International Journal of Innovative Scientific & Engineering Technologies Research 6(2):24-29, April-June, 2018

© SEAHI PUBLICATIONS, 2018 **www.seahipaj.org** **ISSN: 2360-896X**

Analyzing the Relationship between Linear Regression and Coefficient of Correlation Statistical Pattern: An Association of Two Variables Approach

[1]Oyetunde, Bamidele Sunday (Ph.D) & [2]Orime, Okechukwu C. N. (Ph.D)

[1]Department of General Studies (Mathematics and Statistics), Petroleum Training Institute, Warri, Nigeria

[2]Centre for Consultancy, Health, Safety and Security Management Programme, Ignatius Ajuru University of Education, Rumuolumeni, Port Harcourt, Nigeria

ABSTRACT

This study will be analyzing the relationship between Linear Regression and correlation. The aim of this work is to establish the relationship and the degree of association between two variables. The relationship between Linear Regression and Correlation was briefly discussed and illustrated with an example. The results were obtained with a pocket calculator. The value of the coefficient of coefficient (r = 0.75) shows that there is high positive degree of association between the two course. The coefficient of determination r^2 = 0.5625 (56.25%), indicates that the Regression line in the cause of about 56% of total variation in statistics result. The remaining 44% is attributed to factors in the error term. It was also found that the sign of the coefficient of correlation is the sign of the gradient of the Regression line. It was concluded that regression and correlation methods of analyses are quit close and one can easily move from one to the other.

Keywords: Analyses, Relationship, Coefficient, Regression, Correlation.

INTRODUCTION

There has been much confusion on the subject of Regression and Coefficient of Correlation quite frequently, Regression problems are treated as Correlation in scientific literature and the converse is equally true. However, one of the main objectives of science is to estimate values of one factor by reference to the values of an associated factor, and very often we find that researchers are interested in more than one characteristic of a group of individuals and the relationship existing among them.

Regression Analysis

Regression analysis involves the identification of the relationship between a dependent variable Y and one or more independent variable(s) X. A model of the relationship is hypothesized, and estimates of the parameter values where used to develop an estimated regression equation. Various tests are then used to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated Regression equation can be used to predict the value of the dependent variable given values of the independent variable. Hamilton (1997), said that Regression analysis determines the relationship between a dependent variable Y and an independent variable X. According to Ahamefule (1992), regression simply attempts to establish the nature of the relationship between variables, and the relationship may be Linear or a Parabola. Taha (2003), affirms that the simplest form of the regression model assumes that the dependent

variables varies Linearly with the independent variable and that the random error zero mean and a constant standard deviation. John and Frank (2002), posits that the Regression analysis is used for testing Hypothesis about the relationship between a dependent variable Y and an independent (or explanatory) X and for prediction, Maxwell (2014).

Correlation Analysis

Correlation Analysis is concerned with the type and degree of association between two variables, say X and Y.

Taha (2003), affirmed that correlation analysis tests how well the Linear regression model fits the available raw data and Sunder and Richard (2007), said that correlation ana1ysis is concerned with whether two variables are independent or co-vary (i.e. whether they vary together). This is also known as Karl Pearson's product Moment Correlation Coefficient (r).

According to Gaddis and Gary (1990), they posit that Correlation analysis is used to examine the strength of the relationship between two variables. Ahamefule (1992), said that correlation analysis is used to obtain a measure of the degree of association that exists between two variables. Ogbonria (2011); affirmed that correlation coefficient is the measure of the magnitude of the Linear relationship between two variables, say X and Y.

Karl Pearson defined coefficient of correlation r by

$r = \frac{\sum(x-\bar{x})(y-\bar{y}}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$, which is the same as $r = \frac{S_{xy}}{S_x S_y}$ … E1

Where S_{xy} - covariance of x and y respectively. The coefficient of correlation r ranges from - 1 to + 1 (i.e. $-\leq r \leq 1$). A coefficient of correlation of — 1 indicates at the two variables are perfectly related in a negative linear sense while a correlation coefficient of + 1 indicates that the two variables are perfectly related in a positive sense. A correlation of 0 indicates that there is no relationship between the two variables.

Regression and its Relationship with Coefficient of Correlation

Regression and Correlation are quantitative Mathematical tools which concerned with prediction and forecasting. They are related in the sense that both deals with relationship among variables. Neither Regression nor correlation can be interpreted as establishing cause-and-effect relationship. They can indicate c how and to what extent variables are associated with each other. They are widely used in every field of study. Linear regression analysis implies causality between the independent variable X and the dependent variable Y. However, coefficient of correlation simply refers to the type and degree of association between the variables. Indeed, the main use of correlation analysis is to determine the degree of association found in regression analysis. This is given by the coefficient of Determination (r_2) which is the square of the coefficient of correlation r.

Mathematical Relation

In a bi-variant data, if x is the independent variable and y is the dependent variable, the linear regression of y on x is given by the model

$y = \alpha + \beta x$ … E2

Where ct is the intercept on the Y-axis and)3 is the slope of the regression line. The values of cc. and are determined using the least square method. The value of α and β are given by the formulae:

$\beta = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$ … E3

and $\alpha = \bar{y} = \beta\bar{x}$ … E4

If on the other hand x is made the dependent variable and y the independent variable, the regression equation of x any becomes

$$x = \alpha + \beta y \quad \ldots \quad \text{E5}$$

Where $\beta = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2}$

Where x = α is the intercept on the X-axis and β is the slope of the regression line.

If the two regression lines are graphed on the same axes of the Cartesian plane, it will be noticed that they the same when r ±1. They will be at right angle when r = 0, if the variables are standardized, the two regression lines will intersect at the origin (0,0).

Proof

Let y = α + βx be a fitted regression line. The deviation from the regression line is denoted by $e = y - \hat{y}$

Talking the sum of squares of the above deviation gives.

$$\sum e^2 = \sum(y - \hat{y})^2 = \frac{\sum[(x-\bar{x})(y-\hat{y})]^2}{\sum(x-\bar{x})^2} \quad \ldots \quad \text{E6}$$

From E1 we have $\sum(x - \bar{x})^2 \sum(y - \bar{y})^2 = r^2 \left[\sum(x - \bar{x})^2 \sum(y - \bar{y})^2\right] \ldots$ E7

From E6 and E7 we have

$$\sum e^2 = (1 - r^2)\sum(y - \bar{y})^2 \ldots \quad \text{E8}$$

$$\sum e^2 = 0 \Longrightarrow r^2 = 1 \quad \therefore \quad r = +1.$$

Correlation as Square Root of the Coefficient of Determination

The coefficient of correlation (r) measures the degree of association between two or more variables. In the two variables case, the simple linear coefficient of correlation r for a set of sample observations is given by r = $\sqrt{r^2}$, Where r^2 is the coefficient of Determination.

The coefficient of determination r^2 is defined as the proportion of the total variation in y "explained" by the regression of y on x. it is given by

$$r^2 = 1 - \frac{\sum(y-\hat{y})}{\sum(y-\bar{y})} = 1 - \frac{\sum e^2}{\sum y_i^2} \quad \ldots \quad \text{E9}$$

The square root of the coefficient of Determination r^2 is called the coefficient of correlation r or more explicitly PEARSON PRODUCT MOMENT COEFFICIENT OF CORRELATION. It measures the strength of the relationship between two or more variables, thus,

$$r = \pm\sqrt{1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}} \quad \ldots \quad \text{E10}$$

Proof

The total variation in y on x or the total sum of squares (TSS) is given by

$$TSS = \sum(y - \bar{y})^2 = \sum \bar{y}_i^2 \quad \ldots \quad \text{E11}$$

The explained variation in y or Regression sum of squares (RSS) is given by

$$RSS = \sum(\hat{y}_i - \bar{y})^2 = \sum \hat{y}^2 \quad \ldots \quad \text{E12}$$

The residual variation in y or Error sum of squares (ESS) is given by

$$\text{ESS} = \sum(y_i - \hat{y})^2 = \sum e_i^2 \quad \ldots \quad \text{E13}$$

Total sum of squares = Regression sum of squares + Error sum of squares

i.e TSS = RSS + ESS $\Longrightarrow \sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{y})^2 \sum(y_i - \hat{y})^2 \quad \ldots \quad \text{E14}$

Dividing both side by $\sum(y_i - \bar{y})^2$, we have

$$1 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} + \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

$r^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$

Thus, $\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = 1\frac{\sum e^2}{\sum y_i^2}$

This is the Coefficient of Determination. The square root of the coefficient of determination r^2 is the coefficient of correlation r and is given by

$r = \pm \sqrt{1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}, where\ 0 \leq r^2 \leq 1.$

r^2 = 0 when the estimated value of the regression equation explains non of the variation in y.

$r^2 = \pm 1$ when all points lie on the regression line.

Illustration

The table below shows the marks in STA111 (Introduction to Statistics) and STA112 (Descriptive Statistics) obtained by seven (7) statistics students in first semester examination in the year 2011 in the Department of Mathematics and statistics, in Rivers State College of Arts and Science, in Port Harcourt.

Table: 1

STA111(X)	**82**	**80**	**77**	**75**	**74**	**69**	**68**
STA112(Y)	75	74	70	68	66	68	69

Source: Field Survey (2014).

1. Find the regression line of descriptive statistics (STA112:Y) on Introduction to Statistics (STA111:X)
2. Calculate the coefficient of correlation r.

Solution

Table 2:

X	**Y**	**X - □**	**y - □**	**(X -□) (y -) □**	**(X - □)**	**(y - □)^2**
82	75	7	5	35	49	25
80	74	5	4	20	25	16
77	70	2	0	0	4	0
75	68	0	-2	0	0	4
74	66	-1	-4	4	1	16
69	68	-6	-2	12	36	4
68	69	-7	-1	7	49	1
$\sum(X - \square) = 529$	$\sum(y - \square) = 490$			$\sum(X -)(y - \square) = 78$	$\sum(X - \square)^2 = 164$	$\sum(X - \square)^2 = 66$

Source: Field Survey (2014)

$$\square = \frac{\sum X}{N} = \frac{529}{7} = 75\ and\ \square = \frac{\sum y}{N} = \frac{490}{7} = 70$$

With the regression equation of y on x as

$y = \alpha + \beta x\ \ where$

$$\beta = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}\ and\ \alpha = \bar{y} - \beta\bar{x}$$

$\beta = \frac{78}{164} = 0.4756$ and

$\alpha = 70 - (0.4756)(75)$ 34.33

The regression equation of descriptive statistics y on introduction to statistics x is

y = 34.33 + 0.48 x.

The coefficient of correlation r is

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} = \frac{78}{\sqrt{(164)(66)}} = \frac{78}{\sqrt{1024}} = 0.75 = 75\%.$$

The coefficient of determination $r^2 = (0.75)^2 = 0.5625 = 56.25\%$.

Interpretation of Result

There is a high degree of correlation between the two courses. A positive relationship exists between the two courses. This implies that, if a student high in Introduction to Statistics, he/she is likely to scores high in Descriptive Statistics and vice-versa.

Thus, the equation of regression of y on x explains about 56% of the total variation in Statistics result while the remaining 44% is attributed to factors included in the error term.

CONCLUSION AND RECOMMENDATION

According to Alan and Porter (1999) Linear regression and correlation is the alternative ways of examining two variables. Correlation analysis determines the degree of association found in regression analysis. This is given by the coefficient of determination r^2. Neither regression nor Correlation analysis can be interpreted as establishing cause and effect relationships. They can indicate only how or to what extent variables are associated with each other. The coefficient of correlation r = 0.75 calculated, shows that there is a positive high degree of correlation (association) between the two courses. This implies that a student who scores high in Introduction to Statistics is likely to score high in Descriptive statistics and vice-versa.

Since $r^2 = 0.5625$ (56.25%), this implies that the line of regression explains about 56 percent of the total variation to statistics scores while the remaining 44% can be distributed to factors included in the error term.

It is also important to note that the sign (- or +) of the gradient (slope) of the line of regression is also the same sign of the coefficient of correlation r for any set of data values of two variables.

The Mathematical relations between the two methods of analysis are quite close one can move from either method to the other.

REFERENCES

Ahamefule, M.U; (1992); Operations research and quantitative methods for Management Sciences. Frontier Publishers Limited, Okigwe Road, Aba, Nigeria pp; 107-115.

Alan, M. W. and Porter, M. D., (1999); Misuse of concrete and regression Journal of Royal Society of Medicine. 92: 123- — 125.

Gaddis, M. L. and Gery, O. M.; (1990); Annals of Emergency Medicine. Published by Elsviar- inc. 1 (12): 1464- — 1466.

Hamdy, A. T. (2003); Operations Research: An introduction. Seventh edition. Prentice- Hall of India, Private Limited New Delhi pp; 497 - 499.

Hamilton, O. I. (1997); Fundamental of Business Forecasting. Dominican Publishers, Aba, Nigeria pp, 48—50.

John, E. F and Frank, J. W., (2002). Elementary Business Statistics. The modem approach. Third edition. Prentice- Hail international, Inc. India. pp; 400- 404.

Maxwell, A. A.; Elendu, B. N., Elek, A. G. (2014). Analysis of the relationship between linear regression and coefficient of correlation. *International Journal of the Institute For Empirical Research and Sustainable Development*. 11, 75-79

Ogonna, L. N., (2011): African Journal of Professional Research in Human Development. 7(2), pp; 15 - 17.

Sunder, P. S. S. and Richard, J.; (2007); Introduction to I3iostatistics and Research Methods. Prentice - Hall of India, Private Limited, New Delhi. pp; 85-90.