies the best fit.

A very low value of R^2 for a given sample means that;

a) Sample regression line fits the observations rather poorly i.e variations in x leave y un affected.
b) While x is the relevant explanatory variable, its influence on y is weak compared to the influence of the random disturbance.
c) It implies that the regression equation is misspecified.

The coefficient means that a percentage of the sample variation of y can be attributed to the variation of the fitted values of y.

EXAMPLE I

Given the data for price of a commodity and quantity sold in kilogrammes ,

Price (x)	100	90	80	70	70	70	70
Quantity(y)	55	70	90	100	90	105	80
Y^	57.499	69.999	82.499	94.999	94.999	94.999	94.999

i. Obtain SST,SSR and SSE

- total sum of squares (TSS) $=\sum (y - \overline{y})^2 = \sum y^2 - n(\overline{y})^2 = 51550 - 7 * 84.286^2 = 1821.429,$
- error sum of squares (ESS)$=\sum (y - \hat{y})^2$, where $y = a + bx$ such that

b= $\frac{n\sum xy-\sum x\sum y}{n\sum x^2-(\sum x)^2} = \frac{(7*45250)-(550*590)}{(7*44100)-550^2} = \frac{-7750}{6200} = -1.25\ and\ a = 84.286 + (78.571 * 1.25) = 182.499.$

$$y = 182.499 - 1.25x_i$$

$$ESS = 437.500.$$

- regression sum of squares (RSS)=$\beta^2 \sum(x_i - \bar{x})^2 \ \sum(\hat{Y} - \bar{Y})^2 = 1383.929.$

ii. Coefficient of determination R^2 and comment on the results

- **R^2**= SSR/SST = 1383.929/1821. 429 = 0.76. This implies a very good fit to the data.

5.1 **TESTING FOR SIGNIFICANCE OF REGRESSION**

The hypothesis is stated as H_o: model not significance versus H_A: model is significant at a given level of significance. The critical region is given as $F_c \geq F_{\propto,[k-1,N-k]}$ (reject H_o). where k are the parameters estimated and n is the sample size.

Using Analysis of variance (ANOVA), total variation is split into the explained variation and the unexplained variation; SST=SSR+SSE.
Using the ANOVA table for regression, the significance of regression can be determined using the f-test.

Source of variation	Sum of squares	Degrees of freedom	Mean sum of squares	f-computed
Regression	SSR	K-1	SSR/K-1=MSR	
Error	SSE	N-K	SSE/N-K=MSE	F_C=MSR/MSE
Total	SST	N-1		

Compare the computed f-statistic with the tabulated statistic at a level of significance.

Alternatively, for model significance the hypothesis can be stated as follows; H_o: $\beta_1 = \beta_2 = \cdots = \beta_k = 0\ versus\ H_A: not\ all\ \beta' sequal\ to\ zero$. the f-statistic is obtained using $F_c = \frac{\frac{R^2}{(k-1)}}{\frac{(1-R^2)}{(N-K)}}$, where R^2 is the coefficient of determination.

The critical region is $f_c \geq f_{\propto,[k-1,N-K]}$

EXAMPLE II: refer to example I above. Test for significance of regression using the f-test at 5% level of significance.

- H_0: not significant vs H_A: significant
- C.r: $f \geq f_{\propto,[1,n-k]}$, $f_{0.05,[1.5]=6.61}$

Anova table

S.o.v	Degrees of freedom	Sum of squares	Mean sum of squares	f-computed
Regression	1	1383.929	1383.929	1383.929/87.5=15.816
Error	5	437.50	87.5	
Total	6	1821.429		

- Decision: reject H_0. It is significant.

5.2 **TESTS OF HYPOTHESES**

To test the hypothesis that there is no relationship between the variables x and Y using the model $y = \alpha + \beta x$, the null hypothesis is stated as ; H_0: β=0 [no relationship between x and y]. if no prior information about the values of the regression parameters is available , the alternative hypothesis is stated as; H_A: β≠0.

The test statistic is given as $t = \frac{\beta}{s_\beta}$ at n-2 degrees of freedom. For a two tailed test (β≠0), the acceptance region is $-t_{\frac{\propto}{2},n-2} \leq \frac{\beta}{s_\beta} \leq t_{\frac{\propto}{2},n-2}$.

The best test is achieved if we take the alternative hypothesis as β<0, where the rejection region is $-t_{\propto,n-2} \leq \frac{\beta}{s_\beta}$.

If there is prior knowledge about the values of the parameters,

$H_0: \widehat{\beta} = \beta_0 \; and \; the \; test \; statistic \; is \; t = \frac{\widehat{\beta}-\beta_0}{s_\beta} \sim t_{n-2}$.

Alternatively, the F- test can be used to test for a relationship between the variables. The acceptance region for the hypothesis is $\frac{SSR}{(\frac{SSE}{N-2})} \leq F_{(1,N-2)}$.

EXAMPLE

Given the data below for minimum bank deposits in thousands of shillings and number of new accounts opened.

Branch	Minimum deposit (x)	New accounts (y)
A	125	160
B	100	112
C	200	124
D	75	28
E	150	152
F	175	156
G	75	42
H	175	124
I	125	150
J	200	104
K	100	136

i. Estimate regression model of the form $\hat{y}=\beta_0 + \beta_1 x_i$.

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{(11 * 186{,}200) - (1500 * 1288)}{(11 * 226250) - 1500^2} = \frac{116200}{238750}$$
$$= 0.487.$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} = 117.091 - (0.487 * 136.364) = 50.682$$

Equation: $y = 50.682+0.487x_i$.

ii. Variance for the estimator β_1.

$$v(\beta_1) = \frac{\delta^2}{\sum(x-\bar{x})^2}$$

$$where\ \delta^2\ can\ be\ estimated\ using\ s_{yx}^2 = \frac{\sum y^2 - \beta_0 \sum y - \beta_1 \sum xy}{n-2}$$
$$= \frac{170{,}696 - (50.682 * 1288) - (0.487 * 186200)}{11-2}$$
$$= 1637.576$$
$$therefore; v(\beta_1) = \frac{1637.576}{21704.546} = 0.0755.$$

iii. Test the hypothesis that $\beta_1 = 0\ against\ \beta_1 > 0$ at 0.05 level of significance.

- H_o: $\beta_1 = 0\ vs\ H_A$: $\beta_1 > 0$
- L.o.s α=0.05
- C.r: $t_c > t_{\propto,n-2},\ t_{0.05,9} = 1.833$
- $t_c = \frac{\beta_1}{s.e(\beta_1)} = \frac{0.487}{\sqrt{0.0755}} = 1.7724$
- Decision: fail to reject H_0.

4.0 **MULTIPLE REGRESSION**